Notes on Bayesian Statistics

Chiranjit Mukhopadhyay
Indian Institute of Science

1 Introduction

Phenomena involving uncertainties are typically studied using probability models. In all such phenomena usually there are certain observables, called random variables and understanding of such phenomena are considered to be complete when one knows about the kind of distributions (also called "law") these random variables follow. The important problem of interest involving discovering relationship between different random variables are also expressed in uncertain terms through these probability distributions. (This is because, that is the best we can do when faced with uncertainties.)

The problem of statistical inference involves gathering information or evidence about these typically unknown probability distributions (conceptualized as the "true" behavior of the variables under consideration in a real or hypothetical population) based on a finite number of observations from them, called a sample. To put it in mathematical terms, let Y denote an observable random variable with cumulative distribution function or c.d.f.[1] $F(y)$. As data one has a random sample $Y_1, Y_2, \ldots, Y_n$ on Y drawn from its (population) distribution $F(y)$ - typically expressed as $Y_1, Y_2, \ldots, Y_n$ i.i.d. (standing for independent and identically distributed) $F(y)$. The problem of statistical inference usually consists of the following:

1. Estimate $F(y)$ or any of its functional like mean $\mu = \int_{-\infty}^{\infty} y \; dF(y)$, variance $\sigma^2 = \int_{-\infty}^{\infty} (y-\mu)^2 \; dF(y)$ or median $\xi_{0.5}$, which is a solution of the equation $F(y) = 0.5$.
2. Test hypotheses about $F(y)$ or its functionals; like $F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right)$ for some μ and σ, where $\Phi(\cdot)$ is the standard Normal c.d.f.; or $\mu > 5$. OR
3. Predict the behavior of a future observation like Y_{n+1}

based on the random sample $Y_1, Y_2, \ldots, Y_n$ on Y.

Solution of above problems in general is sometimes fairly hard if one allows the c.d.f. $F(\cdot)$ to be completely arbitrary. Methods applicable in situations involving arbitrary $F(\cdot)$'s are called non-parametric or distribution-free methods. While such methods have the appeal of assuming very little about the unknown $F(\cdot)$, the price one typically pays comes in terms of larger sample size and in general increased complexity of analysis. In any case before attacking a problem in its most complex form, it is always instructive to analyze the problem in simpler terms if not for anything else, but at least for getting acquainted with the methodological issues and probably gaining some extra in sight on the way.

One such simplification is attained by assuming the distribution of Y *i.e.* $F(y)$ belongs to some given parametric family of distributions like Binomial, Poisson, Normal, Exponential

[1] C.d.f. of a random variable Y is defined as $F(y) = P(Y \leq y)$. There are several ways of characterizing the distribution of a random variable, but mathematically c.d.f. is the most attractive choice because of its existence for all types of random variables and thus its utility as a unifying concept.

etc. as opposed to coming from an arbitrary non-parametric[2] family. In this parametric set-up the form of $F(\cdot)$ is essentially assumed to be known barring a few unknown parameters denoted by $\boldsymbol{\theta}$[3], and is thus written as $F(y|\boldsymbol{\theta})$. This considerably simplifies the problems of statistical inference. Because now all the quantities of interest (in the population), such as the ones mentioned in 1, 2 and 3 above, can be expressed in terms of the p θ_j's, $j = 1, 2, \ldots, p$, and the problem of inference reduces to drawing inference about only these p unknown θ_j's for $j = 1, 2, \ldots, p$ based on n i.i.d. observations on $Y \sim F(y|\boldsymbol{\theta})$[4].

Of course one can theoretically, and in many important practical applications (such as Time Series Analysis), has observations which are not i.i.d.. Bayesian analysis of such data is going to be exactly same as the methods that are developed in these notes. The only difference in the non-i.i.d. case would be in the form of the likelihood function. This point will again be mentioned elsewhere in appropriate places in these notes. Thus though we shall begin our discussion with the i.i.d. case, these notes are not going to be restricted to this i.i.d. case alone. However we shall confine ourselves only to the case of parametric Bayesian Inference in these notes leaving extremely theoretically intensive Bayesian non-parametrics out from the realm of our discussion. We begin our discussion with some simple review examples of the nature of problems addressed in statistical inference.

Example 1: Suppose we are interested in π, the probability of a consumer choosing brand X of toothpaste. For this problem the underlying probability model is a simple one. Let the random variable Y denote 1 if a consumer chooses brand X of toothpaste and 0 otherwise. Then $Y = \begin{cases} 1 & \text{with probability } \pi \\ 0 & \text{with probability } 1-\pi \end{cases}$ which has the probability mass function or p.m.f.[5] $p(y|\pi) = \pi^y(1-\pi)^{1-y}$ for $y = 0, 1$. Now let us observe the choice of toothpaste brand for n consumers denoted by $Y_1, Y_2, \ldots, Y_n$. Note that for $i = 1, 2, \ldots, n$ each of the Y_i's is 0-1 valued having the same p.m.f. $p(y|\pi)$ as above. Now one simple problem could be obtaining a single valued **point estimate** of π based on the data $Y_1, Y_2, \ldots, Y_n$. But due to uncertainty in this point estimate we might next look for an interval of values called an

[2]Actually the term "non-parametric" is somewhat of a misnomer. It is used in the sense of something which is not parametric. But mathematically in the so-called non-parametric situation the number of parameters is infinite in contrast to the parametric case, where one has to deal with possibly a few but in any case only finitely many parameters, like the probability of success of Binomial, or mean and variance of Normal, or the mean of Poisson, or the failure-rate of Exponential, distributions. Thus many statisticians prefer to refer the methods associated with an arbitrary $F(\cdot)$ not assumed to belong to any particular family of probability models as distribution-free methods instead of non-parametric methods.

[3]In general we shall deal with more than one (but finitely many, say p) unknown parameters $\theta_1, \theta_2, \ldots, \theta_p$. All these unknowns are collectively denoted by the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$. The matrix transposition operator $'$ is used to adhere to the convention, that by default all vectors are column vectors. Thus $\boldsymbol{\theta}$ is a $p \times 1$ vector (matrix) of p unknown θ_j's for $j = 1, 2, \ldots, p$.

[4]The symbol $\sim$ is read as "as distributed as".

[5]For a discrete random variable Y, such as the one in this example, its p.m.f. is defined as $p(y) = P(Y = y)$. P.m.f.'s are a convenient way of characterizing the distributions of discrete random variables. Note that given a p.m.f. $p(y)$ one can construct its c.d.f. $F(y)$ as $F(y) = \sum_{x:x \le y} p(x)$, and given a c.d.f. $F(y)$ its p.m.f. is given by $p(y) = F(y) - \lim_{x \to y-} F(x)$. Thus there is a one-to-one correspondence between the p.m.f. and c.d.f. of a discrete random variable, and one can use either one to characterize a discrete distribution. It is always more convenient to deal with the p.m.f. in the discrete case, for the statistical inference purpose. C.d.f. is typically used while discussing in general terms.

interval estimate which we hope will contain or feel confident about containing the true unknown value of π. The marketing team might claim that more than 10% of the consumers are choosing brand X. Then we have this task of justifying the validity of such a claim or **test** the **hypothesis** that $\pi > 0.10$. All these things we have to do based on our observations $Y_1, Y_2, \ldots, Y_n$ which are i.i.d. $p(y|\pi)$. These are some examples of typical problems that one hopes to tackle in statistical inference. $\bigtriangledown$

Example 2: Let X denote the monthly advertising expenses, in lakhs of Rs., and Y denote the monthly sales, in crores of Rs., of a company. As a first step in modeling a relationship between X and Y it is postulated that if at all X has any effect on Y it is going to be linear in nature, and further the conditional variance of monthly sales given any fixed level of monthly advertising expenses is a constant *i.e.* does not depend on X, and the conditional distribution of monthly sales is Normal. These postulates in a nut-shell can be expressed as the probability model $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$. In this model β_1 represents the expected increase in monthly sales for every unit increase in monthly advertising expenses. That is for instance if $\beta_1 = 0.02$, according to the model, for every lakh of Rs. increase in advertising expense, one can expect the sales to go up by 0.02 crores or Rs.2 lakhs. Thus a β_1 value of less than 0.01 may not be a very profitable proposition for increasing the advertising expenses. If it is and on a certain month the company decides to spend Rs.10 lakh on advertising then we would like to **predict** either in a point or an interval sense (better yet get the distribution itself) what the company's sales is going to be for that month. All these, namely deciding whether $\beta_1 > 0.01$ or not, and getting the distribution of Y when $X = 10$ (for example) has to be done based on the past data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ on n months. (Otherwise how are you to know the values of the parameters β_0, β_1 and σ^2?) These are again some typical problems addressed in statistical inference. $\bigtriangledown$

So how are these typical problems of statistical inference involving **point** and **interval estimation**, **hypothesis testing**, and the problem of **prediction** or **forecasting** handled? For this a quick review of philosophy of the methods within the standard classical or frequentist paradigm may not be very inappropriate.

2 Frequentist Inference

In the frequentist setting, merit of any method is always judged in terms of the method's behavior over repeated sampling. This is done by considering what is called the **sampling distribution** of a **statistic**. A statistic is nothing but whatever is computed using the data $Y_1, Y_2, \ldots, Y_n$ at hand. If $T(Y_1, Y_2, \ldots, Y_n)$ is a statistic, then since it is a function of the random variables $Y_1, Y_2, \ldots, Y_n$, it itself is a random variable. Of course for a given set of data at hand like $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, the statistic T will have an observed value of $T = t$, but conceptually T being a function of random observations (albeit following the law $F(y|\boldsymbol{\theta})$) is itself a random variable and thus one can talk about its distribution which can be derived from the parent distribution $F(y|\boldsymbol{\theta})$. This distribution of a statistic T is called its sampling distributions.

This distribution is called the sampling distribution because of the following interpretation. For a given sample one gets an observed value of T. Now think about obtaining another sample of size n and then computing the value of T. You will most probably get a different value from the one you have obtained earlier. Now imagine repeating this process till all possible samples of size n have been drawn. For each of this sample compute the value of T and then study the distribution of these values of T over all possible samples of size n from $F(y|\boldsymbol{\theta})$. This is same as the (possibly theoretically derived) distribution of the function $T(Y_1, Y_2, \ldots, Y_n)$ where $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $F(y|\boldsymbol{\theta})$. Thus this distribution of T tells us how the values of T would behave over repeated sampling from the same population and is thus called the sampling distribution of the statistic T.

For example as in Example 1, if $Y_1, Y_2, \ldots, Y_n$ are i.i.d. Bernoulli(π) then the sampling distribution of the statistic of interest $\sum_{i=1}^{n} Y_i$ is Binomial(n, π). If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$, the statistics (sample mean, sample variance) denoted by $(\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, s^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2)$ are independent with $\overline{Y} \sim N(\mu, \sigma^2/n)$ and $s^2_{n-1} \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$.

In the rest of this section, where we very briefly take upon the frequentist approach towards addressing the different inference problems, to keep matters simple, we shall only discuss the case of a single parameter θ instead of the vector-valued multi-parameter case of $\boldsymbol{\theta}$. In any case discussions in the following subsections do not get into any in-depth methodological development issues. They are just meant to provide only very brief logical reasonings behind the frequentist approach to statistical inference, and thus nothing is essentially lost if we only consider the case of a scalar θ.

2.1 Point Estimation

Point estimation is concerned with providing a single valued estimate for an unknown population parameter θ. The optimality of such a point estimator of θ is judged by its behavior over repeated sampling or in terms of the properties of its sampling distribution. For instance in Example 1, $\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, the proportion of consumers choosing brand X toothbrush in the sample, is the "best" point estimator of π in the sense that it is the **Uniformly Minimum Variance Unbiased Estimator** (UMVUE) of π. First of all $\hat{\pi}$ as defined above is Unbiased[6] because $E_\pi[\hat{\pi}] = \pi \ \forall 0 < \pi < 1$ (easy to show) and then it can be shown (a tricky task) that among all possible unbiased estimators T of π, $\hat{\pi}$ is the one which possesses the property that $V_\pi[\hat{\pi}] \leq V_\pi[T] \ \forall 0 < \pi < 1$, where the variances $V_\pi[\cdot]$ are computed based on the sampling distributions of the respective statistics, establishing that $\hat{\pi}$ is the UMVUE[7] of π.

[6] A statistic T is said to be an **unbiased** estimator of a parameter θ if $E_\theta[T] = \theta \ \forall \theta \in \Theta$, where Θ, called the parameter space, is the set of possible values the parameter θ can take, and the expectation $E_\theta[T]$ is taken over the sampling distribution of T. Thus it means that if an estimator is unbiased, on an average it hits the target right, in the sense of repeated sampling.

[7] In general an estimator $\hat{\theta}$ of θ is called the **UMVUE** of a parameter θ, if it is unbiased for θ and has the minimum variance (computed w.r.t. their respective sampling distributions) among all other unbiased estimators of $\theta \ \forall \theta \in \Theta$ *i.e.* $E_\theta[\hat{\theta}] = \theta \ \forall \theta \in \Theta$ and $V_\theta[\hat{\theta}] \leq V_\theta[\hat{\theta}'] \ \forall \theta \in \Theta \ \forall \hat{\theta}' \ni E_\theta[\hat{\theta}'] = \theta \ \forall \theta \in \Theta$. The $E_\theta[\cdot]$'s and $V_\theta[\cdot]$'s are subscripted with a θ to emphasize and indicate the fact that, in general the sampling distributions and thus the first and second moments of a statistic depend on the unknown parameter θ.

For the problem of point estimation, in the frequentist paradigm, one usually first strives to obtain an UMVUE. If the UMVUE exists for a problem, the problem of point estimation is considered to be solved. However there is no direct straight-forward algorithmic way of obtaining a UMVUE. That is given a $F(y|\theta)$ and $Y_1, Y_2, \ldots, Y_n$ i.i.d. $F(y|\theta)$ there is no immediate formula or numerical method of obtaining a UMVUE of θ or some function $\phi(\theta)$ that might be of interest. In fact UMVUE may not even exist in some cases. In such situations one usually employs methods of Maximum Likelihood or Moments or Least Squares or Minimum χ^2 to arrive at an estimate. In any case the optimality of any such resulting estimator is judged in terms of its behavior over repeated sampling or in terms of the properties of its sampling distribution. In most of the cases the derivation of the exact sampling distribution of such estimators become tedious or impossible. In such situations, which are practically the norm rather than being exceptions, one tries to obtain at least a large sample approximation (as the sample size $n \to \infty$) of the sampling distribution of the estimators and then judges or compares its behavior with other competing estimators in terms of these approximated sampling distributions. Simulating the sampling distribution using a technique called bootstrap (instead of an analytical large sample approximation) is also widely used. Whatever technique one might use, the bottom line is that, in the frequentist paradigm optimality or desirability of a point estimator is judged in terms of its sampling distribution, which is nothing but the distribution of the estimator over repeated sampling from the same population.

As mentioned above, a point estimator of a parameter yields a single estimated value of the parameter. Now since this value depends on the sample, which is subject to sampling fluctuation and thus uncertain, it is customary to report the amount of *error* that is inherent in the value obtained by using an estimator $\hat{\theta}$. This error is called the **Standard Error** of $\hat{\theta}$ denoted by $SE_\theta(\hat{\theta})$, which is defined as $\sqrt{V_\theta[\hat{\theta}]}$, which is nothing but the standard deviation of the estimator $\hat{\theta}$ calculated according to its sampling distribution. But $SE_\theta(\hat{\theta})$ needs to be interpreted with some caution. It does not mean that the true value of the unknown parameter θ lies somewhere within $\hat{\theta} \pm SE_\theta(\hat{\theta})$. In order to obtain such interval estimators one has to go a step further and introduce the notion of confidence interval in the frequentist set-up.

2.2 Interval Estimation

We begin our discussion of interval estimation with an example. Suppose we have a random sample of size n from a Normal population with known population variance σ^2. Or in other words let $Y_1, Y_2, \ldots, Y_n$ be i.i.d.$N(\mu, \sigma^2)$ with known σ^2. Then it can be shown that $\hat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the UMVUE of μ with $SE_\mu(\hat{\mu}) = \sigma/\sqrt{n}$ ($SE_\mu(\hat{\mu})$ is not subscripted also with σ^2 because σ^2 is not an unknown parameter). But in order to have an interval estimator of μ one has to go beyond simply the mean and variance (standard deviation) of $\hat{\mu}$ and study its sampling distribution in totality. As stated earlier it can be shown that $\overline{Y} \sim N(\mu, \sigma^2/n)$ and using this result one can then make a probability statement like

$$P\left(\overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha, \quad \forall 0 < \alpha < 1 \tag{1}$$

where z_α is the α-th quantile of a standard Normal distribution *i.e.* z_α is such that $P(N(0,1) \leq z_\alpha) = \alpha$. Equation (1) now yields a $100(1-\alpha)\%$ **Confidence Interval** (CI) of μ as the interval $\overline{Y} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$. This CI of μ is of the form $\hat{\mu} \pm z_{1-\alpha/2}SE_\mu(\hat{\mu})$. Thus standard error of an estimator is best interpreted in terms of an interval estimator of the form: (point estimate) $\pm$ (a blow-up factor) $\times$ (its standard error), where the blow-up factor depends on the degree of confidence one wants in one's interval estimator and thus controlling the width of the interval. As intuition suggests, larger the confidence wider is the interval.

A couple of remarks regarding CI are in order. First of all, all CI's are not of the form (point estimate) $\pm$ (a blow-up factor) $\times$ (its standard error) (*e.g.* CI of σ^2 for a random sample from a $N(\mu, \sigma^2)$ population). Thus the standard error of an estimate is not always as neatly tied up with its CI as in the above example. However one class of asymptotic CI (called Wald intervals) of parameters of so-called "regular" models usually have this close tie with their respective asymptotic standard errors, which is based on the large sample approximation of the sampling distribution of the corresponding point estimator.

The second remark concerning the interpretation of CI is more important. For instance, it would be wrong to say that the probability that a $100(1-\alpha)\%$ CI of θ contains its true unknown value is $1-\alpha$. For getting the correct interpretation let us look back at the basic probability statement in equation (1) leading to the CI formula for a Normal mean μ (all CI formulæ are derived from similar probability statements involving the appropriate sampling distribution of an estimator). What is random in the l.h.s. of (1) about which the probability statement is being made? The correct answer is NOT μ but $\overline{Y}$. That is the correct interpretation of the probability statement in (1) is that we have a random interval $[\overline{Y} - z_{1-\alpha/2}\sigma/\sqrt{n}, \overline{Y} + z_{1-\alpha/2}\sigma/\sqrt{n}]$ and the probability that this random interval will capture the true unknown value of μ is $1-\alpha$. Given a realized value of this interval based on a sample at hand, as it is used in applications, the probability that it contains the true unknown value of μ is either 1 or 0 depending on whether μ really falls into it or not - it has nothing to do with the confidence level $100(1-\alpha)\%$. However if you use the same formula for the random interval and apply it over again and again over repeated sampling then the proportion of time this interval will contain μ is going to be $1-\alpha$. Thus the confidence level must be interpreted as the coverage probability of a CI over repeated sampling.

2.3 Hypothesis Testing

Next let us look at the issue of hypothesis testing in a frequentist context. There are essentially two schools of thoughts for the method of testing a statistical hypothesis even within the frequentist paradigm. One is called the fixed significance level testing and the other is the observed significance level testing or the p-value approach. In either case, the problem of testing a statistical hypothesis is formulated as a decision problem of deciding between whether an unknown parameter θ belongs to Θ_0 or Θ_a, where Θ_0 and Θ_a are disjoint subsets of the parameter space Θ (see footnote 6 in page 4) *i.e.* $\Theta_0 \subseteq \Theta$, $\Theta_a \subseteq \Theta$ and $\Theta_0 \cap \Theta_a = \phi$, the null set. The statement $\theta \in \Theta_0$ is called the null hypothesis and is denoted by H_0, whereas the statement $\theta \in \Theta_a$ is called the alternative hypothesis and is denoted by H_a. The problem of hypothesis testing is one of deciding between H_0 and H_a, in light of the observed data $Y_1, Y_2, \ldots, Y_n$ i.i.d. $F(y|\theta)$.

In neither approach, fixed or observed significance level, the null and the alternative hypotheses get a symmetric treatment. In both the approaches the standing is that, the null hypothesis H_0 describes what is called the *status quo* of a situation, while the alternative hypothesis states a point that we want to prove in light of the data. That is unless proven otherwise, the decision goes in favor of H_0 giving it the benefit of doubt. The data has to carry enough evidence beyond any reasonable doubt to establish the truth of the alternative hypothesis H_a, which in it contains a point we wish to prove against the current *status quo* of a situation. The situation is analogous to that in jurisprudence, where a person is not convicted unless proven guilty. That is the null hypothesis is, "the accused is not guilty" and the alternative is its compliment.

In the fixed significance level testing one begins with the consideration of possible consequences of taking a decision in favor or against the null hypothesis H_0, expressed in terms of do not Reject or Reject H_0 respectively, for the two possible true states of nature H_0 and H_a, as summarized in Table 1.

Table 1: Consequences of Different Decisions

Decision Taken $\rightarrow$ The Truth $\downarrow$	Reject H_0	Do not Reject H_0
H_0 is True	Type-I Error	$\surd$
H_a is True	$\surd$	Type-II Error

A $\surd$ in Table 1, indicates that a correct decision has been taken, while there are two possibilities of committing mistakes, classified as the two types of errors. As mentioned in the previous paragraph, since H_0 is supposed to get a favorable treatment, Type-I Error is considered to be more serious than Type-II Error, and hence is the naming classification of the two errors, despite the fact that in either case one rejects the right hypothesis.

Now based on the data $Y_1, Y_2, \ldots, Y_n$ i.i.d. $F(y|\theta)$, one needs to take the decision of whether to Reject H_0. The kind of data which leads to the decision, Reject H_0, is called a **Critical Region**, abbreviated as *CR*. Formally, $CR = \{\boldsymbol{y} : \text{If } \boldsymbol{Y} = \boldsymbol{y} \text{ then Reject } H_0\}$, where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$ is the $n \times 1$ random vector of observations and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ is its realized value in a given sample. Thus $CR \subseteq \mathcal{Y}$, where $\mathcal{Y}$ is the set of all possible values one can observe as a sample, and is thus called the sample space. The problem of taking an optimal decision thus reduces to obtaining an optimal *CR*.

Ideally an optimal *CR* should be such that one does not commit either type of errors. But note that even if the *CR* is completely specified, and one knows the truth about H_0, one still does not know whether one is committing one type of error or other or not, because the decision is being taken based on a random vector $\boldsymbol{Y}$. Thus the best one can do is try to reduce the probabilities of committing either types of errors. Therefore one next systematically introduces these error probabilities as follows. For $\theta \in \Theta_0$ *i.e.* when H_0 is true, the probability of Type-I error is given by $\alpha(\theta) = P(\boldsymbol{Y} \in CR|\theta)$. Note that for $\alpha(\theta)$ to qualify as probability of Type-I error its domain must be Θ_0. Likewise for $\theta \in \Theta_a$ *i.e.* when H_a is true, the probability of Type-II error is given by $\beta(\theta) = 1 - P(\boldsymbol{Y} \in CR|\theta)$, with the domain of $\beta(\theta)$ being Θ_a. Thus ideally an optimal test should be such that both $\alpha(\theta)$ and $\beta(\theta)$ are small for $\theta \in \Theta_0$ and $\theta \in \Theta_a$ respectively.

But now observe that both $\alpha(\theta)$ and $\beta(\theta)$ involve $P(\boldsymbol{Y} \in CR|\theta)$ with their signs occurring in reverse directions. Thus in general reducing one leads to an increase in the other. This problem is solved by appealing to the original philosophy of inherent bias towards H_0. According to this reasoning, Type-I error is considered to be more serious than Type-I error. Thus the problem of deciding on the optimal CR starts with first fixing a small maximal probability of Type-I error, say α_0. Now among all tests or CR's satisfying $\sup_{\theta \in \Theta_0} \alpha(\theta) = \alpha_0$, one chooses that test or CR which has the uniformly smallest $\beta(\theta)$ $\forall \theta \in \Theta_a$. Fortunately, such optimal CR's exist and may be explicitly obtained for many important practical cases, and the test thus obtained is called a **uniformly most powerful** test of **size** α_0.

The definition of the **size** of a test is the maximal probability of Type-I error. **Power** of a test, or more appropriately the **power function** of a test is same as the probability of Rejecting H_0 for a given value of θ, which is same as $P(\boldsymbol{Y} \in CR|\theta) = 1 - \beta(\theta)$. Thus a test having uniformly smaller $\beta(\theta)$ compared to other tests is same as saying that it is uniformly more powerful than other tests. This explains why such a test as above is called a **uniformly most powerful** test of **size** α_0. In many (so-called two-tailed alternative) situations, a uniformly most powerful test may not exist. Then together with the size constraint one introduces another requirement called unbiasedness[8] and attempts to obtain **uniformly most powerful unbiased** test of **size** α_0.

In the above approach since one fixes the size, also called the level of significance, of a test, the resulting optimal tests are called fixed significance level tests. However the point that is to be noted for these fixed significance level tests is that, whether one talks about the size or power of a test, these probabilities pertain to the observed data $\boldsymbol{Y}$, and thus a small size or large power refers to the behavior of the test over repeated sampling from the same population, as in the earlier cases of point and interval estimation.

The observed significance level or the p-value approach to testing statistical hypothesis, though frequentist in nature, comes from a different philosophical perspective than that of the fixed significance level testing, though in practical implementation one might appear to be a very close cousin of the other. Here one does not view the problem of testing a statistical hypothesis as a decision taking problem as depicted in Table 1. Rather one tries to assess the strength of evidence the data is exhibiting for or against the null hypothesis H_0. In order to do this it first proposes a test-statistic $T(\boldsymbol{Y})$ such that larger the value of T more is the evidence against H_0, and then it defines the observed significance level or p-value as $P(T \geq T_{\text{observed}}|H_0)$, where T_{observed} is the observed value of the statistic T *i.e.* $T_{\text{observed}} = T(\boldsymbol{y})$.

The idea behind this definition is that p-value in a nut-shell gives the amount of evidence the data is carrying against the null hypothesis in a 0-1 scale. Smaller the p-value more is the evidence against H_0. This is because for computing the p-value one starts with the assumption that H_0 is true. Then under this assumption, one first assesses the kind of behavior to expect out of the test statistic T in terms of its sampling distribution under H_0. Then one sees how far in the right-tail (this is because larger the value of T more is

[8] A test is said to be **unbiased** if its Power$\geq$Size. For uniformly most powerful tests, this condition is automatically satisfied.

the evidence against H_0) of this distribution is the observed value of T yielded by the data set at hand is sitting. This is quantified in terms of the p-value. Small p-value indicates that T_{observed} is sitting way out there in the right-tail, while a large p-value indicates that T_{observed} is not large enough to raise suspicion against H_0. If the p-value is small, it means that according to H_0, it is fairly unlikely to observe a value of T such as T_{observed} (or more), but since such an instance has happened that means that there is very little evidence in support of H_0. On the other hand a large p-value indicates that it is not at all unlikely to observe a value of T such as T_{observed} (or more) according to H_0, and thus such values of T are only to be expected under H_0.

Again the point to be noted is that, even if looked from the point of view of strength or degree of evidence against H_0, p-value after all is a frequentist probability in the sense that it says, if H_0 were true how likely is it for T to exceed its observed value over repeated sampling.

2.4 Prediction

This is one area where the frequentist inference is at its worst. Some prediction problem involves estimation of a parametric function. Like for instance, in Example 2 if the company is interested in estimating its *average* sales for all those months in which it had spent Rs.10 lakhs on advertising, then according to the model this equals $\beta_0+10\beta_1$, which is a parametric function and the usual UMVU point estimation or CI interval estimation may be carried out within the frequentist logic as in §2.1 or §2.2. Likewise based on past observations one can attempt to forecast the mean or variance of a future observation according to some ARIMA times series model. But when it comes to predict or forecast a random variable itself frequentist methods lead to incoherency. Like in Example 2 if the company is interested in predicting or forecasting its sales for a certain month in which it is spending Rs.10 lakhs on advertising, mathematically we are interested in the random variable $Y|X = 10$. The methods that are provided for handling such situations in the frequentist set-up is at best ad-hoc and at worst wrong. Thus we shall not even bother to review such methods under the frequentist paradigm.

3 Why Bayesian?

In the previous section time and again it was emphasized that all the frequentist methods rely on sampling distribution which involve the behavior of a statistic over repeated sampling. But if one gives it a moment's thought it should be clear that it is sort of uncalled for. Given a set of data at hand we should possibly take all our decisions based on this data set alone, without bothering about what other data set *we could have observed* and base our decision on that. But frequentist inference dictates one to do just that. For example suppose you have to run some pathological test on a tumor removed from a patient. You can send them either to Lab-I or Lab-II, both of which are equally competent. Suppose you toss a coin and send it to Lab-II and get the results back. Common sense dictates that you should

possibly go by the results sent to you by Lab-II. But frequentist inference demands that you include (hypothetical) test results from Lab-I as well in your analysis! A couple of numerical examples will hopefully drive the point home.

Example 3: Consider the following population p.m.f. $p(y|\theta) = \begin{cases} \theta & \text{with probability } 0.5 \\ \theta+1 & \text{with probability } 0.5 \end{cases}$. That is the population random variable Y takes two possible values θ and $\theta+1$ with equal probability, where θ is the unknown parameter of interest. Now suppose we have 2 i.i.d. observations Y_1 and Y_2 on this $Y \sim p(y|\theta)$. Consider the interval estimate $\hat{\theta} = \begin{cases} \text{Minimum}\{Y_1, Y_2\} & \text{if } Y_1 \neq Y_2 \\ Y_1 & \text{if } Y_1 = Y_2 \end{cases}$. (Agreed that it is a degenerate interval of the form $[\hat{\theta}, \hat{\theta}]$, but it is still an interval estimate nonetheless.) Now consider the confidence level (coverage probability) of this interval estimate. It equals, $P(\hat{\theta} = \theta) = P(\text{Minimum}\{Y_1, Y_2\} = \theta | Y_1 \neq Y_2)P(Y_1 \neq Y_2) + P(Y_1 = \theta | Y_1 = Y_2)P(Y_1 = Y_2) = 1 \times 0.5 + 0.5 \times 0.5 = 0.75$. But what use does this 75% confidence have? When indeed $Y_1 \neq Y_2$ we are 100% certain that we have got the value of θ using $\hat{\theta}$, while we are only 50% certain about $\hat{\theta}$ when $Y_1 = Y_2$. It is true that over repeated use $\hat{\theta}$ will capture the value of θ 75% of the time, but we exactly know what our degree of confidence is for a given data set, and that is what what should be reported (100% or 50%) instead of the repeated measure 75%. $\bigtriangledown$

Example 4: Consider the following hypothesis testing problem where the task is to choose between $\theta = 0$ or $\theta = 1$ from a $\Theta = \{0, 1\}$. Suppose the observable Y is discrete taking values 1, 2 and 3 and its distribution characterized in terms of its p.m.f. $p(y|\theta)$ depends on θ as in the following table:

$y \rightarrow$	1	2	3
$p(y\|0)$	0.0050	0.0050	0.99
$p(y\|1)$	0.0051	0.9849	0.01

Now it can be shown that the most powerful test for $H_0 : \theta = 0$ versus $H_a : \theta = 1$ based on sample of size 1 is given by the $CR = \{1, 2\}$ *i.e.* one Rejects H_0 if the observed Y happens to be a 1 or 2 and does not Reject H_0 if the observed Y happens to be a 3. The probability of Type-I error for this test/decision rule/CR is 0.005+0.005=0.01, and the probability of Type-II error is also 0.01. But when the observed Y is 1, then really there is very little to choose between $\theta = 0$ or $\theta = 1$, while the most powerful test will recommend Rejecting H_0 with the false security of probability of committing either type of error of a small 1% associated with this decision rule. $\bigtriangledown$

Example 1 (Continued): Suppose in a store it is observed that 3 customers chose brand X of toothbrush while 9 did not. Based one this observation we ant to establish that π, the probability that a customer chooses toothbrush of brand X is more than 0.1. Since this is the point we wish to prove it goes in the alternative and we formulate this hypothesis testing problem as $H_0 : \pi \leq 0.1$ versus $H_a : \pi > 0.1$. Now a unique frequentist solution to this problem cannot be found unless more is stated about how these observations were obtained.

First consider a sampling scheme, where we observe the choice of toothbrush brand of 12 consumers among whom 3 happened to choose brand X and 9 did not. In this situation the underlying observable random variable of interest Y = Number of consumers choosing

brand X $\sim$ Binomial(12,π). Obviously larger the value of Y more is the evidence against H_0 and thus under this sampling scheme we shall compute our p-value as $P(B(12, 0.1) \geq 3) = \sum_{k=3}^{12} \left(\begin{array}{c} 12 \\ k \end{array} \right) 0.1^k 0.9^{12-k} = 0.1109$.

Now consider an alternative sampling scheme where we keep on observing the choice of toothbrush brand of consumers till we find 9 that did not choose brand X. Under this sampling scheme the same Y as above will now have Negative Binomial distribution with probability of success 1-π and number of successes one waits for = 9. Here again larger the value of Y more is the evidence against H_0. But now according to this sampling scheme the p-value equals $P(NB(9, 0.9) \geq 3) = \sum_{k=3}^{\infty} \left(\begin{array}{c} 8+k \\ k \end{array} \right) 0.1^k 0.9^9 = 0.0896$.

If one is working with $\alpha = 0.1$ one will take two opposite decisions (saying not enough evidence for $\pi > 0.1$ in the former and concluding that $\pi > 0.1$ in the later) under the two schemes yielding the same data. $\bigtriangledown$

Apart from the obvious draw-backs as pointed out in the above examples there are some serious philosophical problems with the frequentist reasoning of statistical inference. Even without getting into these philosophical discussions it would be worthwhile to point out a few interpretational difficulties with frequentist methods. If you have already faced difficulty in swallowing the arguments put forth in §2 for selling the methods (but not the logic, which is rational and clear) then welcome to the Bayesian club!

First let us look at the problem of **point estimation**. The main point there is we are uncertain about the value being provided by an estimator. If this uncertainty is summarized in terms of its sampling distribution, then it does not say anything about the uncertainty we are suffering with the sample at hand. It goes about addressing the issue in a round-about fashion about what to expect in repeated sampling in which we possibly hardly have any interest in. We could not care less about what would have happened in other hypothetical samples, while dealing with current uncertainties regarding the value provided by an estimator. A more direct approach of dealing with this uncertainty about the value of the parameter itself would clearly be more than welcome.

The confidence coefficient of a **confidence interval** is a misleading quantity. When faced with a statement like 95% confidence interval for μ is $[2.3, 5.1]$, most users mistakenly tend to interpret it as, there is a 95% chance that the true unknown value of μ will lie between 2.3 and 5.1. One can hardly blame an user for doing this. It is not as much a fault of the user as it is with the circuitous arguments that lead one to a confidence interval. Here again the definition is silent about the numerical interval we have at our hand regarding its degree of credibility in containing the value of the unknown parameter. But it gives an elusive probability like number which has nothing to do with the sample we have at hand, but as usual with what would have happened in other phantom samples. A method which allows the user an interpretation s/he intuitively understands (like the chance of the unknown μ lying between 2.3 and 5.1 is 0.95) is clearly far more desirable.

The logic of the topic which beginners in statistics possibly find most difficult to understand is **hypothesis testing**. Be its lopsided treatment of the null hypothesis, or the kind of

situation one exactly is in after taking a decision looking at the two error probabilities, or by looking at the p-value and subliminally attempting to interpret it as the chance of H_0 being true. The whole thing is basically a mess with counter-examples galore showing its pit-falls in its every nook and corner. What one really wants is a direct approach. Instead of a round-about p-value the user actually wants a direct probability of a hypothesis being true without any partial treatment meted out to one hypothesis or other. Moreover, in the same vein, many a times we are faced with not one (only the null), not two (a null and an alternative) but multiple hypotheses simultaneously for one to choose from. This is one major triumph of Bayesian statistics over the frequentist paradigm apart from its coherent and logical treatment of the **prediction** problem mentioned in §2.4.

4 The Posterior Distribution

The turning point of Bayesian statistics is the way in which it handles uncertainty. In the Bayesian paradigm, for drawing inference about an unknown population parameter $\boldsymbol{\theta}$ with the data $Y_1, Y_2, \ldots, Y_n$ i.i.d. $F(y|\boldsymbol{\theta})$, one starts with a prior distribution $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$, which is a probability distribution defined on the parameter space $\boldsymbol{\Theta}$. The idea behind the prior distribution is as follows. An experimenter collects data to gather information about the parameter $\boldsymbol{\theta}$ because s/he is uncertain about its value, otherwise there would have been no reason to collect any observation. However even before s/he starts collecting data, the experimenter though uncertain about the exact value of $\boldsymbol{\theta}$, might have some approximate idea about the kind of values $\boldsymbol{\theta}$ is expected to take. This approximate idea about the value of $\boldsymbol{\theta}$ and more importantly the uncertainty around it, before collecting data or conducting an experiment, is expressed in the prior distribution of $\boldsymbol{\theta}$. This is because the mathematical language of uncertainty is probability and thus there is no better way than expressing ones uncertainty about the value of an unknown parameter $\boldsymbol{\theta}$ in terms of a probability distribution on it, with its support spread over the parameter space $\boldsymbol{\Theta}$.

Once one agrees with the viewpoint expressed in the last sentence of the above paragraph regarding handling uncertainties, things become very smooth sailing. Of course, there is the initial hiccup of specifying the prior distribution and we shall delve into the issue further later in these notes. But once we can somehow express our initial uncertainty about $\boldsymbol{\theta}$ in terms of its prior distribution $\pi(\boldsymbol{\theta})$ before looking at the data, the way to update it, in light of the collected observations, has been explored at enormous depth and length in the literature. Starting from intuitive arguments to deep axiomatic developments, the answer to how to upgrade one's uncertainty about $\boldsymbol{\theta}$ expressed in terms of the prior $\pi(\boldsymbol{\theta})$, before collecting the data, to post data situation, is summarized in close to two and half century old Bayes' theorem. Thus we start our discussion with a quick review of Bayes' theorem in its elementary form.

Bayes' Theorem: Let $\Theta_1, \Theta_2, \ldots, \Theta_k$ be k mutually exclusive *i.e.* $\Theta_i \cap \Theta_j = \phi \; \forall i \neq j$, and exhaustive *i.e.* $\bigcup_{i=1}^{k} \Theta_i = \Theta$, the entire parameter space; states of nature, for which one has a prior distribution $(P(\Theta_1), P(\Theta_2), \ldots, P(\Theta_k))$ such that $\forall i = 1, 2, \ldots k$, $0 \leq P(\Theta_i) \leq 1$ and $\sum_{i=1}^{k} P(\Theta_i) = 1$. Then in light of the observed data Y the **posterior** probability of Θ_i is

given by:

$$P(\Theta_i|Y) = \frac{P(Y|\Theta_i)P(\Theta_i)}{\sum_{j=1}^{k} P(Y|\Theta_j)P(\Theta_j)} \quad \forall i = 1, 2, \ldots k \tag{2}$$

where the likelihood or "model" probabilities $P(Y|\Theta_i)$ for $i = 1, 2, \ldots k$ representing the probabilities of observing the data Y under each of the k possible states of nature, are assumed to be known.

Proof:

$$\begin{aligned}
& P(\Theta_i|Y) \\
&= \frac{P(\Theta_i \cap Y)}{P(Y)} && \text{because } P(A|B) = P(A \cap B)/P(B) \\
&= \frac{P(Y|\Theta_i)P(\Theta_i)}{P\left(Y \bigcap \left\{\bigcup_{j=1}^{k} \Theta_j\right\}\right)} && \text{because } P(A \cap B) = P(B|A)P(A) \text{ and } \Theta_j\text{'s are exhaustive} \\
&= \frac{P(Y|\Theta_i)P(\Theta_i)}{P\left(\bigcup_{j=1}^{k} \{Y \cap \Theta_j\}\right)} \\
&= \frac{P(Y|\Theta_i)P(\Theta_i)}{\sum_{j=1}^{k} P(Y \cap \Theta_j)} && \text{because } \Theta_j\text{'s are mutually exclusive} \\
&= \frac{P(Y|\Theta_i)P(\Theta_i)}{\sum_{j=1}^{k} P(Y|\Theta_j)P(\Theta_j)} && \text{because } P(A|B) = P(A \cap B)/P(B) \qquad \bigtriangledown
\end{aligned}$$

Thus according to Bayes' theorem one should upgrade one's uncertainty about $\boldsymbol{\theta}$, expressed in terms of the prior $\pi(\boldsymbol{\theta})$ before looking at the data, after observing the data $\boldsymbol{Y} = \boldsymbol{y}$ through the posterior distribution of $\boldsymbol{\theta}$ denoted by $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. Now the exact form of the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, though is essentially derived using the Bayes' theorem, depends on whether the parameter $\boldsymbol{\theta}$ and the observable random variable Y are discrete or continuous. For the time being, we shall also assume that given $\boldsymbol{\theta}$ we have n i.i.d. observations $Y_1, Y_2, \ldots, Y_n$ on the population random variable Y. The formulæ for $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ for the four different cases are as follows.

Case of discrete Y and discrete $\boldsymbol{\theta}$: As mentioned in footnote 5 in page 2, distributions of discrete random variables are most conveniently handled using their p.m.f.'s. Thus let the parameter space $\boldsymbol{\Theta}$ be discrete with $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots\}$. For $k = 1, 2, \ldots$ given $\boldsymbol{\theta} = \boldsymbol{\theta}_k$ let the population random variable Y be also discrete with the support $\mathcal{Y}^k = \left\{y_1^k, y_2^k, \ldots\right\}$ and respective probabilities $\left\{p_1^k, p_2^k, \ldots\right\}$ such that $\forall k = 1, 2, \ldots$ and $\forall j = 1, 2, \ldots$ $p_j^k \geq 0$ and $\sum_{j \geq 1} p_j^k = 1$ $\forall k = 1, 2, \ldots$. Now let the prior p.m.f. $\pi(\boldsymbol{\theta})$ be given by $\pi_k = P(\boldsymbol{\theta} = \boldsymbol{\theta}_k)$ and let $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$ be the observed data, denoted by $\boldsymbol{Y} = \boldsymbol{y}$, and given $\boldsymbol{\theta}$ all of which are i.i.d. Y. Then a routine use of (2) yields the posterior distribution of $\boldsymbol{\theta}$ as:

$$P(\boldsymbol{\theta} = \boldsymbol{\theta}_k|\boldsymbol{Y} = \boldsymbol{y}) = \frac{\pi_k \prod_{i=1}^{n} p_{\sum_{j \geq 1} j I[y_i = y_j^k]}^{k}}{\sum_{l \geq 1} \pi_l \prod_{i=1}^{n} p_{\sum_{j \geq 1} j I[y_i = y_j^k]}^{l}} \qquad \forall k \geq 1 \tag{3}$$

where $I[A]$ is the indicator function of an event or statement A, meaning that $I[A]$ is 1 if A is true and 0 otherwise. The term $\prod_{i=1}^{n} p_{\sum_{j \geq 1} j I[y_i = y_j^k]}^{k}$ might require some explanation.

The role of this term is same as that of $P(Y|\Theta_i)$ in (2). That is here we are to evaluate $P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|\boldsymbol{\theta} = \boldsymbol{\theta}_k)$. Since given $\boldsymbol{\theta}$, $Y_1, Y_2, \ldots Y_n$ are independent, $P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|\boldsymbol{\theta} = \boldsymbol{\theta}_k) = \prod_{i=1}^{n} P(Y_i = y_i|\boldsymbol{\theta} = \boldsymbol{\theta}_k)$. Now $P(Y_i = y_i|\boldsymbol{\theta} = \boldsymbol{\theta}_k)$ is going to be exactly one of $\left\{p_1^k, p_2^k, \ldots\right\}$. Actually it is that p_j^k such that $y_i = y_j^k$. $\sum_{j\geq 1} jI[y_i = y_j^k]$ precisely churns out that j. $\bigtriangledown$

Example 5: Let Y denote the number of logical bugs per thousand lines of codes written by a certain programmer. It is reasonable to assume that $Y \sim$Poisson(λ). The problem is to empirically determine the value of λ given some observations on Y. You believe *apriori* that λ for this programmer is 2, 2.5 or 3 with probabilities 0.2, 0.4 and 0.4 respectively. Now suppose you randomly select 2 different thousand line codes written by this programmer and upon examination you found that $Y_1 = 0$ and $Y_2 = 1$. The problem is to update your initial uncertainty about λ expressed as the prior p.m.f.

λ	2	2.5	3
$\pi(\lambda)$	0.2	0.4	0.4

in light of the observed data $Y_1 = 2$ and $Y_2 = 3$. This posterior computation is summarized in the following table

λ	$\pi(\lambda)$	$P(Y_1 = 0, Y_2 = 1\|\lambda)$	(2)×(3)	$\pi(\lambda\|\boldsymbol{Y} = \boldsymbol{y})$
(1)	(2)	(3)	(4)	(5)
2	0.2	0.0366	0.00732	0.4306
2.5	0.4	0.0168	0.00672	0.3953
3	0.4	0.0074	0.00296	0.1741
Total	1.0	-	0.01700	1.0

Column (5) is obtained as (4) divided by its total. Column (3) is obtained from Poisson probabilities, like for 5example $P(Y_1 = 0, Y_2 = 1|\lambda = 2) = \{e^{-2}\}\{(2/1!)e^{-2}\} = 0.1353\times 0.2707 = 0.0366$ etc.. Thus you update your belief about λ as in column (5) above from your initial belief of (0.2,0.4,0.4), after observing the data. Note that since in the data the only values of Y that were observed were 0 and 1, that tilts the scale substantially towards the smallest λ value considered *apriori viz.* 2, even with the *apriori* belief that 2 is least likely of the values considered for λ. $\bigtriangledown$

Case of discrete Y and continuous $\boldsymbol{\theta}$: Distribution of continuous random variables are most conveniently handled using their probability density functions or p.d.f.[9]. Thus let the parameter vector $\boldsymbol{\theta}$ be continuous taking values in the parameter space $\boldsymbol{\Theta}$. The prior distribution in this case will be specified in terms of a p-dimensional (see footnote 3 in page 2) joint p.d.f. of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$, say $\pi(\boldsymbol{\theta})$. Given $\boldsymbol{\theta}$, let the population random variable Y be discrete with the support $\mathcal{Y}^{\boldsymbol{\theta}} = \left\{y_1^{\boldsymbol{\theta}}, y_2^{\boldsymbol{\theta}}, \ldots\right\}$ and respective probabilities $\left\{p_1^{\boldsymbol{\theta}}, p_2^{\boldsymbol{\theta}}, \ldots\right\}$ such that $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\forall j = 1, 2, \ldots$ $p_j^{\boldsymbol{\theta}} \geq 0$ and $\sum_{j\geq 1} p_j^{\boldsymbol{\theta}} = 1$ $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let $\boldsymbol{Y} = \boldsymbol{y}$ be the

[9] A random variable Y is said to be continuous if its c.d.f. $F(y|\boldsymbol{\theta})$ is a continuous function of y. Barring a few pathological cases the c.d.f. $F(y|\boldsymbol{\theta})$ of a continuous random variable usually admits a first derivative w.r.t. y. This derivative $\frac{d}{dy}F(y|\boldsymbol{\theta})$ is called the p.d.f. of Y which is denoted by $f(y|\boldsymbol{\theta})$. It is called a probability density because at a given point y, it gives the amount of probability the random variable Y gives at a neighborhood of y per unit length of that neighborhood as the length of the neighborhood shrinks to 0. This interpretation directly follows from the definitions of derivative and c.d.f.. Furthermore given the p.d.f. $f(y|\boldsymbol{\theta})$ the c.d.f. can be obtained as $\int_{-\infty}^{y} f(t|\boldsymbol{\theta})\, dt$ and more generally for any set A, $P(Y \in A) = \int_A f(y|\boldsymbol{\theta})\, dy$.

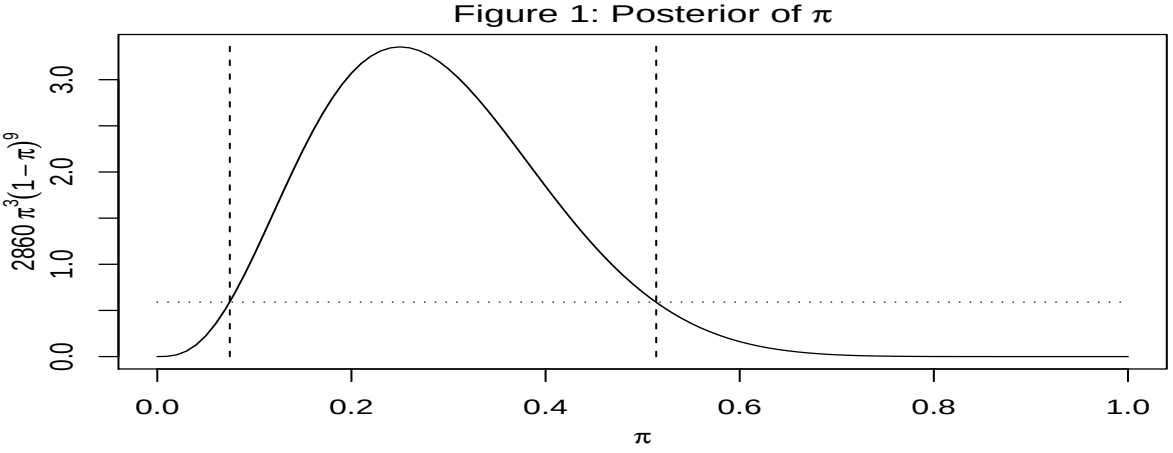
observed data, and given $\boldsymbol{\theta}$ let $Y_1, Y_2, \ldots Y_n$ be i.i.d. Y. Then the posterior of $\boldsymbol{\theta}$ given the observed data $\boldsymbol{Y} = \boldsymbol{y}$ is found as the conditional density of $\boldsymbol{\theta}$ given $\boldsymbol{Y} = \boldsymbol{y}$, which is denoted by $\pi(\boldsymbol{\theta}|\boldsymbol{Y} = \boldsymbol{y})$. The formula for $\pi(\boldsymbol{\theta}|\boldsymbol{Y} = \boldsymbol{y})$ is as follows:

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y} = \boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta}) \prod_{i=1}^{n} p^{\boldsymbol{\theta}}_{\sum_{j\geq 1} jI[y_i = y_j^k]}}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\phi}) \prod_{i=1}^{n} p^{\boldsymbol{\phi}}_{\sum_{j\geq 1} jI[y_i = y_j^k]}\, d\boldsymbol{\phi}} \qquad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{4}$$

Note that equation (4) is same as equation (3) except that now the summation has been replaced by the integral. The proof of the above result is same as that of the Bayes' theorem. The numerator of (4) is nothing but the joint density (used in a slightly broader context) of $(\boldsymbol{\theta}, \boldsymbol{Y})$ $g(\boldsymbol{\theta}, \boldsymbol{y})$, which is given by $P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, which equals $\pi(\boldsymbol{\theta}) \prod_{i=1}^{n} p^{\boldsymbol{\theta}}_{\sum_{j\geq 1} jI[y_i = y_j^k]}$, following the same logic as in the previous case. The denominator of (4) is $P(\boldsymbol{Y} = \boldsymbol{y})$ which is obtained by integrating $\boldsymbol{\theta}$ out over its domain $\boldsymbol{\Theta}$ from $g(\boldsymbol{\theta}, \boldsymbol{y})$, which is same as the numerator of (4), and thus the result follows. $\triangledown$

Example 1 (Continued): Suppose we are interested in the probability π of a consumer choosing toothbrush of brand X. For this we observe the brand choice of 12 consumers and the choice is coded using the random variable Y of Example 1. Thus we have 12 0-1 valued $Y_1, Y_2, \ldots, Y_{12}$ which are i.i.d. with p.m.f. $\pi^y(1-\pi)^{1-y}$ for $y = 0, 1$. Now suppose (as in Example 1 (Continued) in page 10) we find 3 Y_i's to be 1 and the remaining 9 0's (it does not matter for which i's because of the following). Then $\prod_{i=1}^{12} p^{\pi}_{\sum_{j\geq 1} jI[y_i = y_j^k]} = \pi^3(1-\pi)^9$. Now if we say that before looking at the data we had no idea about the kind of value π will have, then a flat Uniform[0,1] prior for π might be quite appropriate for modeling this ignorance *i.e.* let us take $\pi(\pi) = \begin{cases} 1 & \text{if } 0 \leq \pi \leq 1 \\ 0 & \text{otherwise} \end{cases}$ as the prior of π. Then in order to get the exact expression for the posterior density $\pi(\pi|\boldsymbol{Y} = \boldsymbol{y})$ of π, by (4), all we have to do is find $\int_0^1 \pi^3(1-\pi)^9\, d\pi$ for the denominator. Using the so-called β-integrals it may be shown that $\int_0^1 \pi^3(1-\pi)^9\, d\pi = \frac{3!9!}{13!} = 3.4965 \times 10^{-4}$, so that the exact expression of the posterior density of π based on the observed data becomes $\pi(\pi|\boldsymbol{Y} = \boldsymbol{y}) = \begin{cases} 2860\pi^3(1-\pi)^9 & \text{if } 0 \leq \pi \leq 1 \\ 0 & \text{otherwise} \end{cases}$. This posterior density of π is plotted in Figure 1 below: s

Figure 1: Posterior of π



Now essentially all the questions pertaining to π will be answered in terms of this posterior density, which has all the uncertainties regarding the value of π directly packed into this distribution given the observed data. Like for instance (as in Example 1 (Continued) in page 10) if one wishes to obtain the probability of the hypothesis $\pi > 0.1$, unlike the frequentist case, now one can say that the chance of this is $\int_{0.1}^{1} 2860\pi^3(1-\pi)^9 \, d\pi$, which by numerical integration equals 0.9664. But be careful before jumping into the conclusion that then the data is saying that almost certainly $\pi > 0.1$. This is because *apriori* we are assuming that there is a 90% chance of $\pi > 0.1$ even before observing the data and this prior value has to be suitably taken into account before concluding the truth about any hypothesis. Bayesian hypothesis testing precisely does that, which will be taken up in the next section.

Before closing the discussion on this example, we shall provide one more use of the posterior distribution to emphasize its key role and importance in Bayesian inference. According to the posterior in Figure 1, a 95% Bayesian interval estimate of π is given by $[0.0747, 0.5140]$. This interval has been indicated by the pair of vertical dashed lines in Figure 1. Unlike the frequentist case, now this interval estimate has the direct interpretation that, given the data, there is a 95% chance of π falling between the two numbers 0.0747 and 0.5140. $\bigtriangledown$

Case of continuous Y and discrete $\boldsymbol{\theta}$: Let $\boldsymbol{\theta}$ be discrete taking values in $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots\}$ and given $\boldsymbol{\theta} = \boldsymbol{\theta}_k$ let the population random variable Y have the p.d.f. $f(y|\boldsymbol{\theta}_k)$ with support $\mathcal{Y}^k$. Now as usual let the prior on $\boldsymbol{\theta}$ be given by $\pi_k = P(\boldsymbol{\theta} = \boldsymbol{\theta}_k)$ and $Y_1, Y_2, \ldots, Y_n$ be i.i.d. Y. Then given the data $\boldsymbol{Y} = \boldsymbol{y}$ the posterior distribution of $\boldsymbol{\theta}$ is given by:

$$P(\boldsymbol{\theta} = \boldsymbol{\theta}_k | \boldsymbol{Y} = \boldsymbol{y}) = \frac{\pi_k \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}_k)}{\sum_{l \geq 1} \pi_l \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}_l)} \qquad \forall k \geq 1 \tag{5}$$

Again the proof of (5) follows by imitating the proof of Bayes' Theorem. The numerical steps involved in computing the posterior in this case is essentially exactly same as that of the first one *viz.* discrete Y and discrete $\boldsymbol{\theta}$, except that now for computing the column (3) as in Example 5, one would use $\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}_k)$, instead of the earlier $\prod_{i=1}^{n} P(Y_i = y_i|\boldsymbol{\theta}_k)$. Since computationally these two cases are almost identical we shall skip the numerical example part for this case and move on to the last case. $\bigtriangledown$

Case of continuous Y and continuous $\boldsymbol{\theta}$: For most practical applications one encounters this and the second case *viz.* discrete Y and continuous $\boldsymbol{\theta}$, because it is fairly rare to have a discrete parameter space and thus the first and the third cases are essentially included here for illustrative purposes. However for parameters which enter the model in a very complicated form, sometimes one puts a discrete prior on it to make things tractable. Coming back to the case of interest, let $Y_1, Y_2, \ldots Y_n$ be i.i.d. Y where given $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, Y is assumed to have the p.d.f. $f(y|\boldsymbol{\theta})$ with support $\mathcal{Y}^{\boldsymbol{\theta}}$. Let $\pi(\boldsymbol{\theta})$ be the prior on $\boldsymbol{\theta}$, which is a joint density of $(\theta_1, \theta_2, \ldots, \theta_p)'$ with support $\boldsymbol{\Theta}$. Given observations $\boldsymbol{Y} = \boldsymbol{y}$, let $\pi(\boldsymbol{\theta}|\boldsymbol{Y} = \boldsymbol{y})$ denote the posterior density of $\boldsymbol{\theta}$, which is nothing but the conditional density of $\boldsymbol{\theta}$ given $\boldsymbol{Y} = \boldsymbol{y}$. The formula for this conditional density is as follows:

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y} = \boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta}) \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\phi}) \prod_{i=1}^{n} f(y_i|\boldsymbol{\phi}) \, d\boldsymbol{\phi}} \qquad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{6}$$

Since equation (6) together with (4) will be heavily used in the sequel we first present a formal proof of (6), which is extremely straight-forward and is just a continuous analogue of

the proof of the Bayes' Theorem.

$$
\begin{aligned}
&\pi(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y}) \\
&\quad = \quad g(\boldsymbol{\theta},\boldsymbol{y})/m(\boldsymbol{y}) \qquad \left(\begin{array}{l} \text{where } g(\boldsymbol{\theta},\boldsymbol{y}) \text{ is the joint density of } (\boldsymbol{\theta},\boldsymbol{Y}) \text{ and } m(\boldsymbol{y}) \\ \text{is the marginal density of } \boldsymbol{Y}. \end{array} \right) \\
&\quad = \quad \frac{\pi(\boldsymbol{\theta}) \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} g(\boldsymbol{\phi},\boldsymbol{y})\, d\boldsymbol{\phi}} \qquad \left(\begin{array}{l} \text{because the joint density of } (\boldsymbol{\theta},\boldsymbol{Y}) \text{ is the product of} \\ \text{the conditional density of } \boldsymbol{Y}|\boldsymbol{\theta}, \text{ given by } \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) \\ \text{(since } Y_1, Y_2, \ldots Y_n \text{ are i.i.d. } f(y|\boldsymbol{\theta})\text{) and the marginal} \\ \text{density of } \boldsymbol{\theta}, \text{ given by } \pi(\boldsymbol{\theta}); \text{ and the marginal density} \\ m(\boldsymbol{y}) \text{ of } \boldsymbol{Y} \text{ is obtained by integrating } \boldsymbol{\theta} \text{ out of } g(\boldsymbol{\theta},\boldsymbol{y}), \\ \text{the joint density of } (\boldsymbol{\theta},\boldsymbol{Y}). \end{array} \right) \\
&\quad = \quad \frac{\pi(\boldsymbol{\theta}) \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\phi}) \prod_{i=1}^{n} f(y_i|\boldsymbol{\phi})\, d\boldsymbol{\phi}} \qquad \left(\text{because as explained above, } g(\boldsymbol{\phi},\boldsymbol{y}) = \pi(\boldsymbol{\phi}) \prod_{i=1}^{n} f(y_i|\boldsymbol{\phi}) \,. \right)
\end{aligned}
$$

proving (6). ▽

Example 6: Let $Y_1, Y_2, \ldots Y_n$ be i.i.d. $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Let $\boldsymbol{\theta} = (\mu, \sigma^2)$. So far we have only been dealing with proper priors such that $\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})\ d\boldsymbol{\theta} = 1$ (or for discrete $\boldsymbol{\Theta}$, $\sum_{k\geq 1} P(\boldsymbol{\theta} = \boldsymbol{\theta}_k) = \sum_{k\geq 1} \pi_k = 1$). Such priors are also sometimes called informative priors, which essentially contain an experimenter's subjective gut-feeling about the unknown $\boldsymbol{\theta}$. While this is fine and sometimes desirable (the frequentist paradigm does not allow one to do this and since in many occasions important prior information may be available, among many others, this argument is also put forth to advocate the Bayesian methods), there are situations where the experimenter may have no idea whatsoever about the unknown value of $\boldsymbol{\theta}$. In such situations instead of falling back upon frequentist ideas, the Bayesian solution is to use non-informative priors. Such non-informative priors are also called vague or improper or diffuse or reference and more recently default priors. Different kinds of arguments are given for appropriately choosing a default prior in a given situation and we shall briefly outline them in a later section. But for the $N(\mu, \sigma^2)$ model, it is now fairly well-established that the default prior for (μ, σ) should be

$$\pi(\mu, \sigma) \propto \frac{1}{\sigma} \qquad \text{for } -\infty < \mu < \infty \text{ and } \sigma > 0. \tag{7}$$

Note that the prior specified in (7) implies

1. $\pi(\mu, \sigma)$ is improper in the sense that it is not a legitimate density function, since its integral is ∞.
2. μ and σ are assumed to be independent *apriori*, which in turn implies any function of σ such as the variance σ^2 or $\tau = 1/\sigma^2$, also called the precision parameter are also independent of μ.
3. Prior of μ is flat over the entire real line, which is intuitively very appealing as a non-informative prior for μ.
4. Prior of $\log(\sigma)$ is flat over the entire real line, yielding the prior on σ as $1/\sigma$ for $\sigma > 0$, using the change of variable formula.
5. Using the change of variable formula, the prior on (μ, σ^2), $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$ for $-\infty < \mu < \infty$ and $\sigma^2 > 0$; and the prior on (μ, τ), $\pi(\mu, \tau) \propto \frac{1}{\tau}$ for $-\infty < \mu < \infty$ and $\tau > 0$.

For convenience, we shall work with the reparameterized version (μ, τ) called the (mean, precision) instead of the original (mean,variance) or (mean, standard deviation) parameterization. Thus though our task is to draw inference on (μ, σ^2), we shall do so through the posterior p.d.f.'s of (μ, τ), which is slightly easier in terms of their resemblance with the standard probability distributions. Thus let us begin deriving the joint posterior p.d.f. of (μ, τ), denoted by $\pi(\mu, \tau | \boldsymbol{Y} = \boldsymbol{y})$.

Digression: Since it is a case of continuous Y and continuous $\boldsymbol{\theta}$ the formula used for the posterior computation is that given in (6). Now we shall employ a tactic which is routine in Bayesian posterior calculation. In all the four posterior formulæ given in (3), (4), (5) and (6) the major hurdle in posterior computation is the denominator. But what is the role of the denominator in these formulæ? A little closer examination reveals that the main thing of interest, namely the form of the posterior as a function of $\boldsymbol{\theta}$, is determined by the numerator. But the numerator by itself is not a legitimate p.m.f. or a p.d.f. because it does not add or integrate to 1. The role of the denominator is to just do that. That is the denominator is chosen in such a manner that the r.h.s. of all these formulæ add or integrate to 1. That is the only role the denominator plays in the posterior computation is that of a normalizing constant, which is free of the parameter of interest $\boldsymbol{\theta}$, such that the sum or integral of the function of $\boldsymbol{\theta}$ in the numerator equals the denominator so that the r.h.s. in its entirety becomes a proper p.m.f. or a p.d.f..

Thus an explicit determination of the denominator, which is a constant free of $\boldsymbol{\theta}$, is seldom carried out at the outset for determining the posterior. The form of the posterior as function of $\boldsymbol{\theta}$ is first studied by writing down the numerator, and very often it is the case that the form reveals semblance with some known distributions, which are then used to figure out the normalizing constant if required at all. In the process since the denominator is ignored, instead of "=" one uses "$\propto$" for the posterior and the numerator. At this juncture a couple of words about the numerator is also in order. The numerators in all the four cases in (3), (4), (5) and (6) involve two terms - one involves the prior and the other involves the p.m.f. or p.d.f. of $\boldsymbol{Y}$ evaluated at $\boldsymbol{Y} = \boldsymbol{y}$ *e.g.* $\prod_{i=1}^{n} p^{k}_{\sum_{j\geq 1} jI[y_i = y_j^k]}$ in (3) and (4) and $\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$ in (5) and (6). This is called the **likelihood function** of $\boldsymbol{\theta}$ given the data $\boldsymbol{Y} = \boldsymbol{y}$ and is denoted by $L(\boldsymbol{\theta}|\boldsymbol{y})$[10]. Now with the introduction of the likelihood function and elimination of the denominator, we have a unified way of writing the formula for the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ as

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\boldsymbol{y}) \tag{8}$$

Now it is possible to give a further succinct formula for the posterior density than (8) in many cases where the raw data $\boldsymbol{y}$ can be reduced in dimension in terms of what are called **sufficient** statistics. This reduction technique is very important because it provides us an easy handle on analytically tackling the posterior, which otherwise becomes extremely messy in terms of the raw data $\boldsymbol{y}$. Presenting this technique at this juncture will lead to a long digression. Thus this technique has been deferred to Appendix A, and has been referred to in the text, whenever required. $\bigtriangledown$

[10] A formal definition and interpretation of the **likelihood function** are provided in Definition A2 and the paragraph following it, in Appendix A in page 40.

Coming back to the example, here since $Y_1, Y_2, \ldots Y_n$ i.i.d. with p.d.f. $f(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\} = (2\pi)^{-1/2}\tau^{1/2}\exp\left\{-\frac{1}{2}\tau\left(y-\mu\right)^2\right\} = f(y|\mu,\tau)$ (say), given the observations $\boldsymbol{Y} = \boldsymbol{y}$, the likelihood

$$L\left(\mu,\tau|\boldsymbol{y}\right) \propto \tau^{n/2}\exp\left\{-\frac{1}{2}\tau\left[\nu s_{n-1}^2 + n\left(\overline{y}-\mu\right)^2\right]\right\} \tag{9}$$

where $\nu = n-1$, $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \overline{y})^2$. Note that the same likelihood function has also been obtained in (A1) in Appendix A in terms of (μ,σ^2). The term in the square bracket in (9) is obtained by expanding the square in $\sum_{i=1}^n \left(y_i-\mu\right)^2 = \sum_{i=1}^n \left\{(y_i-\overline{y}) + (\overline{y}-\mu)\right\}^2$ and observing that $\sum_{i=1}^n \left(y_i-\overline{y}\right) = 0$. Following (8) we get the form of the posterior $\pi(\mu,\tau|\boldsymbol{Y}=\boldsymbol{y})$ by multiplying the likelihood (9) by the prior of (μ,τ) mentioned in point 5 following the discussion of the prior on (μ,σ) given in (7). Thus,

$$\pi\left(\mu,\tau|\boldsymbol{Y}=\boldsymbol{y}\right) \propto \tau^{n/2-1}\exp\left\{-\frac{1}{2}\tau\left[\nu s_{n-1}^2 + n\left(\overline{y}-\mu\right)^2\right]\right\} \text{ for } -\infty < \mu < \infty \text{ and } \tau > 0. \tag{10}$$

Equation (10) gives the expression for the bivariate joint posterior p.d.f. of (μ,τ) given the data $\boldsymbol{Y} = \boldsymbol{y}$. It is not normalized but no matter what the normalizing constant might be its shape is completely determined by the r.h.s. of (10). A three-dimensional plot of this surface and its respective contours are provided in Figures 2 and 3 below for a hypothetical data set with $n = 10$, $\overline{y} = 0$ and $s_{n-1}^2 = 1$.

Figure 2: Joint p.d.f.

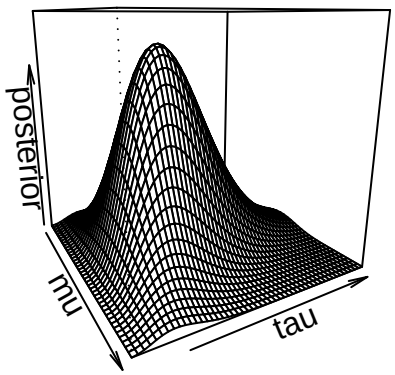
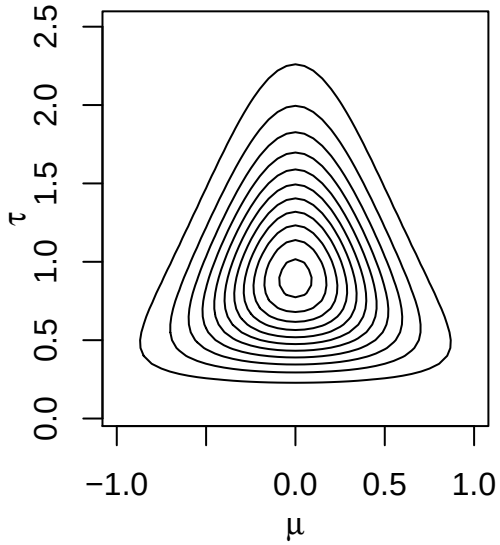


Figure 3: Contour Plot



Though these plots might be useful in getting some idea about the joint behavior of the two parameters, inference on individual parameters are typically based on their marginal posteriors. We shall first derive the marginal posterior of τ.

In order to find the marginal posterior of τ one needs to integrate μ out from the r.h.s. of (10). Note that the term involving μ in the r.h.s. of (10) equals $\exp\left\{-(n\tau/2)\left(\overline{y}-\mu\right)^2\right\}$. With μ as the variable and $\overline{y}$ as a fixed constant, which it is, given the data at hand,

this term is easily recognized as the p.d.f. of a Normal distribution with mean $\overline{y}$ and variance $1/(n\tau)$.[11] Thus using the fact that integral of a Normal p.d.f. equals 1 *i.e.* $\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\,dx=1$, we get that $\int_{-\infty}^{\infty}\exp\left\{-(n\tau/2)(\overline{y}-\mu)^2\right\}\,d\mu=(2\pi/n)^{1/2}\tau^{-1/2}$, and therefore the marginal posterior p.d.f. of τ is given by

$$\pi(\tau|\boldsymbol{Y}=\boldsymbol{y})\propto\tau^{\nu/2-1}\exp\left(-\tau\nu s_{n-1}^2/2\right)\qquad\text{for }\tau>0. \tag{11}$$

The density given in (11) is immediately recognized as a Gamma density with $\alpha=\nu/2$ and $\lambda=\nu s_{n-1}^2/2$ or equivalently since Gamma($\alpha=\nu/2$,$\lambda=1/2$) is same as χ_ν^2, a quick change of variable $\tau\longleftrightarrow\tau\nu s_{n-1}^2$ in (11) yields the important and interesting result that the marginal posterior of $\nu s_{n-1}^2/\sigma^2\sim\chi_\nu^2$. Of course here given the data, s_{n-1}^2 is a fixed constant and σ^2 is the random quantity the distribution of a function of which is χ^2 with ν d.f.. But it is worth noticing that in the frequentist theory we have the same result that $\nu s_{n-1}^2/\sigma^2\sim\chi_\nu^2$ as the sampling distribution of the sample variance from a Normal population, which is subsequently used for inference purpose like confidence interval for σ^2. The formula for a 100(1-α)% Bayesian interval estimate using the posterior distribution of σ^2 is thus going to be identical to that of the 100(1-α)% CI of σ^2 *viz.* $\left[\frac{\nu s_{n-1}^2}{\chi^2_{\nu,1-\alpha/2}},\frac{\nu s_{n-1}^2}{\chi^2_{\nu,\alpha/2}}\right]$, where $\chi^2_{\nu,\alpha}$ is the α-th quantile of χ^2 distribution with ν d.f., but as is also mentioned in footnote 11, the interpretation of the obtained interval here is more direct. If the 95% interval estimate for σ^2 equals $[0.8,1.4]$ now we can say that given the data, there is a 95% chance of σ^2 falling between 0.8 and 1.4, without any reference to other data sets or repeated use of the formula.

Now let us turn our attention to the marginal posterior of μ. Just as the marginal posterior of τ was obtained by integrating μ out of the r.h.s. of (10), likewise the marginal posterior of μ will be obtained by integrating τ out of the r.h.s. of (10). The r.h.s. of (10) as a function of τ is again immediately recognized as a Gamma density with $\alpha=n/2$ and $\lambda=\left[\nu s_{n-1}^2+n(\overline{y}-\mu)^2\right]$. Now since the integral of the p.d.f. of the Gamma(α,λ) distribution equals 1 *i.e.* $\frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}e^{-\lambda x}\,dx=1$, we get that $\int_0^\infty\tau^{n/2-1}\exp\left\{-\frac{1}{2}\tau\left[\nu s_{n-1}^2+n(\overline{y}-\mu)^2\right]\right\}\,d\tau$ $=\Gamma(n/2)\Big/\left[\nu s_{n-1}^2+n(\overline{y}-\mu)^2\right]^{n/2}$. Thus now, ignoring the constants, the marginal posterior p.d.f. of μ can be written as

$$\pi(\mu|\boldsymbol{Y}=\boldsymbol{y})\propto 1\Big/\left[1+\frac{1}{\nu}\left(\frac{\mu-\overline{y}}{s_{n-1}/\sqrt{n}}\right)^2\right]^{(\nu+1)/2}\qquad\text{for }-\infty<\mu<\infty \tag{12}$$

Since a t-distribution with ν d.f. has the p.d.f. $\propto(1+t^2/\nu)^{-(\nu+1)/2}$, the r.h.s of (12) is immediately recognized as t_ν density. Thus the marginal posterior of μ is such that $\frac{\mu-\overline{y}}{s_{n-1}/\sqrt{n}}\sim t_\nu$ which can now be used for inference purposes. Again it is worthwhile to draw parallel with the corresponding frequentist result. For unknown σ^2, in the frequentist set-up inference about μ is drawn using the fact that $\frac{\overline{y}-\mu}{s_{n-1}/\sqrt{n}}$ has a t_ν distribution. In the

[11] At this point it is worth mentioning that if σ^2 and thus τ were known, inference about μ would have proceeded with its posterior distribution which in this case would have been $\propto\exp\left\{-(n\tau/2)(\overline{y}-\mu)^2\right\}$. Thus the posterior of μ would have been $N(\overline{y},\sigma^2/n)$ and a 100(1-α)% interval estimate for μ using this posterior distribution would have been same as the one provided in (1) *viz.* $\overline{y}\pm z_{1-\alpha/2}\sigma/\sqrt{n}$. But the logic of this interval estimate, with the same formula, here is completely different, with a direct interpretation.

frequentist set-up the statistics $\overline{y}$ and s_{n-1} are the ones which are random and thus this t distribution refers to a sampling distribution. But here the t-distribution refers to the posterior distribution of a function of the parameter μ, where μ is the random variable and given the data at hand, the statistics $\overline{y}$ and s_{n-1} are non-random known constants, as they should be. Thus though one will get for example an identical interval estimate of μ for a given set of data, its Bayesian interpretation is the one with which most people will feel most comfortable with. $\bigtriangledown$

In this section we have discussed the notion of the posterior distribution in detail. At this point it is worthwhile to pause for a moment and look ahead to see where we are heading. In the Bayesian paradigm, the posterior distribution plays a pivotal role, in the sense that it is argued that whatever an observed set of data has to say about the unknown parameters, all this information is packed into the posterior distribution and one really does not need to know anything else. However there are specific inferential problems like estimation, hypothesis testing and prediction which needs to be solved, and how one can use the posterior distribution in addressing these issues is yet to be discussed. We take up these issues in section 6 after a quick introduction to statistical decision theory in the following section. Then there is this thorny issue about the choice of the prior. While much is available in the literature, here we shall very lightly touch upon the topic in §7.

Finally there are serious computational issues involving numerical calculation of the posterior distribution. Essentially the posterior computation requires numerical calculation of typically high-dimensional integrals. Recent years have witnessed an explosion of development and application of such numerical methods, broadly coming under the umbrella of what are called Markov Chain Monte Carlo or MCMC methods. These methods bring with themselves their own problems like their implementations and convergence issues. We shall take these up as a subject in itself, for which there will be a separate set of small lecture notes and other concise but very well-written study materials.

5 Statistical Decision Theory

Most of the problems statisticians indulge in like estimation, hypothesis testing, model selection, prediction etc. can be mathematically formulated as one of taking a decision in the face of uncertainty. In this section we shall formulate and provide (Bayesian) solutions to these statistical inferential problems as one of Decision Theory after providing a brief overview and some examples of the elements of Statistical Decision Theory.

A decision problem is formulated in terms of the triplet $(\Theta, \mathcal{A}, L)$, where Θ denotes the set consisting of all possible "states of nature" (but which one is typically unknown), $\mathcal{A}$ denotes the set of all possible actions or decisions that one might take for the decision problem at hand, and L is a real valued function with a finite lower bound defined on the domain $\boldsymbol{\theta} \times \mathcal{A}$ with the interpretation that for $\theta \in \Theta$ and $a \in \mathcal{A}$, $L(\theta, a)$ denotes the amount of loss one will incur if one takes action a when the state of nature is θ.

Example 7: Suppose one is considering investing Rs.1000 for one year either in the stock market or by depositing it in the post-office for a fixed one-year term deposit at an interest rate of 6% per annum. The stock market will yield a net of either 20% gain or 10% loss on investments after one year. The decision that needs to be taken is whether to invest that Rs.1000 in stock market or post-office. This problem in the above notation can be formulated as follows. Here there are two unknown states of nature *viz.* the stock market will yield a gain, say θ_1, or it might net a loss, say θ_2 (the return from the term-deposit in post-office is guaranteed, thus it need not figure in the unknown states of nature). Let $\Theta = \{\theta_1, \theta_2\}$. Now there are two possible actions *viz.* action a_1: invest in the stock-market or action a_2: invest in post-office. Thus let $\mathcal{A} = \{a_1, a_2\}$. Now we need to figure out the loss function for each possible (θ_1, a_j) pair for $i, j = 1, 2$. This loss function $L(\theta, a)$ is given in the following table:

$a \rightarrow$ $\theta \downarrow$	a_1	a_2
θ_1	-200	-60
θ_2	100	-60

We shall get back to the issue of taking the optimal decision shortly after providing a couple more examples on decision problem formulation in terms of the triplet $(\Theta, \mathcal{A}, L)$. $\bigtriangledown$

Example 8: Suppose before getting out of your home in the morning on a certain day you are to decide whether you should carry your umbrella along with you or not. You envisage that getting wet is twice as inconvenient as needlessly carrying an umbrella around. For this problem, $\Theta = \{\theta_1, \theta_2\}$, where θ_1 denotes the state of nature that it will rain and θ_2 denotes that it will not rain; $\mathcal{A} = \{a_1, a_2\}$, where a_1 denotes the decision, carry the umbrella and a_2 denotes, do not carry it; and the loss function $L(\theta, a)$ is given in the following table: $\bigtriangledown$

$a \rightarrow$ $\theta \downarrow$	a_1	a_2
θ_1	0	2
θ_2	1	0

Example 9: Suppose a pharmaceutical company is considering launching a new pain-reliever in the market. But before doing so, it needs to get an idea about the proportion of the market this new pain-reliever is going to capture. Let θ denote the actual proportion of the market that this new pain-reliever is going to capture and let a denote the company's estimate of this θ. Thus here $\Theta = \mathcal{A} = [0, 1]$. The company elicits that the loss will be 1.5 times more for over-estimation in terms of cost of production, unsold inventory etc. than under-estimation, in which case the only loss is in terms of missed opportunity. Thus the company puts forth the loss function $L(\theta, a) = \begin{cases} (\theta - a) & \text{if } \theta \geq a \\ 1.5(a - \theta) & \text{if } \theta < a \end{cases}$. $\bigtriangledown$

Now how does one take the optimal decision? There are essentially two approaches. The frequentist approach called the Minimax approach advocates taking that decision which minimizes the maximum loss, calculated over all possible states of nature. Mathematically the minimax decision rule is that action which equals Arg. $\text{Min}_{a \in \mathcal{A}} \text{Maximum}_{\theta \in \Theta} L(\theta, a)$. In

Example 7, the maximum losses one can encounter for a_1 and a_2 are 100 and -60 respectively. Of these two since the minimum occurs for a_2 the minimax decision for Example 7 would be to invest Rs.1000 in the post-office. Like wise the minimax decision in Example 8 would be to carry an umbrella, which is easy to see, and the minimax decision in Example 9 would be $a = 0.4$. For Example 9, for a fixed $a \in \mathcal{A} = [0, 1]$, a simple geometric plot of $L(\theta, a)$ against $\theta \in \Theta = [0, 1]$ shows that the maximum loss is Maximum$\{1.5a, 1 - a\}$ which after a small step of algebra can be shown $= \begin{cases} 1 - a & \text{if } a \leq 0.4 \\ 1.5a & \text{if } a > 0.4 \end{cases}$, and this maximum loss is minimized when $a = 0.4$.

Bayesians always deal with uncertainties by modeling it using a probability distribution. In a decision problem characterized by $(\Theta, \mathcal{A}, L)$ one is uncertain about the state of nature θ which takes values in Θ. Thus it is only but natural to model one's uncertainty regarding θ using a probability distribution, say $\pi^*(\theta)$, on θ with support Θ and then chose that action $a \in \mathcal{A}$ which minimizes the *Bayesian Expected Loss* given by $E_{\pi^*(\theta)}[L(\theta, a)]$, where $E[\cdot]$ denotes the mean or expectation operator. $E[\cdot]$ is subscripted with $\pi^*(\theta)$ to indicate the fact that the expectation of $L(\theta, a)$ is taken over θ using its distribution $\pi^*(\theta)$, and thus depends on $\pi^*(\theta)$ for every fixed $a \in \mathcal{A}$. Thus minimizing the Bayesian Expected Loss criterion calls for calculating $E_{\pi^*(\theta)}[L(\theta, a)]$ for every fixed $a \in \mathcal{A}$, once the uncertainty about the value of θ has been modeled using $\pi^*(\theta)$, and then it suggests taking that decision a which has the smallest Bayesian Expected Loss. Mathematically this decision is same as Arg. Min$_{a \in \mathcal{A}} E_{\pi^*(\theta)}[L(\theta, a)]$. This is the Bayesian solution. From now on by expected loss we shall mean Bayesian Expected Loss.

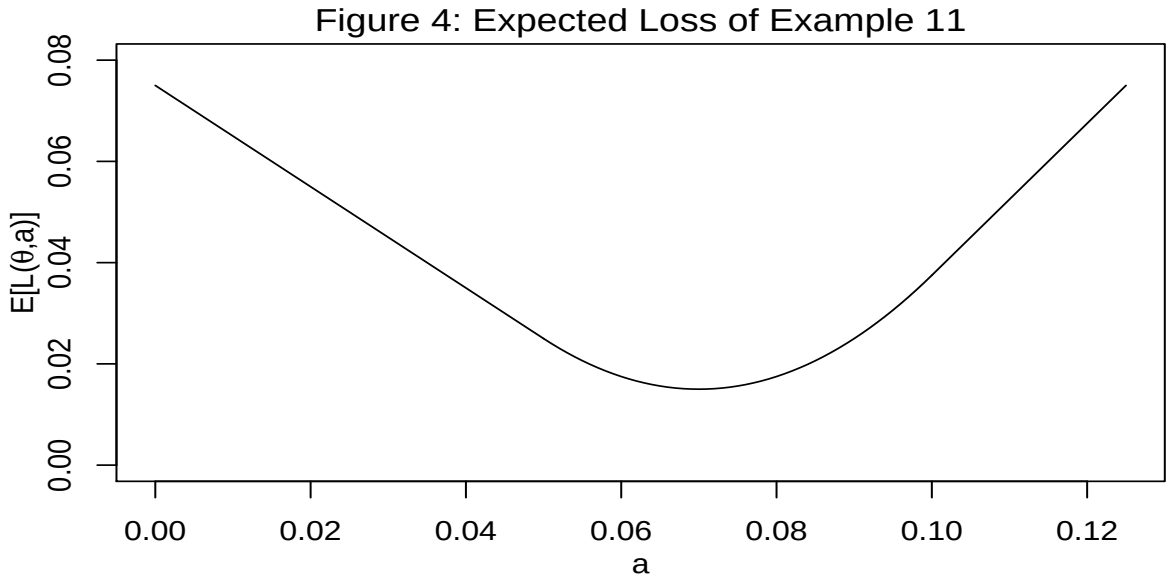
Example 7 (Continued): Suppose based on one's experience with the stock market one thinks that there is 60% chance of it going up next year and thus there is a 40% chance of incurring a loss. Based one these quantitative uncertainties one finds that the expected loss of action a_1 equals -200×0.6+100×0.4=-80 and that of action a_2 equals -60×0.6-60×0.4=-60. Thus the optimal Bayesian decision would be to invest that Rs.1000 in the stock market. ▽

Example 8 (Continued): Suppose the weather forecast states that day that there is a 20% chance of rain. Then the expected loss of carrying an umbrella or action a_1 would be 0×0.2+1×0.8 =0.8, and that for not carrying an umbrella or action a_2 would be 2×0.2+0×0.8 =0.4. Since the expected loss in not carrying an umbrella is smaller, the optimal Bayesian decision would be to not to carry an umbrella. ▽

Example 9 (Continued): Based on previous experience it has been found that a newly introduced pain-reliever typically tends to capture between 5% and 10% of the market. Also it appears that any value in this range is equally likely to occur. Based on these information the uncertainty regarding the value of θ may be modeled using a Uniform$[0.05, 0.1]$ distribution which has the p.d.f. $\pi^*(\theta) = \begin{cases} 20 & \text{if } 0.05 \leq \theta \leq 0.1 \\ 0 & \text{otherwise} \end{cases}$. Thus the expected loss is given by

$$E_{\pi^*(\theta)}[L(\theta, a)] = \begin{cases} 20 \int_{0.05}^{0.1} (\theta - a)\, d\theta & = & 0.075 - a & \text{if } 0 \leq a < 0.05 \\ 20 \left[1.5 \int_{0.05}^{a} (a - \theta)\, d\theta + \int_{a}^{0.1} (\theta - a)\, d\theta\right] & = & 25a^2 - 3.5a + 0.1375 & \text{if } 0.05 \leq a \leq 0.1 \\ 30 \int_{0.05}^{0.1} (a - \theta)\, d\theta & = & 1.5a - 0.1125 & \text{if } a > 0.1 \end{cases}$$

This expected loss is plotted in Figure 4 below:



The entire right tail has not been plotted because it is clear from the above expression (and also the plot) that it is monotonically increasing. Thus the expected loss is going to be minimum for some some $a \in [0.05, 0.1]$. The exact value of this a can be found by differentiating $25a^2 - 3.5a + 0.1375$, which is $E_{\pi^*(\theta)}[L(\theta, a)]$ for $0.05 \leq a \leq 0.1$, w.r.t. a and equating it to 0, yielding the optimal Bayesian solution as $a = 3.5/50 = 0.07$ as an the estimate of θ. $\bigtriangledown$

Examples such as above involve what is called a "no-data" situation, while the subject matter of statistics deals with the problem of taking decisions in face of uncertainty with some data at hand. In decision theoretic parlance this business of collecting observations, which carry some information regarding the unknown state of nature θ is called sneaking into nature. The real difference between the two paradigms (frequentist and Bayesian) become apparent in presence of data. It is not just a methodological issue of Minimax versus Bayesian Expected Loss. The philosophical standing itself of the two paradigms are different.

Thus now for the decision problem $(\Theta, \mathcal{A}, L)$ suppose we have sneaked into nature and collected some observation Y[12] whose c.d.f. $F(y|\theta)$ depends on the unknown θ. Let $\mathcal{Y}$ denote the sample space of possible values Y can take. Now a decision δ is defined as a function from the sample space $\mathcal{Y}$ to the action space $\mathcal{A}$ denoted as $\delta : \mathcal{Y} \longrightarrow \mathcal{A}$, with the understanding that if one observes $Y = y$, then one takes the action or decision $\delta(y)$, which is a member of $\mathcal{A}$. In the frequentist set-up the problem now shifts from taking an optimal action a to that of choosing an optimal decision $\delta(y)$, and thus complicating the problem further, while not much changes in the corresponding Bayesian set-up.

In the Bayesian frame-work since the criterion of choosing an optimal decision is the one which minimizes $E_{\pi^*(\theta)}[L(\theta, a)]$, the Bayesian Expected Loss, the solution is straight forward. In the "no-data" case we used the prior $\pi(\theta)$ for $\pi^*(\theta)$. When we have the observation $\boldsymbol{Y} = \boldsymbol{y}$,

[12]This notation and also θ instead of $\boldsymbol{\theta}$ might look like we are implying a single observation and a scalar θ. But that is not the case, the theory applies equally well for multiple observations with a vector valued $\boldsymbol{\theta}$.

since we have updated our uncertainty regarding $\boldsymbol{\theta}$ in terms of its posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{Y} = \boldsymbol{y})$, all one needs to do now is compute the Bayesian Expected Loss using this posterior distribution instead *i.e.* use $\pi(\boldsymbol{\theta}|\boldsymbol{Y} = \boldsymbol{y})$ for $\pi^*(\theta)$ in the formula of the Bayesian Expected Loss, and then take that decision which yields the smallest expected loss.

Example 8 (Continued): Let the loss function and the prior be as before. Now suppose you sneak into nature and check-out whether it is cloudy or not. This now becomes your observation or data. In order to compute the posterior probabilities of whether it is going to rain or not, now you need a model. Recall that the two states of nature that we are concerned with are θ_1 and θ_2 denoting "Rain" and "No Rain" respectively. Suppose you sneak out of your window and find that it is cloudy, and let this event be denoted by C. Thus now as a model you have to specify the probabilities of the observation C for each of the two states of nature. Since rain cannot happen without cloud it is only natural to assume that $P(C|\theta_1) = 1$. Now suppose based on your past experience you have found that on 40% of the days when it did not rain, the sky was still cloudy. This gives $P(C|\theta_2) = 0.4$. Now with all the elements in place, let us compute the posterior probabilities of θ_1 and θ_2 given the observation C. By (2), and the prior probabilities of "Rain" and "No Rain", which are based on the weather forecast of that day, which stated that there is a 20% chance of rain,

$$P(\theta_1|C) = \frac{P(C|\theta_1)P(\theta_1)}{P(C|\theta_1)P(\theta_1) + P(C|\theta_2)P(\theta_2)} = \frac{1 \times 0.2}{1 \times 0.2 + 0.4 \times 0.8} = 0.3846$$

and thus $P(\theta_2|C) = 1 - 0.3846 = 0.6154$. Now based on this posterior distribution, the expected loss of action a_1, that of carrying an umbrella, is 0.6154; and that of action a_2, not carrying an umbrella, is $2 \times 0.3846 = 0.7692$. Therefore the optimal Bayesian action to take would be to carry an umbrella, because it has smaller expected loss. $\bigtriangledown$

Example 6 (Continued): Given $Y_1, Y_2, \ldots Y_n$ i.i.d. $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown, suppose we wish to obtain a point estimate of the unknown σ^2. In order to formulate this as a problem of taking a decision we need to introduce its three elements, of which obviously $\Theta = \mathcal{A} = [0, \infty)$. Now for the point estimation problem, where we wish to estimate an unknown scalar θ by a, the standard loss function that is employed is given by $L(\theta, a) = (\theta - a)^2$, called the **squared error loss**[13]. For the optimal Bayesian point estimate, we choose that a as an estimate, which minimizes $E_{\pi(\theta|\boldsymbol{y})}[(\theta - a)^2]$, which is the expectation of $(\theta - a)^2$ taken w.r.t. $\pi(\theta|\boldsymbol{y})$, the posterior distribution of θ given $\boldsymbol{Y} = \boldsymbol{y}$. It is shown in §6, that the a which minimizes $E_{\pi(\theta|\boldsymbol{y})}[(\theta - a)^2]$ is the posterior expectation of θ denoted by $E_{\pi(\theta|\boldsymbol{y})}[\theta] = \int_\Theta \theta\pi(\theta|\boldsymbol{y})\, d\theta$. If we work with the non-informative prior on (μ, σ) as in (7), then according to equation (10) since the posterior distribution of $\tau = 1/\sigma^2$

[13]This squared error loss function is the mother of the Mean Square Error (MSE) criterion for point estimation in the frequentist framework. If a is replaced by the decision rule $\delta(\boldsymbol{Y})$ and then one takes the expectation of the squared error loss w.r.t. the (sampling) distribution of $\boldsymbol{Y}$ then one gets the MSE of the decision rule or estimator $\delta(\boldsymbol{Y})$. However in the Bayesian framework since one does not average over the observation $\boldsymbol{Y}$, the criterion of minimization is not MSE but the expectation of the squared error loss w.r.t. the posterior distribution of θ given $\boldsymbol{Y} = \boldsymbol{y}$.

is Gamma($\alpha = \nu/2$,$\lambda = \nu s^2_{n-1}/2$) the posterior mean of $\sigma^2 = 1/\tau$ is given by

$$\frac{\left(\nu s^2_{n-1}/2\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)} \int_0^\infty \tau^{(\nu/2-1)-1} \exp\left(-\tau\nu s^2_{n-1}/2\right)\, d\tau = \frac{\left(\nu s^2_{n-1}/2\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)} \frac{\Gamma\left(\frac{\nu}{2}-1\right)}{\left(\nu s^2_{n-1}/2\right)^{\nu/2-1}} = s^2_{n-2}$$

Thus the optimal Bayesian point estimate of σ^2 under the squared error loss is given by $s^2_{n-2} = \frac{1}{n-2}\sum_{i=1}^n (y_i - \overline{y})^2$. $\bigtriangledown$

Example 1 (Continued): Based on the data that 3 out 12 consumers chose brand X toothbrush, we wish to find a 95% Bayesina interval estimate of π. Actually the numerical solution to this problem *viz.* $[0.0747, 0.5140]$ has already been mentioned in page 16, with the interval shown with two vertical dashed lines in the posterior of π (given this data and a non-informative flat prior on π) in Figure 1. Here we provide a decision theoretic justification of this interval estimate.

Thus as usual let $\Theta = [0, 1]$, the set of values the unknown parameter π can take. In this case the action space needs to be defined with some caution. Since we are interested in an interval estimate our eventual decision would be in the form of a subset A of $[0, 1]$. However since the interval estimate must be such that it has a posterior probability content of 95%, we shall confine ourselves only to those subsets of $[0, 1]$ such that $\int_A \pi(\pi|\boldsymbol{y})\, d\pi = 0.95$. One can meaningfully talk about this integral, only for "measurable" A's (do not bother if you do not know what "measurable" means, it just ensures that we can talk about integrals over these sets). Thus $\mathcal{A} = \{\text{measurable } A \subseteq [0, 1] : \int_A \pi(\pi|\boldsymbol{y})\, d\pi = 0.95\}$. A natural loss function in the case of interval estimate is its length. Thus let $L(\pi, A) = \int_A d\pi$, which gives the length of the set A. Note that this loss function does not depend on π (that π in the integral is just a dummy variable) and thus the Bayesian expected loss is same as the loss function itself. We shall show in §6, that the solution of this problem *i.e.* the set $A \in \mathcal{A}$ with minimum $\int_A d\pi$ is of the form $A = \{\pi : \pi(\pi|\boldsymbol{y}) \geq k\}$ where the constant k is determined by the equation $\int_{\{\pi:\pi(\pi|\boldsymbol{y})\geq k\}} \pi(\pi|\boldsymbol{y})\, d\pi = 0.95$. Such an interval estimate is called 95% Highest Posterior Density (HPD) Credible set.

An algorithm can be easily worked out for unimodal densities which simultaneously gives the value of k together with the optimal 100(1- α)% HPD credible set for any given α. A numerical implementation of this algorithm for this example churned out the value of k as 0.59 and the interval as $[0.0747, 0.5140]$ for $\alpha = 0.05$. This value of k is indicated by the horizontal dotted line in Figure 1. $\bigtriangledown$

We shall end this section after providing a very brief account of what one does in the frequentist decision theoretic set up. Since the frequentist is interested in the behavior of a decision $\delta(\boldsymbol{Y})$ over repeated sampling, s/he first computes the quantity $E_{\boldsymbol{\theta}} L[\boldsymbol{\theta}, \delta(\boldsymbol{Y})]$, called the **Frequentist Risk**, denoted by $R(\boldsymbol{\theta}, \delta)$ w.r.t. the sampling distribution of $\boldsymbol{Y}$. Since this is a function of $\boldsymbol{\theta}$, unlike the Bayesian Expected Loss, for a given decision, the task of choosing an optimal decision is in general a much harder problem for the frequentist. At this point the frequentist either chooses a decision rule following the Minimax principle, which seeks to minimize the maximum risk; or invokes some other principle like invariance to deal with $R(\boldsymbol{\theta}, \delta)$; or at last s/he puts a prior $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$, computes what is called **Bayes Risk**, defined as $r(\pi, \delta) = \int_{\boldsymbol{\Theta}} R(\boldsymbol{\theta}, \delta)\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, and then chooses that decision which has

the smallest Bayes risk, which is possible because unlike the frequentist risk, Bayes risk is a single number. Such a decision rule which minimizes the Bayes risk, is called **Bayes rule**. Obviously for the "no-data" problem the two criteria of minimizing the Bayesian Expected Loss and Bayes risk are one and the same and thus both of these two criteria yield the same optimal decision. This connection goes even deeper and in general it is fairly easy to show that both these criteria yield the same optimal decision for any given set of data, providing a connection between the two paradigms. However it is usually computationally much easier to determine a rule that minimizes the Bayesian Expected Loss than Bayes risk. In any case from this point on we shall stop mentioning frequentist criteria and will solely concentrate on the Bayesian methods, with only pointing out their connections, when there is one

6 Bayesian Inference

Structurally this section is going to be similar to §2. Thus we shall pick up the standard inference problems one by one and will provide general Bayesian solutions to these problems. Also §2 just gave glimpses of the frequentist methods without going into any kind of details. But this being a notes on Bayesian Statistics, here we shall provide a more detailed discussion of the methods with proofs wherever needed. Though we have already had a brush with some of these inference problems under the Bayesian umbrella in some of the examples in the previous sections, here we shall discuss the methods of obtaining **Bayes rule**, which as mentioned above is equivalent to minimizing the Bayesian Expected Loss, in a more systematic manner for general problems in a Bayesian decision theoretic frame-work, collected at one place. Also without loss of generality here we shall assume that the unknown parameter $\boldsymbol{\theta}$ is continuous. For discrete $\boldsymbol{\theta}$ one just replaces the integrals by summation and all the results essentially follow.

6.1 Point Estimation

Philosophically many Bayesians do not believe in point estimation. Because to a Bayesian, you are uncertain about the value of $\boldsymbol{\theta}$, and this uncertainty has been capsuled in the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ of $\boldsymbol{\theta}$. Thus any real further action should use the posterior distribution as a whole instead of recommending a single value from it. Nonetheless there are situations where the real action itself might be choosing a value from the posterior distribution. Also there is this illustrative point of how can a Bayesian tackle this particular inference problem, if challenged with. These considerations make the discussion of the topic of Bayesian Point Estimation worthwhile.

Proposition 1: Let θ be a scalar with posterior p.d.f. $\pi(\theta|\boldsymbol{y})$ and the loss $L(\theta, a)$ be squared error *i.e.* $L(\theta, a) = (\theta - a)^2$ with $\mathcal{A} = \Theta$. Then the Bayes rule for a is given by $E_{\pi(\theta|\boldsymbol{y})}[\theta]$, which is the expectation or mean of θ according to its posterior distribution.

Proof: In order to obtain the Bayes rule, we are to find the a which minimizes $E_{\pi(\theta|\boldsymbol{y})}[(\theta - a)^2]$.

Let $\delta(\boldsymbol{y}) = E_{\pi(\theta|\boldsymbol{y})}[\theta]$. Then $\forall a \in \mathcal{A}$

$$
\begin{aligned}
& E_{\pi(\theta|\boldsymbol{y})}\left[(\theta - a)^2\right] \\
&= E_{\pi(\theta|\boldsymbol{y})}\left[\{(\theta - \delta(\boldsymbol{y})) + (\delta(\boldsymbol{y}) - a)\}^2\right] \\
&= E_{\pi(\theta|\boldsymbol{y})}\left[(\theta - \delta(\boldsymbol{y}))^2\right] + (\delta(\boldsymbol{y}) - a)^2 + (\delta(\boldsymbol{y}) - a)\, E_{\pi(\theta|\boldsymbol{y})}[(\theta - \delta(\boldsymbol{y}))] \\
&\quad \text{(because } (\delta(\boldsymbol{y}) - a) \text{ is a constant free of } \theta.\text{)} \\
&= E_{\pi(\theta|\boldsymbol{y})}\left[(\theta - \delta(\boldsymbol{y}))^2\right] + (\delta(\boldsymbol{y}) - a)^2 \\
&\quad \text{(because } E_{\pi(\theta|\boldsymbol{y})}[(\theta - \delta(\boldsymbol{y}))] = E_{\pi(\theta|\boldsymbol{y})}[\theta] - E_{\pi(\theta|\boldsymbol{y})}[\delta(\boldsymbol{y})] = \delta(\boldsymbol{y}) - \delta(\boldsymbol{y}) = 0.\text{)} \\
&\geq E_{\pi(\theta|\boldsymbol{y})}\left[(\theta - \delta(\boldsymbol{y}))^2\right] \quad \text{(because } (\delta(\boldsymbol{y}) - a)^2 \text{ is always } \geq 0.\text{)}
\end{aligned}
$$

Thus $\forall a \in \mathcal{A}$, $E_{\pi(\theta|\boldsymbol{y})}[(\theta - a)^2] \geq E_{\pi(\theta|\boldsymbol{y})}\left[(\theta - \delta(\boldsymbol{y}))^2\right]$ with the equality attaining when $a = \delta(\boldsymbol{y}) = E_{\pi(\theta|\boldsymbol{y})}[\theta]$. Therefore the decision which minimizes the expected squared error loss is the posterior mean. $\bigtriangledown$

Example 6 (Continued) in pages 25-26 gives an example as an illustration of this result. Actually the above result is valid even for the multi-parameter case with general quadratic loss, as stated in the following Proposition.

Proposition 2: Let $\boldsymbol{\theta}$ be a $p \times 1$ vector of unknown parameters with posterior p.d.f. $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. Let the loss in estimating $\boldsymbol{\theta}$ by $\boldsymbol{a} \in \boldsymbol{\mathcal{A}} = \boldsymbol{\Theta}$, $L(\boldsymbol{\theta}, \boldsymbol{a}) = (\boldsymbol{\theta} - \boldsymbol{a})'\boldsymbol{Q}(\boldsymbol{\theta} - \boldsymbol{a})$ for some $p \times p$ positive definite matrix $\boldsymbol{Q}$. Such loss functions are called **Quadratic Loss**. Then the Bayes rule for $\boldsymbol{a}$ is given by $E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[\boldsymbol{\theta}]$, which is the co-ordinate-wise expectation or mean of the components of $\boldsymbol{\theta}$ according to its posterior distribution.

Proof: In order to obtain the Bayes rule, we are to find the $\boldsymbol{a}$ which minimizes $E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[(\boldsymbol{\theta} - \boldsymbol{a})^2]$. Let $\boldsymbol{\delta}(\boldsymbol{y}) = E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[\boldsymbol{\theta}]$. Then $\forall \boldsymbol{a} \in \boldsymbol{\mathcal{A}}$

$$
\begin{aligned}
& E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[(\boldsymbol{\theta} - \boldsymbol{a})'\boldsymbol{Q}(\boldsymbol{\theta} - \boldsymbol{a})] \\
&= E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}\left[\{(\boldsymbol{\theta} - \boldsymbol{\delta}(\boldsymbol{y})) + (\boldsymbol{\delta}(\boldsymbol{y}) - \boldsymbol{a})\}'\, \boldsymbol{Q}\, \{(\boldsymbol{\theta} - \boldsymbol{\delta}(\boldsymbol{y})) + (\boldsymbol{\delta}(\boldsymbol{y}) - \boldsymbol{a})\}\right] \\
&= E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}\left[(\boldsymbol{\theta} - \boldsymbol{\delta}(\boldsymbol{y}))'\, \boldsymbol{Q}\, (\boldsymbol{\theta} - \boldsymbol{\delta}(\boldsymbol{y}))\right] + (\boldsymbol{\delta}(\boldsymbol{y}) - \boldsymbol{a})'\, \boldsymbol{Q}\, (\boldsymbol{\delta}(\boldsymbol{y}) - \boldsymbol{a}) \\
&\text{(because as in the scalar case, } (\boldsymbol{\delta}(\boldsymbol{y}) - \boldsymbol{a}) \text{ is a constant free of } \boldsymbol{\theta} \text{ and } E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[(\boldsymbol{\theta} - \boldsymbol{\delta}(\boldsymbol{y}))] = 0.\text{)} \\
&\geq E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}\left[(\boldsymbol{\theta} - \boldsymbol{\delta}(\boldsymbol{y}))'\, \boldsymbol{Q}\, (\boldsymbol{\theta} - \boldsymbol{\delta}(\boldsymbol{y}))\right] \\
&\quad \text{(because } (\boldsymbol{\delta}(\boldsymbol{y}) - \boldsymbol{a})'\, \boldsymbol{Q}\, (\boldsymbol{\delta}(\boldsymbol{y}) - \boldsymbol{a}) \text{ is always } \geq 0 \text{ as } \boldsymbol{Q} \text{ is positive definite.)}
\end{aligned}
$$

showing as before that the minimum expected loss getting attained by $\boldsymbol{\delta}(\boldsymbol{y}) = E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[\boldsymbol{\theta}]$, the co-ordinate-wise posterior mean. Note that this is true for any $p \times p$ positive definite matrix $\boldsymbol{Q}$ and thus the estimator $\boldsymbol{\delta}(\boldsymbol{y})$ does not depend on $\boldsymbol{Q}$. $\bigtriangledown$

Now instead of quadratic loss sometimes one might envisage an absolute error loss. The following proposition pertain to such situations for scalar-valued parameters.

Proposition 3: Let θ be a scalar with posterior p.d.f. $\pi(\theta|\boldsymbol{y})$ and the loss function $L(\theta, a)$ be absolute error *i.e.* $L(\theta, a) = |\theta - a|$ with $\mathcal{A} = \Theta$. Then the Bayes rule for a is the posterior median, denoted by $\tilde{\theta}$, where $\tilde{\theta}$ is such that $P(\theta < \tilde{\theta}|\boldsymbol{y}) = P(\theta > \tilde{\theta}|\boldsymbol{y}) = 0.5$.

Proof: Let $a \in \mathcal{A}$ be such that $a > \tilde{\theta}$. Then

$$L(\theta, \tilde{\theta}) - L(\theta, a) = \begin{cases} \tilde{\theta} - a & \text{if } \theta < \tilde{\theta} \\ 2\theta - (\tilde{\theta} + a) & \text{if } \tilde{\theta} \leq \theta \leq a \\ a - \tilde{\theta} & \text{if } \theta > a \end{cases}$$

Now for $\theta \leq a$ since $2\theta - (\tilde{\theta} + a) \leq a - \tilde{\theta}$ it follows that

$$L(\theta, \tilde{\theta}) - L(\theta, a) \leq (\tilde{\theta} - a) I_{(-\infty, \tilde{\theta})}(\theta) + (a - \tilde{\theta}) I_{[\tilde{\theta}, \infty)}(\theta)$$

where the function $I_A(x)$ with x as its argument is the indicator function of the set A defined as $I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$. Now taking expectation w.r.t. $\pi(\theta|\boldsymbol{y})$ on either side yields

$$E_{\pi(\theta|\boldsymbol{y})} \left[L(\theta, \tilde{\theta}) - L(\theta, a) \right] \leq 0.5(\tilde{\theta} - a) + 0.5(a - \tilde{\theta}) = 0$$

establishing that the posterior expected loss of $\tilde{\theta}$ is as small as a. A similar argument shows the same for $a < \tilde{\theta}$. Thus the expected absolute error loss is minimized for the posterior median. $\bigtriangledown$

Example 1 (Continued): Under the absolute error loss the best estimate is the posterior median. For the posterior depicted in Figure 1 it may be numerically computed that the median = 0.2753. For comparison, under the squared error loss the mean of the posterior is 0.285714, which for this example can be analytically computed. $\bigtriangledown$

An extension of the above result deals with the problem of general linear loss, such as one encountered in Example 9, which we state without proof.

Proposition 4: Let θ be a scalar with posterior p.d.f. $\pi(\theta|\boldsymbol{y})$ and let the loss function $L(\theta, a) = \begin{cases} c_0(\theta - a) & \text{if } \theta \geq a \\ c_1(a - \theta) & \text{if } \theta < a \end{cases}$, for some known constants c_0 and c_1. Then the Bayes rule for a is given by the $c_0/(c_0 + c_1)$-th quantile of $\pi(\theta|\boldsymbol{y})$. $\bigtriangledown$

Example 9 (Continued): In this new pain-reliever introduction example, where a company is interested in estimating the proportion of market the new pain-reliever is going to capture, we had a loss function which was same as that in Proposition 4, with $c_0 = 1$ and $c_2 = 1.5$. Thus a routine application of the above result states that the optimal estimate in this case is given by the 1/2.5=0.4-th quantile of the posterior or in case there is no data, as in Example 9, the prior distribution. Recall that the prior for this example was Uniform$[0.05, 1]$ and thus its 0.4-th quantile or the 40-th percentile is given by 0.07, which is same as the answer we had already obtained (see page 24) using a direct method which did not use Proposition 4 above. $\bigtriangledown$

6.2 Interval Estimation

Though we have named this subsection as above, it is more of a legacy of the reminiscent frequentist past. This is because what we are interested in doing here is provide a set of

values which has some pre-assigned (typically high) probability of containing the value of $\boldsymbol{\theta}$ given $\boldsymbol{y}$. Such sets are called **credible sets** and the chance coefficient associated with it is called its **credibility**. Formally a set $\boldsymbol{C} \subseteq \boldsymbol{\Theta}$, is called a 100(1-$\alpha$)% credible set if $P(\boldsymbol{\theta} \in \boldsymbol{C}|\boldsymbol{y}) = \int_{\boldsymbol{C}} \pi(\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{\theta}$ = 1-α, for $0 < \alpha < 1$. Now there is no reason why such a credible set would be an interval. In principle, it could be any arbitrary set, thus a more appropriate name for this subsection would be "Credible Set Determination".

The issue of the structure of the Action space $\boldsymbol{\mathcal{A}}$ in this kind of credible set determination problem was briefly touched upon in Example 1 (Continued) in page 26. Here we essentially provide the same formulation in a systematic general context. The final action or decision that will be taken in the credible set determination problem is going to be in the form of a set $\boldsymbol{C} \subseteq \boldsymbol{\Theta}$. Since we are only interested in credible sets with credibility 1-α, we need only consider those subsets of $\boldsymbol{C} \subseteq \boldsymbol{\Theta}$ with $\int_{\boldsymbol{C}} \pi(\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{\theta}$ = 1-α. However one can meaningfully talk about the credibility of a set $\boldsymbol{C}$ as expressed in terms of an integral as above if and only if, it is measurable. Thus formally for the credible set determination problem, $\boldsymbol{\mathcal{A}}$ $=\Big\{\text{measurable } \boldsymbol{C} \subseteq \boldsymbol{\Theta} : \int_{\boldsymbol{C}} \pi(\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{\theta} = 1 - \alpha\Big\}$.

Now for some fixed $\alpha \in (0, 1)$ we are interested in obtaining an "optimal" Bayes rule of choosing a $\boldsymbol{C}$ from $\boldsymbol{\mathcal{A}}$, where the optimality is determined in terms of some loss function. Since the credibility can be increased by enlarging the size of a set, for a fixed credibility, it is thus natural to seek for a set whose size is small. Thus it is very natural to express the loss in terms of the size of the action $\boldsymbol{C}$, with the understanding that larger the size more is the loss and an action will be considered to be "optimal" if it has the smallest loss. For scalar θ the size of a credible set C is its length, for a two-dimensional $\boldsymbol{\theta}$ the size of $\boldsymbol{C}$ would be its area, for a three-dimensional $\boldsymbol{\theta}$ the size of $\boldsymbol{C}$ is its volume and in general for a $\boldsymbol{\theta}$ of arbitrary finite dimension p, the size of a credible set $\boldsymbol{C}$ is simply given by $\int_{\boldsymbol{C}} d\boldsymbol{\theta}$.

However this notion of size can be generalized and one can define an abstract size of a set in terms of a real-valued size function $s(\boldsymbol{\theta})$ defined on $\boldsymbol{\Theta}$ and size of a set $\boldsymbol{C}$ as $\int_{\boldsymbol{C}} s(\boldsymbol{\theta})\ d\boldsymbol{\theta}$. In most practical applications, as in Example 1 (Continued) in page 26, $s(\boldsymbol{\theta}) \equiv 1\ \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ yielding the usual size of set in terms of its length/area/volume/hyper-volume. But there is no harm in deriving the Bayes rule for this more generalized notion of the size of a set. With this abstract definition of size, now the loss function for this credible set determination problem is defined as $L(\boldsymbol{\theta}, \boldsymbol{C}) = \int_{\boldsymbol{C}} s(\boldsymbol{\theta})\ d\boldsymbol{\theta}$. Note that this loss function does not depend on $\boldsymbol{\theta}$ (the $\boldsymbol{\theta}$ in the integral is just a dummy variable of integration), and thus the expected loss is same as the loss itself. Thus now the problem of optimal credible set determination is formulated as determining a $\boldsymbol{C}$ from $\boldsymbol{\mathcal{A}}$= $\Big\{\text{measurable } \boldsymbol{C} \subseteq \boldsymbol{\Theta} : \int_{\boldsymbol{C}} \pi(\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{\theta} = 1 - \alpha\Big\}$ such that $\int_{\boldsymbol{C}} s(\boldsymbol{\theta})\ d\boldsymbol{\theta}$ is minimum, a solution of which is provided in Proposition 5 below.

Proposition 5: Let $\boldsymbol{C}^* \in \boldsymbol{\mathcal{A}}$ be as $\boldsymbol{C}^* = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \pi(\boldsymbol{\theta}|\boldsymbol{y}) \geq k s(\boldsymbol{\theta})\}$, where the constant k is determined from the equation $\int_{\boldsymbol{C}^*} \pi(\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{\theta} = 1 - \alpha$, which $\boldsymbol{C}^*$ has to satisfy if it were to belong to $\boldsymbol{\mathcal{A}}$. Now let $\boldsymbol{C}$ be any arbitrary set in $\boldsymbol{\mathcal{A}}$. Then $\int_{\boldsymbol{C}} s(\boldsymbol{\theta})\ d\boldsymbol{\theta} \geq \int_{\boldsymbol{C}^*} s(\boldsymbol{\theta})\ d\boldsymbol{\theta}$.

Proof:

$$\int_{\boldsymbol{C}} s(\boldsymbol{\theta})\ d\boldsymbol{\theta} - \int_{\boldsymbol{C}^*} s(\boldsymbol{\theta})\ d\boldsymbol{\theta}$$

$$
\begin{aligned}
&= \int_{\boldsymbol{C} \bigcap \boldsymbol{C}^{*^c}} s(\boldsymbol{\theta})\, d\boldsymbol{\theta} - \int_{\boldsymbol{C}^* \bigcap \boldsymbol{C}^c} s(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&\quad \text{(by eliminating the common part } \int_{\boldsymbol{C} \bigcap \boldsymbol{C}^*} s(\boldsymbol{\theta})\, d\boldsymbol{\theta} \text{ present in both the integrals.)} \\
&\geq \frac{1}{k}\left[\int_{\boldsymbol{C} \bigcap \boldsymbol{C}^{*^c}} \pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} - \int_{\boldsymbol{C}^* \bigcap \boldsymbol{C}^c} \pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}\right] \\
&\quad \text{(because, in } \boldsymbol{C}^{*^c},\ s(\boldsymbol{\theta}) > \pi(\boldsymbol{\theta}|\boldsymbol{y})/k, \text{while in } \boldsymbol{C}^*,\ -s(\boldsymbol{\theta}) \geq -\pi(\boldsymbol{\theta}|\boldsymbol{y})/k.) \\
&= \frac{1}{k}\left[\int_{\boldsymbol{C}} \pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} - \int_{\boldsymbol{C}^*} \pi(\boldsymbol{\theta}|\boldsymbol{y})\right] \\
&\quad \text{(by adding the common part } \int_{\boldsymbol{C} \bigcap \boldsymbol{C}^*} \pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} \text{ to both the integrals.)} \\
&= \frac{1}{k}\left[(1-\alpha) - (1-\alpha)\right] \\
&\quad \text{(since both } \boldsymbol{C} \text{ and } \boldsymbol{C}^* \text{ are in } \boldsymbol{\mathcal{A}},\ \int_{\boldsymbol{C}} \pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} = \int_{\boldsymbol{C}^*} \pi(\boldsymbol{\theta}|\boldsymbol{y}) = 1-\alpha.) \\
&= 0
\end{aligned}
$$

thus establishing $\int_{\boldsymbol{C}} s(\boldsymbol{\theta})\, d\boldsymbol{\theta} \geq \int_{\boldsymbol{C}^*} s(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ $\bigtriangledown$

Proposition 5 thus shows that the optimal $\boldsymbol{C}$ in the action space $\boldsymbol{\mathcal{A}}$ that minimizes the loss, defined as the size of $\boldsymbol{C}$, is of the form $\boldsymbol{C}^* = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \pi(\boldsymbol{\theta}|\boldsymbol{y}) \geq k s(\boldsymbol{\theta})\}$, where the constant k is determined from the equation $\int_{\boldsymbol{C}^*} \pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} = 1\text{-}\alpha$. When $s(\boldsymbol{\theta}) \equiv 1\ \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, as is the case in most practical applications, this optimal $\boldsymbol{C}$ is called a $100(1\text{-}\alpha)\%$ Highest Posterior Density (HPD) Credible Set, as it consists of those points in $\boldsymbol{\Theta}$ where the posterior density is higher than the points which do not belong to it. The question, "How high is high?" is determined by the constant k, whose exact value depends on the amount of credibility $1\text{-}\alpha$ associated with the HPD credible set. In general larger the credibility smaller is the value of k and thus larger is the size of the HPD credible set.

An algorithmic implementation for computing such $100(1\text{-}\alpha)\%$ HPD credible sets for unidimensional θ is as follows. Let $C(k) = \{\theta \in \Theta : \pi(\theta|\boldsymbol{y}) \geq k\}$ and $P(k) = \int_{C(k)} \pi(\theta|\boldsymbol{y})\, d\theta$. Let $\pi_{\max} = \text{Maximum}_{\theta \in \Theta} \pi(\theta|\boldsymbol{y})$. If $\pi(\theta|\boldsymbol{y})$ is unimodal $\pi_{\max}$ is same as $\pi(\theta|\boldsymbol{y})$ evaluated at its mode, otherwise if $\pi(\theta|\boldsymbol{y})$ is multimodal, evaluate $\pi(\theta|\boldsymbol{y})$ at each of its local modes and then chose the maximum of these $\pi(\theta|\boldsymbol{y})$'s as $\pi_{\max}$. Note that $P(0) = 1$ and $P(\pi_{\max}) = 0$ and for intermediate values of k, first $C(k)$ can be determined by solving for $\pi(\theta|\boldsymbol{y}) = k$ and then considering the (in the multimodal case, union of) interval(s) between these solutions, and then $P(k)$ can be computed by numerically integrating $\pi(\theta|\boldsymbol{y})$ on $C(k)$. Thus using the extreme values 0 and $\pi_{\max}$ and the above algorithm for computing $P(k)$ for intermediate k's, the $100(1\text{-}\alpha)\%$ HPD credible set can be determined by solving $P(k)$=$1\text{-}\alpha$ using *regula falsi*.

Example 6 (Continued): Suppose we are interested in obtaining a $100(1\text{-}\alpha)\%$ HPD credible set for μ with $Y_1, Y_2, \ldots, Y_n$ i.i.d. $N(\mu, \sigma^2)$ with both μ and σ^2 unknown, using the non-informative prior on (μ, σ) given in (7). The marginal posterior p.d.f. of μ is given in equation (12). Since this p.d.f. is symmetric about and unimodal at $\overline{y}$ the $100(1\text{-}\alpha)\%$ HPD credible set for μ must be of the form $\overline{y} \pm k$ for some constant k, where the constant

k is determined from the equation $\int_{\overline{y}-k}^{\overline{y}+k} \pi(\mu|\boldsymbol{y})\ d\mu$ = 1-α. Now since we know that the posterior of μ is such that $(\mu - \overline{y}) / (s_{n-1}/\sqrt{n}) \sim t_{n-1}$, a simple change of variable in the above integral equation yields $k = t_{n-1,1-\alpha/2}\,(s_{n-1}/\sqrt{n})$, where $t_{n-1,\alpha}$ is the α-th quantile of a t distribution with $(n-1)$ d.f.. Thus the 100(1-α)% HPD credible set for μ is given by $\overline{y} \pm t_{n-1,1-\alpha/2}\,(s_{n-1}/\sqrt{n})$. $\bigtriangledown$

100(1-α)% HPD credible set for the π of a Binomial has to be found numerically as mentioned in Example 1 (Continued) in page 26, and likewise for the σ^2 of a $N(\mu, \sigma^2)$ model.

6.3 Hypotheses Testing

As has been mentioned once before, in the Bayesian frame-work one need not choose between only two possible hypotheses the null and the alternative. Here one can simultaneously test or choose among any finite assortment of mutually exclusive or disjoint hypotheses. However since frequentist hypotheses testing has been developed for the null and alternative pair, we shall first develop Bayesian hypothesis testing in this classical set-up.

In order to provide a decision theoretic solution for testing $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ versus $H_1 : \boldsymbol{\theta} = \boldsymbol{\Theta}_1$, we need to introduce the triplet $(\boldsymbol{\Theta}, \mathcal{A}, L)$, where $\boldsymbol{\Theta}$ is already given. In the hypothesis testing situation since one can take only one of the two possible final decisions, $\mathcal{A} = \{a_0, a_1\}$, where a_0 denotes the decision, "Accept H_0" and a_1 denotes the decision, "Accept H_1". Now intuitively in order to give a symmetric treatment to the two hypotheses, one should work with a so-called 0-1 loss function defined as $L(\boldsymbol{\theta}, a_i) = \begin{cases} 1 & \text{if } \boldsymbol{\theta} \notin \boldsymbol{\Theta}_i \\ 0 & \text{if } \boldsymbol{\theta} \in \boldsymbol{\Theta}_i \end{cases}$ for $i = 0, 1$. However if one really wants to met out a favorable treatment to one of the hypotheses, as in the frequentist set-up, one can work with a slightly more generalized version of the 0-1 loss function given by $L(\boldsymbol{\theta}, a_i) = \begin{cases} C_i & \text{if } \boldsymbol{\theta} \notin \boldsymbol{\Theta}_i \\ 0 & \text{if } \boldsymbol{\theta} \in \boldsymbol{\Theta}_i \end{cases}$ for some $C_i > 0$ for $i = 0, 1$. For instance, as in the frequentist set-up, if Accepting H_1 when it is false (Type-I error) is considered to be more serious than Accepting H_0 when it is false (Type-II error), one should choose $C_1 > C_0$, with their relative values chosen depending on the relative degree of severity of the two types of errors. Note that even how this asymmetric treatment to the two hypotheses can be directly handled in a quantitative manner in the Bayesian set-up, without resorting to indirect measures like error probabilities.

The Bayes rule for choosing the optimal decision is determined by choosing that action a_i which has smaller expected loss. According to the general 0-C_i loss function $E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}\,[L(\boldsymbol{\theta}, a_0)]$ $= C_0(1 - P(\boldsymbol{\theta}\boldsymbol{\Theta}_0|\boldsymbol{y}))$ and $E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}\,[L(\boldsymbol{\theta}, a_1)] = C_1(1 - P(\boldsymbol{\theta}\boldsymbol{\Theta}_0|\boldsymbol{y}))$. If $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$ are such that $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \bigcup \boldsymbol{\Theta}_1$, as would usually be the case with two hypotheses, then according to the expected loss computed above, one would take action a_1 or "Accept H_1" if

$$\frac{P(\boldsymbol{\Theta}_1|\boldsymbol{y})}{P(\boldsymbol{\Theta}_0|\boldsymbol{y})} > \frac{C_1}{C_0} \text{ or equivalently if } P(\boldsymbol{\Theta}_1|\boldsymbol{y}) > \frac{C_1}{C_0 + C_1}. \tag{13}$$

Above decision looks fairly intuitive, in the sense that it calls for taking the action a_1, if the posterior probability of $\boldsymbol{\Theta}_1$ is large or its value is large compared to that of the posterior

probability of $\mathbf{\Theta}_0$, and the issue of how large is large settled by the loss function.

Example 6 (Continued): Consider a situation where $Y_1, Y_2, \ldots, Y_n$ i.i.d. $N(\mu, \sigma^2)$ with known σ^2, and we are interested in testing the hypotheses $\begin{matrix} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{matrix}$ for some known μ_0. That is here $\Theta_0 = (-\infty, \mu_0]$ and $\Theta_1 = (\mu_0, \infty)$. As mentioned in footnote 11 in page 20, the posterior of μ in this case is $N(\overline{y}, \sigma^2/n)$. Thus according to (13) one should, "Accept H_1" or "Reject H_0" if

$$P(\Theta_1|\boldsymbol{y}) = 1 - \Phi\left(\frac{\mu_0 - \overline{y}}{\sigma/\sqrt{n}}\right) > \frac{C_1}{C_0 + C_1},$$

where $\Phi(\cdot)$ is the standard Normal c.d.f., or equivalently if

$$\overline{y} > \mu_0 + z_{C_0/(C_0+C_1)}\frac{\sigma}{\sqrt{n}}$$

where z_α is the α-th quantile of the standard Normal distribution. Note the similarity of the above Bayesian decision rule with the frequentist most powerful fixed significance level test. Both of them are exactly of the same form. In the frequentist case, the fixed significance level is determined subjectively in an ad hoc manner, while in the Bayesian case it is determined by the loss function, and a systematic subjective input which is specified in terms of the prior distribution. The prior input is not clear in this example as we are dealing with a non-informative prior, but in general the above comment holds true. We close this example after drawing a parallel with the observed significance level or p-value and the posterior probability of the null hypothesis. For this example the p-value is given by

$$P(\overline{Y} \geq \overline{y}|H_0) = 1 - \Phi\left(\frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\mu_0 - \overline{y}}{\sigma/\sqrt{n}}\right) = P(\Theta_0|\boldsymbol{y}).$$

The last but one equality follows from the fact that $1 - \Phi(z) = \Phi(-z)$. Thus at least in this example the p-value gives the posterior probability of the null hypothesis, and thus one can give a Bayesian justification of its interpretation as credibility of the H_0. However such parallels to the frequentist results for Bayesian hypothesis testing is more of a fluke than a rule. $\bigtriangledown$

Although the Bayesian decision theoretic solution to standard statistical hypothesis testing, specified in terms of a null and an alternative, is as simple as the rule specified in (13), and for the case of choosing between a set of multiple hypotheses one should still use a decision theoretic solution; a special Bayesian technique is available for the standard statistical hypothesis testing problem involving a H_0 and a H_1. Recall the hypothesis testing example mentioned in Example 1 (Continued) in page 16, where the posterior probability of the hypothesis $H : \pi > 1$ was found to be 0.9664, and it was mentioned that part of the reason for such a high posterior probability was its prior probability which was 0.9. Now we shall discuss a Bayesian method which attempts to eliminate this prior bias in deciding between the truth of the two hypotheses H_0 and H_1, based on the data.

Let $\alpha_0 = P(\mathbf{\Theta}_0|\boldsymbol{y})$ and $\alpha_1 = P(\mathbf{\Theta}_1|\boldsymbol{y})$ respectively denote the posterior probabilities of the two hypotheses H_0 and H_1. Also let π_0 and π_1 denote their respective prior probabilities.

Then the posterior (prior) odds ratio of H_0 to H_1 is given by α_0/α_1 (π_0/π_1). Now let B denote the ratio of this posterior odds to the prior odds *i.e.* let $B = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1}$. B is called the **Bayes factor** in favor of H_0. While deciding between two hypotheses the odds ratio gives a slightly clearer picture than the two raw probabilities. Because if $\alpha_0/\alpha_1 = 5$, it means that given the data, H_0 is 5 times more likely to be true than H_1. Now the Bayes factor can sometimes be interpreted as this odds ratio as purely given by the data.

This interpretation is easiest to see when both H_0 and H_1 are simple hypotheses *i.e.* when $\Theta_0 = \{\boldsymbol{\theta}_0\}$ and $\Theta_1 = \{\boldsymbol{\theta}_1\}$. In this case by (2), $\alpha_0 = \pi_0 L(\boldsymbol{\theta}_0|\boldsymbol{y})/\{\pi_0 L(\boldsymbol{\theta}_0|\boldsymbol{y}) + \pi_1 L(\boldsymbol{\theta}_1|\boldsymbol{y})\}$ and $\alpha_1 = \pi_1 L(\boldsymbol{\theta}_1|\boldsymbol{y})/\{\pi_0 L(\boldsymbol{\theta}_0|\boldsymbol{y}) + \pi_1 L(\boldsymbol{\theta}_1|\boldsymbol{y})\}$, where $L(\boldsymbol{\theta}|\boldsymbol{y})$ is the likelihood function of $\boldsymbol{\theta}$ given the data $\boldsymbol{y}$. Thus here $B = L(\boldsymbol{\theta}_0|\boldsymbol{y})/L(\boldsymbol{\theta}_1|\boldsymbol{y})$, which is a number free of the prior.

In general however the Bayes factor depends on the prior distribution given over Θ_0 and Θ_1. To see this, let $\Theta = \Theta_0 \bigcup \Theta_1$ and write the prior $\pi(\boldsymbol{\theta})$ on Θ as $\pi(\boldsymbol{\theta}) = \pi_0 g_0(\boldsymbol{\theta}) I_{\Theta_0}(\boldsymbol{\theta}) + \pi_1 g_1(\boldsymbol{\theta}) I_{\Theta_1}(\boldsymbol{\theta})$, where the indicator function $I_A(x)$ is as defined in page 29, $\pi_0 + \pi_1 = 1$, and each $g_i(\boldsymbol{\theta})$ is a proper density on Θ_i $i = 0, 1$. Then a simple calculation shows that $B = \int_{\Theta_0} L(\boldsymbol{\theta}|\boldsymbol{y}) g_0(\boldsymbol{\theta})\, d\boldsymbol{\theta} \Big/ \int_{\Theta_1} L(\boldsymbol{\theta}|\boldsymbol{y}) g_1(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. Because of the involvement of g_0 and g_1 in b, this cannot be viewed as a measure of the relative support for the hypotheses provided solely by the data. In many practical applications however B will be relatively insensitive to reasonable choices of the g_i's (as in Example 1 (Continued) in page 16) and then such an interpretation is reasonable.

A rule of thumb in using B is that, a $B < 3$ is not much of an evidence at all in favor of H_0, while a B between 3 and 6 gives a moderate amount evidence about the truth of H_0 and $B > 6$ gives an overwhelming support that H_0 is true. Using this yardstick, coming back to Example 1, B in favor the hypothesis $\pi > 1$ equals $\frac{0.9664}{0.0336} \Big/ \frac{0.9}{0.1} = 3.19$, indicating that the data renders a moderate amount of support to this hypothesis.

We close our discussion on hypotheses testing after mentioning how to handle a point null hypothesis. Since a continuous posterior gives probability 0 to a single point, this point merits some discussion. Suppose one is interested in testing $\begin{matrix} H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\ H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \end{matrix}$. In such situations one first specifies the prior π_0 for H_0 and $\pi_1 = 1 - \pi_0$ for H_1. In the next step one specifies a proper prior $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Note that since $\boldsymbol{\theta}_0$ is a single point, $\pi(\boldsymbol{\theta})$ might as well be a proper density on Θ. However since the prior on $\boldsymbol{\theta}$ now has both discrete and continuous components, some caution is required in deriving the posterior and eventually the Bayes factor. The marginal density of $\boldsymbol{Y}$ is given by $m(\boldsymbol{y}) = \pi_0 L(\boldsymbol{\theta}_0|\boldsymbol{y}) + (1 - \pi_0) m_1(\boldsymbol{y})$, where $m_1(\boldsymbol{y}) = \int_{\Theta - \{\boldsymbol{\theta}_0\}} L(\boldsymbol{\theta}|\boldsymbol{y}) \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. Again note that the domain of integration for determining $m_1(\boldsymbol{y})$ could have as well been whole of Θ. Thus the posterior probability α_0 of H_0 is given by $\alpha_0 = \pi_0 L(\boldsymbol{\theta}_0|\boldsymbol{y})/m(\boldsymbol{y})$ and α_1 is found as 1-α_0. Then after a couple of steps of algebra it can be shown that $B = L(\boldsymbol{\theta}_0|\boldsymbol{y})/m_1(\boldsymbol{y})$.

Example 1 (Continued): Suppose with the same observation, 3 choosing and 9 not choosing we wish to test the hypothesis that $\begin{matrix} H_0 : \pi = 0.1 \\ H_1 : \pi \neq 0.1 \end{matrix}$. Let π_0 and π_1 denote the respective prior probabilities of H_0 and H_1. Next under H_1 let us again assign a flat prior for π. That is

for $A \subseteq [0,1]$ the prior probability of $\pi \in A$ is given by $\pi(A) = \begin{cases} 0.5 + 0.5 \int_A d\pi & \text{if } 0.1 \in A \\ 0.5 \int_A d\pi & \text{if } 0.1 \notin A \end{cases}$. Then the marginal joint p.m.f. of the 12 0-1 valued $\boldsymbol{Y}$ at the observed $\boldsymbol{y}$ is given by $m(\boldsymbol{y}) = \pi_0 0.1^3 0.9^9 + (1-\pi_0) \int_0^1 \pi^3 (1-\pi)^9 \, d\pi = \pi_0 3.874 \times 10^{-4} + (1-\pi_0) 3.4965 \times 10^{-4}$. Therefore the posterior probability of the null hypothesis H_0 is given by $\alpha_0 = \pi_0 3.874 \times 10^{-4} / (\pi_0 3.874 \times 10^{-4} + (1-\pi_0) 3.4965 \times 10^{-4})$. For example under the natural choice of $\pi_0 = 0.5$, $\alpha_0 = 0.5256$, that is now there is only slightly more chance of the null being true. However for testing this hypothesis we shall depend on the Bayes factor, which does not depend on π_0 (and that is why we have been scrupulously avoiding a specific value of π_0 such as 0.5) and is given by $3.874 \times 10^{-4} / 3.4965 \times 10^{-4}) = 1.11$, which actually does not provide any strong evidence in favor of neither of the two hypotheses. Even after observing the data both the hypotheses still remain almost equally likely to be true. But note that using the two sided p-value, which is twice the numbers reported in page 11, the frequentist would not have Rejected H_0 for either model. $\bigtriangledown$

6.4 Predictive Inference

As has been mentioned a couple of times before, one major triumph of the Bayesian paradigm is the way in which it coherently handles the problem of prediction. For one of the motivations of the the predictive density formula we are going to present, given in equation (15) below, consider a slightly different problem of inference about a parametric function $\phi(\boldsymbol{\theta})$. Of course given the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ we shall derive $\pi(\phi|\boldsymbol{y})$, the posterior of $\phi(\boldsymbol{\theta})$ given $\boldsymbol{y}$ and then base all our inference like estimation, hypothesis testing etc. regarding $\phi(\boldsymbol{\theta})$ on this $\pi(\phi|\boldsymbol{y})$. In particular if one is interested in estimating $\phi(\boldsymbol{\theta})$ against a squared error loss, then the estimate will be given by its posterior mean. But note that one need not compute $\pi(\phi|\boldsymbol{y})$ for this posterior mean, because the posterior mean of $\phi(\boldsymbol{\theta})$ can be easily computed from the already obtained posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ of $\boldsymbol{\theta}$ as

$$E[\phi(\boldsymbol{\theta})|\boldsymbol{y}] = \int_{\boldsymbol{\Theta}} \phi(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\boldsymbol{y}) \, d\boldsymbol{\theta}. \tag{14}$$

Now suppose we have observed the data $\boldsymbol{Y} = \boldsymbol{y}$ and have obtained the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. Now we are to predict a new random variable Z whose distribution depends on the unknown $\boldsymbol{\theta}$ and might also depend of $\boldsymbol{Y}$. Thus let the conditional density of Z given $\boldsymbol{\theta}$ and $\boldsymbol{Y} = \boldsymbol{y}$ be denoted by $g(z|\boldsymbol{\theta}, \boldsymbol{y})$. If the distribution of Z does not depend on $\boldsymbol{Y}$ given $\boldsymbol{\theta}$, as in the case of i.i.d. observations, $\boldsymbol{y}$ will not appear in the conditional density of Z, but in general there is no harm in letting Z depend on $\boldsymbol{Y}$ as well. Now in the face of uncertainty regarding the value of $\boldsymbol{\theta}$ (which of course is capsuled in its posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$), if one views the problem of obtaining this conditional density $g(z|\boldsymbol{\theta}, \boldsymbol{y})$ as a problem of point estimation of the function $g(z|\boldsymbol{\theta}, \boldsymbol{y})$, point by point for each z, then for a fixed z, $g(z|\boldsymbol{\theta}, \boldsymbol{y})$ may be viewed as parametric function $\phi(\boldsymbol{\theta})$, a point estimator of which under squared error loss may be obtained using (14) as

$$p(z|\boldsymbol{y}) = \int_{\boldsymbol{\theta}} g(z|\boldsymbol{\theta}, \boldsymbol{y}) \pi(\boldsymbol{\theta}|\boldsymbol{y}) \, d\boldsymbol{\theta}. \tag{15}$$

The density $p(z|\boldsymbol{y})$ given in equation (15) is called the **predictive density** of Z given $\boldsymbol{Y} = \boldsymbol{y}$. (15) can of course be viewed from another purely distributional angle. In order to obtain the conditional density of Z given $\boldsymbol{Y} = \boldsymbol{y}$, we first note that apart from the innate dependence of Z on $\boldsymbol{Y}$, if at all there is one, Z essentially depends on $\boldsymbol{\theta}$, which in turn depends on $\boldsymbol{Y}$. Thus the conditional density of Z given $\boldsymbol{Y} = \boldsymbol{y}$ can be obtained by looking at the joint density of Z and $\boldsymbol{\theta}$, conditional on $\boldsymbol{Y} = \boldsymbol{y}$, and then integrating $\boldsymbol{\theta}$ out. The joint density of Z and $\boldsymbol{\theta}$, conditional on $\boldsymbol{Y} = \boldsymbol{y}$ is same as the product of the conditional density of Z given $\boldsymbol{\theta}$ and $\boldsymbol{Y}$, and the conditional density of $\boldsymbol{\theta}$ given $\boldsymbol{Y}$; which given $\boldsymbol{Y} = \boldsymbol{y}$, is same as $g(z|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})$. This is another argument for the predictive density in (15).

Example 6 (Continued): Suppose $Y_1, Y_2, \ldots, Y_n$ i.i.d. $N(\mu, \sigma^2)$ with known σ^2 and based on this data we are interested in predicting the behavior of a future observation Y_{n+1}. As mentioned in footnote 11 in page 20, given $\boldsymbol{Y} = \boldsymbol{y}$ the posterior of μ is $N(\overline{y}, \sigma^2/n)$ *i.e.* $\pi(\mu|\boldsymbol{y}) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp\left\{-\frac{n}{2\sigma^2}(\mu - \overline{y})^2\right\}$. Now the conditional density of Y_{n+1} given μ and $\boldsymbol{Y}$ does not depend on $\boldsymbol{Y}$ and is given by $f(y_{n+1}|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_{n+1} - \mu)^2\right\}$. Thus using (15), the predictive density of Y_{n+1} given $\boldsymbol{Y}$ is given by

$$
\begin{aligned}
& p(y_{n+1}|\boldsymbol{y}) \\
&= \frac{\sqrt{n}}{2\pi\sigma^2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}\left[(n+1)\mu^2 - 2(y_{n+1} + n\overline{y})\mu + (y_{n+1}^2 + n\overline{y}^2)\right]\right\}\, d\mu \\
&= \frac{\sqrt{n}}{2\pi\sigma^2} \exp\left\{-\frac{n+1}{2\sigma^2}\left(\frac{y_{n+1}^2 + n\overline{y}^2}{n+1} - \frac{y_{n+1}^2 + n^2\overline{y}^2 + 2ny_{n+1}\overline{y}}{(n+1)^2}\right)\right\} \times \\
&\quad \int_{-\infty}^{\infty} \exp\left\{-\frac{n+1}{2\sigma^2}\left(\mu - \frac{y_{n+1} + n\overline{y}}{n+1}\right)^2\right\}\, d\mu \\
&= \frac{\sqrt{n}}{2\pi\sigma^2}\sqrt{\frac{2\pi\sigma^2}{n+1}} \exp\left\{-\frac{n}{2\sigma^2(n+1)}\left(y_{n+1} - \overline{y}\right)^2\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2(1+n^{-1})}} \exp\left\{-\frac{1}{2\sigma^2(1+n^{-1})}\left(y_{n+1} - \overline{y}\right)^2\right\}
\end{aligned}
$$

implying that $Y_{n+1}|\boldsymbol{Y} \sim N\left(\overline{y}, \left(1 + \frac{1}{n}\right)\sigma^2\right) x1$. Thus for instance a 100(1-α)% highest predictive density credible interval for Y_{n+1}, after observing the data $\boldsymbol{Y} = \boldsymbol{y}$ is given by $\overline{y} \pm z_{1-\alpha/2}\sigma\sqrt{1 + \frac{1}{n}}$. $\bigtriangledown$

Example 2 (Continued): This is the first time we are getting back to this simple linear regression example since its introduction in page 3. That is suppose we have observations $\{(X_1 = x_1, Y_1 = y_1), (X_2 = x_2, Y_2 = y_2), \ldots, (X_n = x_n, Y_n = y_n)\}$, denoted by data $\boldsymbol{d}$, on monthly sales Y and advertising expenses X for n months, where $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$. Now based on this data and the assumed model we are to predict the sales Y_h for some month when the advertising expense has been x_h. For simplicity assume that σ^2 is known and put a flat non-informative prior on (β_0, β_1) given by

$$\pi(\beta_0, \beta_1) \propto 1 \text{ for } -\infty < \beta_0 < \infty \text{ and } -\infty < \beta_1 < \infty$$

Now the likelihood function for (β_0, β_1) is given by

$$L(\beta_0, \beta_1|\boldsymbol{d}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(n\beta_0^2 + \beta_1^2\sum_{i=1}^n x_i^2 - 2n\beta_0\overline{y} + 2\beta_1(\beta_0 - 1)\sum_{i=1}^n x_i y_i\right)\right\}.$$

In the regression set-up since the x_i's are considered to be given constants, statistics are the ones which are functions of the y_i's. Thus from the above likelihood function it may be seen that $(\overline{y}, \sum_{i=1}^n x_i y_i)$ are sufficient statistics. Since $(\overline{y}, \sum_{i=1}^n x_i y_i) \longleftrightarrow \left(\hat{\beta}_0, \hat{\beta}_1\right)$ is a one to one function where $\hat{\beta}_1 = S_{xy}/S_{xx}$ (where $S_{xy} = \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$ and $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2$) and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$ are maximum likelihood (least squares) estimates of (β_0, β_1), $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ is sufficient for (β_0, β_1). Thus by (A5), posterior of (β_0, β_1) is proportional to the sampling distribution of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$. Since $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \frac{\sigma^2}{S_{xx}}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix}\right)$ and we have assumed a flat prior on (β_0, β_1), its posterior is obtained by just reversing the role of (β_0, β_1) and $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ *i.e.* it then follows that the posterior of $\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ is $N_2\left(\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}, \frac{\sigma^2}{S_{xx}}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix}\right)$.

Now we are ready to handle the predictive distribution problem. As the predictive density of Y_h we want the conditional density of $Y_h|\boldsymbol{d}$. (Since the x's are given constants and σ^2 is assumed to be known, we simplify the notation by suppressing them in the given quantities.) In order to get this conditional density, note that the conditional density of $Y_h|(\beta_0, \beta_1, \boldsymbol{d})$ is Normal with mean $\beta_0 + \beta_1 x_h$ and variance σ^2, while the conditional density of $(\beta_0, \beta_1)|\boldsymbol{d}$ is bivariate Normal. Therefore the (marginal) conditional density of $Y_h|\boldsymbol{d}$ must also be Normal, according to the multivariate Normal distribution theory. Thus in order to get this Normal distribution all we need to do is figure out the (marginal) conditional mean $E[Y_h|\boldsymbol{d}]$ and the (marginal) conditional variance $V[Y_h|\boldsymbol{d}]$ of $Y_h|\boldsymbol{d}$.

$$E[Y_h|\boldsymbol{d}] = E[E[Y_h|\beta_0, \beta_1, \boldsymbol{d}]] = E[\beta_0 + \beta_1 x_h|\boldsymbol{d}] = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

and

$$V[Y_h|\boldsymbol{d}] = E[V[Y_h|\beta_0, \beta_1, \boldsymbol{d}]] + V[E[Y_h|\beta_0, \beta_1, \boldsymbol{d}]] = E\left[\sigma^2|\boldsymbol{d}\right] + V[\beta_0 + \beta_1 x_h|\boldsymbol{d}]$$

$$= \sigma^2 + V[\beta_0|\boldsymbol{d}] + x_h^2 V[\beta_1|\boldsymbol{d}] + 2x_h Cov(\beta_0, \beta_1|\boldsymbol{d}) = \frac{\sigma^2}{S_{xx}}\left[S_{xx} + \frac{1}{n}\sum_{i=1}^n x_i^2 + x_h^2 - 2x_h\overline{x}\right]$$

$$= \frac{\sigma^2}{S_{xx}}\left[S_{xx} + \frac{1}{n}\left\{\sum_{i=1}^n x_i^2 - n\overline{x}^2\right\} + (x_h - \overline{x})^2\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_h - \overline{x})^2}{S_{xx}}\right]$$

Thus the predictive density of $Y_h|\boldsymbol{d} \sim N\left(\hat{\beta}_0 + \hat{\beta}_1 x_h, \sigma^2\left[1 + \frac{1}{n} + \frac{(x_h - \overline{x})^2}{S_{xx}}\right]\right)$ $\bigtriangledown$

7 Prior Distribution

Now we have come to the final section of these notes. We will wrap these Bayesian notes up after discussing the crucial issue of choosing the prior distribution for a given inference

problem. Even among the Bayesians there are at least two schools of believers. The first school, called the *subjective Bayesians* strongly advocate that a prior should only reflect one's subjective belief about the uncertainty in $\boldsymbol{\theta}$ and thus the prior must be a proper distribution on $\boldsymbol{\Theta}$. In deriving this prior one uses the laws of subjective probability. The laws of subjective probability are basically same as the ones one is usually familiar with, however they are derived from scratch from subjective considerations. These laws of subjective probability have been derived in Appendix B of these notes. One can then subjectively specify the prior by drawing either a histogram or a c.d.f.. Then one can proceed towards the posterior computation using these raw priors using numerical methods. We shall take up the issue of numerical computation of posterior using MCMC separately. A second more popular approach is to subjectively specify the prior quantiles or moments and then match these quantities with those of a standard distribution. These standard distributions which are used as priors for a given model, are chosen in such a manner that the posterior becomes analytically tractable. Such priors, when they exist for a given probability model, are called conjugate priors. We shall discuss these conjugate priors in detail for the standard probability models in §7.1.

The second school of Bayesians called the *objective Bayesians* insist on providing an objective analysis even for the cases where a lot of subjective prior information is available, as a supplementary analysis, if not for anything else, at least for the purpose of comparison of how much the analysis got distorted because of one's subjective inputs. Such objective Bayesian analyzes are performed using non-informative priors. Still now there is no clear-cut automatic method of determining an appropriate non-informative prior for a given situation, although some guidelines may be provided in choosing them in certain cases. Some such methods and results are discussed in §7.2.

7.1 Conjugate Priors

As mentioned above, conjugate priors are those which yield analytically tractable posterior distributions. Loosely speaking, a conjugate prior $\pi(\boldsymbol{\theta})$ for a likelihood $L(\boldsymbol{\theta}|\boldsymbol{y})$ is such that both the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and the prior $\pi(\boldsymbol{\theta})$ belong to the same family of distributions. The form of a conjugate prior is usually determined after studying the likelihood function. A few examples will help clarify the concept.

Example 1 (Continued): For the case of $Y_1, Y_2, \ldots, Y_n$ i.i.d. Bernoulli(π) which has the p.m.f. $p(y|\pi) = \pi^y(1-\pi)^{1-y}$ for $y = 0, 1$ the likelihood function $L(\pi|\boldsymbol{y}) = \pi^t(1-\pi)^{n-t}$ where $T(\boldsymbol{Y}) = \sum_{i=1}^n Y_i$ is the sufficient statistic, and t is its observed value. A simple look at this likelihood is enough to come up with the conjugate prior for π, which is given by a Beta distribution Beta(α, β) with $\alpha, \beta > 0$ having p.d.f. $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}$ for $0 < \pi < 1$. Using (8) (or (A5) it is then clear that the posterior of π under this β-prior is Beta($\alpha+t, \beta+n-t$), falling in the same family. Thus Beta distribution is a conjugate prior for the problem of unknown proportion π. $\bigtriangledown$

Example 6 (Continued): For $Y_1, Y_2, \ldots, Y_n$ i.i.d. $N(\mu, \sigma^2)$ the likelihood function of (μ, τ), where $\tau = 1/\sigma^2$ is the precision parameter, is given in (9) in page 19. This likelihood

reveals that conditional on τ if one puts a $N(\theta, 1/(\psi\tau))$ on μ and then puts a Gamma(α, λ) prior on τ then these can be absorbed in the r.h.s of (9) with the posterior having the same prior form. That is let

$$\pi(\mu, \tau) = \pi(\mu|\tau)\pi(\tau) = \frac{\lambda^{\alpha}\psi^{1/2}}{\Gamma(\alpha)\sqrt{2\pi}}\tau^{\alpha-1/2}\exp\left\{-\frac{1}{2}\tau\left[2\lambda + \psi\left(\mu-\theta\right)^2\right]\right\} \tag{16}$$

The prior given in (16) is called the Normal-Gamma prior. Now by (9) and (8) the posterior of (μ, τ) under this Normal Gamma prior is given by

$$\pi(\mu, \tau|\boldsymbol{y}) \propto \tau^{\nu/2+\alpha}\exp\left\{-\frac{1}{2}\tau\left[\left(2\lambda + \nu s_{n-1}^2\right) + (n+\psi)\left(\mu - \frac{n\overline{y}+\psi\theta}{n+\psi}\right)^2\right]\right\} \tag{17}$$

which reveals that given τ the posterior of mu is $N\left(\frac{n\overline{y}+\psi\theta}{n+\psi}, \frac{1}{(n+\psi)\tau}\right)$ and the marginal posterior of τ is Gamma$\left(\alpha + n/2, \lambda + \nu s_{n-1}^2/2\right)$ and thus the posterior is also of the Normal-Gamma form. Thus the Normal-Gamma prior given in (16) is a conjugate prior for (μ, τ).

A couple of remarks regarding the posterior with this informative prior is in order. First note that for known or given τ (and thus σ^2) the mean of the Normal posterior of μ is a weighted average of the sample mean and the prior mean, where the weight of the sample mean is the sample size n and that of the prior mean is prior precision divided by the model precision. This intuitively looks very appealing. For priors with larger uncertainties about the value of μ will have smaller prior precision, where the degree of smallness is judged against the backdrop of model precision in terms of ψ. In such situations the location of the posterior of μ is going to be largely determined by the sample mean. On the other hand with sharp prior information about μ expressed in terms large prior precision, the location of the posterior of μ is going to be largely dictated by its prior value θ. The posterior variance of μ is also very nicely interpretable. It is same as the harmonic total (we got into this harmonic business, because we are working in the transformed reciprocal scale in terms of precision) of the sample variance of the sample mean and prior variance of mu. Thus the posterior precision increases if either the sample size or the prior precision increases. Similar interpretations can be made with the posterior of μ after integrating out τ, which will again be a t distribution but with possibly a non-integer degree of freedom and location and scale parameters a judicious mixture of the prior information and the information carried by the data. $\bigtriangledown$

Example 10: Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. Poisson(λ). Then its likelihood function is proportional to $\lambda^t e^{-n\lambda}$, where $t = \sum_{i=1]^n} y_i$ which immediately reveals that a conjugate prior for λ can be obtained using a Gamma prior. Thus if $\pi(\lambda) \propto \lambda^{\alpha-1}e^{-\lambda\beta}$, then the posterior of λ is Gamma$(\alpha + t, \beta + n)$. The point estimate of λ under squared error loss will then be $\frac{\alpha+t}{\beta+n}$, which again has the same nice interpretation as in the above example. $\bigtriangledown$

7.2 Non-Informative Priors

As mentioned above such priors are not very easy to specify for an arbitrary problem. However satisfactory solutions can be given for some special cases, leading to an approximate

general solution for regular models. However here we shall confine ourselves to the case of a scalar parameter θ, because the solution for the multi-parameter case, though more or less well-understood is way beyond the scope of these notes.

7.2.1 Location Parameters

Let $f(y|\theta)$, the p.d.f. of observable Y be of the form $f(y-\theta)$. Such families of distributions are said to belong to location families of distributions, and θ is called a location parameter. Now suppose instead of observing Y one observes $X = Y + c$ for some constant c. Then the p.d.f. of X is same as $f(x-\eta)$ where $\eta = \theta + c$. One way of interpreting X is that it is same as observing the original Y except now in a different unit of measurement, like say for example Centigrade and Kelvin. Since both the problems (Y, θ) and (X, η) are of identical structure, $\pi_\eta(A) = \pi_\theta(A)\ \forall A \subseteq \mathbf{\Theta}$, where $\pi_\phi(A)$ denotes the prior probability of the set A given by the prior of parameter ϕ. Now $\pi_\eta(A) = \pi_\theta(A - c)$, where $A - c = \{a - c : a \in \mathbf{\Theta}\}$. Thus $\pi_\theta(A) = \pi_\theta(A - c)$ implying

$$\int_A \pi(\theta)\ d\theta = \int_{A-c} \pi(\theta)\ d\theta = \int_A \pi(\theta - c)\ d\theta$$

where $\pi(\theta)$ is prior p.d.f. of θ. Note that, the above equality must hold $\forall A$. Now that can happen if and only if $\pi(\theta) = \pi(\theta - c)$. Now since c is arbitrary, setting it equal to θ yields $\pi(\theta) = \pi(0)$, a constant.

Example 6 (Continued): If $Y \sim N(\mu, \sigma^2)$, with known σ^2, then the distribution of Y falls in a location family. Thus a reasonable (location invariant, to be precise) non-informative prior for μ is given by $\pi(\mu) \propto 1$ for $-\infty < \mu < \infty$. $\triangledown$

7.2.2 Scale Parameters

Let $f(y|\theta)$, the p.d.f. of observable Y be of the form $(1/\theta)f(y/\theta)$. Such families of distributions are said to belong to scale families of distributions, and θ is called a scale parameter. Now suppose instead of observing Y one observes $X = cY$ for some constant c. Then a simple change of variable yields the p.d.f. of X as $(1/\eta)f(x/\eta)$ where $\eta = c\theta$. One way of interpreting X is that it is same as observing the original Y except now in a different unit of measurement, like say for example foot and meter. Since both the problems (Y, θ) and (X, η) are of identical structure, $\pi_\eta(A) = \pi_\theta(A)\ \forall A \subseteq \mathbf{\Theta}$, where $\pi_\phi(A)$ denotes the prior probability of the set A given by the prior of parameter ϕ. Now $\pi_\eta(A) = \pi_\theta(c^{-1}A)$, where $c^{-1}A = \{c^{-1}a : a \in \mathbf{\Theta}\}$. Thus $\pi_\theta(A) = \pi_\theta(c^{-1}A)$ implying

$$\int_A \pi(\theta)\ d\theta = \int_{c^{-1}A} \pi(\theta)\ d\theta = \int_A \pi(c^{-1}\theta)c^{-1}\ d\theta$$

where $\pi(\theta)$ is prior p.d.f. of θ. Note that, the above equality must hold $\forall A$. Now that can happen if and only if $\pi(\theta) = c^{-1}\pi(c^{-1}\theta)$. Now since c is arbitrary, setting it equal to θ yields $\pi(\theta) = \theta^{-1}\pi(1)$. Since $\pi(1)$ is a constant this argument gives a non-informative prior for a scale parameter θ as $\pi(\theta) \propto 1/\theta$.

Example 6 (Continued): If $Y \sim N(\mu, \sigma^2)$, with known μ, then the distribution of $Y - \mu$ falls in a scale family. Thus a reasonable (scale invariant, to be precise) non-informative prior for μ is given by $\pi(\mu) \propto 1$ for $-\infty < \mu < \infty$. $\bigtriangledown$

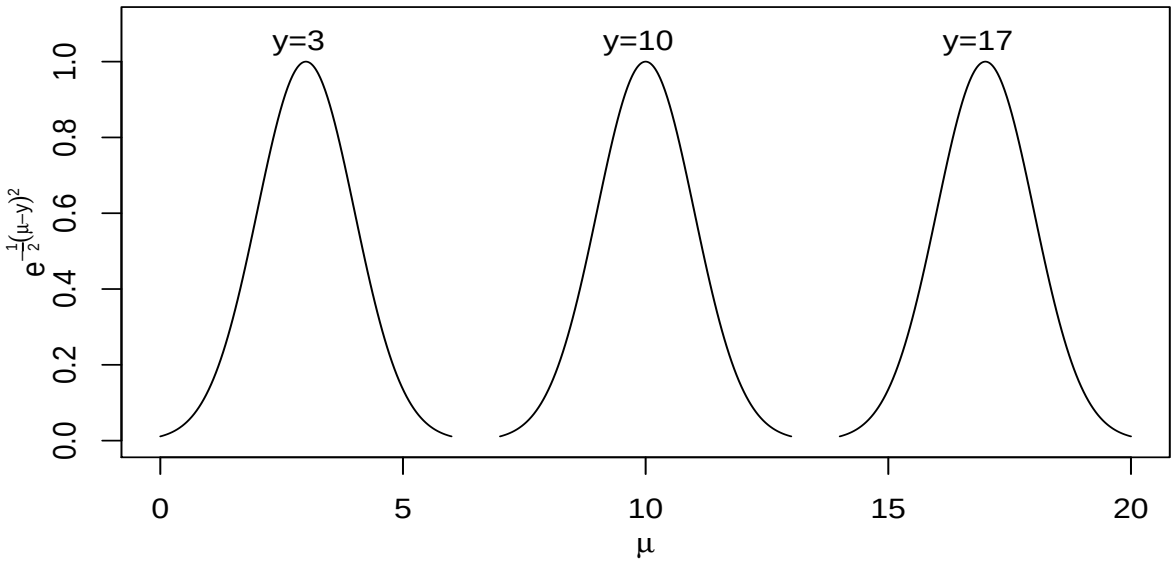
Example 11: Let $Y \sim$exp(λ), the exponential distribution with failure rate λ. For this Y its p.d.f. is given by $\lambda e^{-\lambda y} = (1/\theta) f(y/\theta)$, where $\theta = 1/\lambda$ and $f(z) = e^{-z}$. Thus this belongs to a scale family, and thus a non-informative prior for θ is given by $\pi(\theta) \propto 1/\theta$, which after change of variable yields the prior for λ as $\pi(\lambda) \propto 1/\lambda$. $\bigtriangledown$

7.2.3 Data Translated Likelihood

Since the time of Bayes and Laplace, since intuitively uniform prior appears to be a natural choice as a non-informative prior, it is worth examining the situations where such uniform prior might be appropriate as a non-informative prior. Let the likelihood function, written as a function of $\phi(\theta)$, where θ is the original parameter of interest, be such that it is completely known *apriori* except for its location, which depends on the data yet to be observed.

For example consider the likelihood function of μ for the Normal model given in (9) for known $\sigma^2 = \tau = 1$, $n = 1$ and yet to observe data $Y = y$. The likelihood function is $\propto \exp\{-(1/2)(\mu - y)^2\}$ and is exactly known how it will look (as a function of μ) before observing the data, except its location. This situation is depicted in Figure 5 below, where we have plotted the likelihood function for 3 values of y, namely 3, 10 and 17.

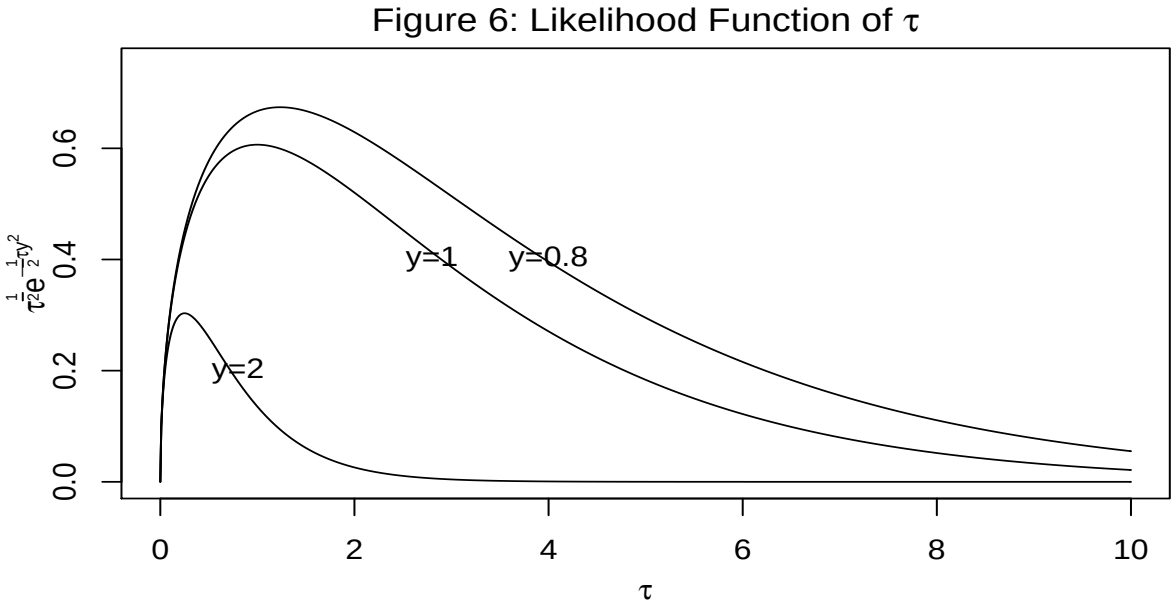
Figure 5: Likelihood Function of μ



All the three likelihood functions are identical except their location, which of course depends on the value of y. When we have a situation of this type, where the only effect the data has got is to translate the likelihood, written as a function of $\phi(\theta)$, to a proper location, then such a likelihood is called **data translated** w.r.t. $\phi(\theta)$. In the above example the likelihood is data translated w.r.t the original parameter μ.

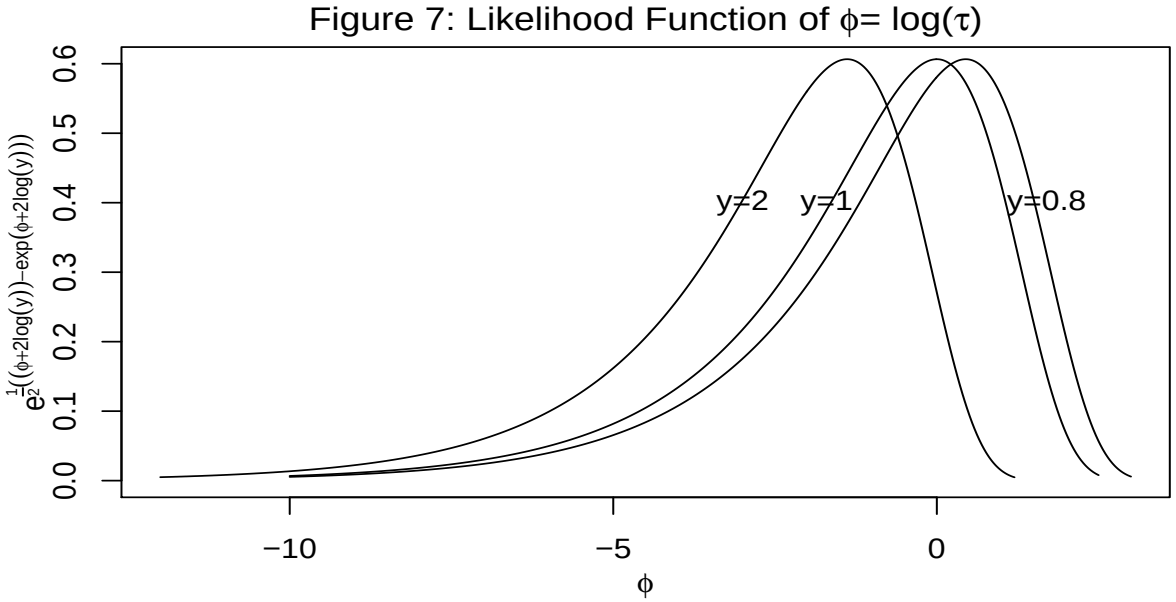
Now again consider the likelihood function (9), this time for known $\mu = 0$ and $n = 1$. The likelihood function, as function of τ is $\propto \tau^{1/2} \exp\{-(1/2)\tau y^2\}$. This likelihood function, as

function of τ, has been plotted for 3 values of y, namely 0.8, 1 and 2 in Figure 6 below.



This time of course the likelihood is not data translated w.r.t. τ as its shape is not known *apriori* before observing the data.

Now re-consider this case again, but this time consider the likelihood function as function of $\phi = \log \tau$. $L(\phi|y) \propto \exp\{(1/2)(\phi + 2\log y) - (1/2)\exp(\phi + 2\log y)\}$. $L(\phi)$ has been plotted in Figure 7 below for the same 3 values of y *viz.* 0.8, 1 and 2.



Now it is clear from the above plot, as well as algebraically from the expression of $L(\phi|y)$, that this likelihood is data translated w.r.t. ϕ.

When the likelihood is data translated w.r.t. $\phi(\theta)$, that means the exact shape of the likelihood is known *apriori* even before the data is observed, and the only role the data plays is translate the likelihood to an appropriate place and thus pin-pointing its location. In such a situation to say we know a little *apriori* relative to what the data is going to tell

us, may be expressed by saying that we are almost equally willing to accept one value of $\pi(\theta)$ as another. This state of indifference may be expressed by imposing a uniform prior on $\phi(\theta)$, so that a non-informative prior for θ is given by $\pi(\theta) \propto \phi'(\theta)$. Thus for the last example since it was found that the likelihood is data translated w.r.t. ϕ=$\log \tau$, we should put a uniform prior on ϕ. Thus by change of variable, a non-informative prior for τ should be $\propto 1/\tau$, which agrees with the prior we have already obtained in 7.2.2. In general it is fairly hard to come up with a $\phi(\cdot)$ such that the likelihood, as a function of ϕ, is going to be exactly data translated. However for regular models it is possible to find such a ϕ such that the likelihood is at least approximately data translated, so that a non-informative prior for θ may be found. This method is discussed in the next sub-section.

7.2.4 Jeffreys' Prior

Consider the likelihood function $L(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\theta)$ for a regular[14] model. Let $\ell(\theta) = \log L(\theta|\boldsymbol{y})$ and $\hat{\theta}$ denote the Maximum Likelihood Estimator (MLE) of θ. Since it is well known that $\hat{\theta} \xrightarrow{P} \theta$, ignoring terms of third order or more in the Taylor series expansion of $\ell(\theta)$ about $\hat{\theta}$ one can write,

$$L(\theta|\boldsymbol{y}) = \exp\{\ell(\theta)\} \approx \exp\left\{\ell(\hat{\theta}) - (1/2)n(\theta-\hat{\theta})^2 J(\boldsymbol{y},\hat{\boldsymbol{\theta}})\right\} \approx\propto \exp\left\{-(1/2)n(\theta-\hat{\theta})^2 I(\hat{\theta})\right\} \tag{18}$$

where $J(\boldsymbol{y},\theta) = -\frac{1}{n}\frac{d^2}{d\theta^2}\ell(\theta)$ and $I(\theta) = E_\theta\left[-\frac{d^2}{d\theta^2}\log f(Y|\theta)\right]$ [15] with $J(\boldsymbol{y},\hat{\theta})$ and $I(\hat{\theta})$ being the values of $J(\boldsymbol{y},\theta)$ and $I(\theta)$ respectively evaluated at $\theta = \hat{\theta}$. A couple of explanations are required for understanding the above string of equations in (18). First of all the second term $(\theta-\hat{\theta})\,\frac{d}{d\theta}\ell(\theta)\Big|_{\theta=\hat{\theta}}$ in the Taylor series expansion of $\ell(\theta)$ about $\hat{\theta}$ is missing because by virtue of $\hat{\theta}$ being the MLE, this derivative is 0. The $\propto$ reflects dropping of $e^{\ell(\hat{\theta})}$ in the final expression. And finally, since $J(\boldsymbol{y},\theta) = -\frac{1}{n}\sum_{i=1}^{n}.\frac{d^2}{d\theta^2}\log f(y_i|\theta)$, by the law of large numbers, $J(\boldsymbol{y},\theta) \xrightarrow{P} I(\theta)$.

Above arguments justify approximating a regular likelihood by a Normal p.d.f. with mean $\hat{\theta}$ and standard deviation $n^{-1/2}I^{-1/2}(\hat{\theta})$. Obviously this likelihood is not data translated w.r.t. θ, because both the mean and standard deviation of the approximating Normal density depend on the data through $\hat{\theta}$. And with the approximations over, we are now in search of

[14] Formally a model is called **regular** if it satisfies the regularity conditions required for satisfying the Cramer-Rao lower bound. These regularity conditions ensure that derivatives of $f(y|\theta)$ w.r.t. θ exist at least up to the second order and one can freely interchange these derivatives w.r.t.θ and integrals w.r.t. y whenever required.

[15] The quantity $I(\theta)$ is called **Fisher Information** of a model and plays a very critical role in frequentist statistics. Intuitively, $I(\theta)$ gives the negative of the expected curvature of the log-likelihood for a single observation. It can be shown that under the regularity conditions that $I(\theta) > 0\ \forall\theta \in \Theta$. Thus the negative sign only ensures that we are dealing with a positive quantity that is to be called "Information". A large value of $I(\theta)$ indicates that the likelihood function is sharply pointed. In which case the model is very informative about θ. On the other hand if the likelihood is flat, then the model carries very little information about θ, and in this case the value of $I(\theta)$ will be very close to 0.

a $\phi(\theta)$ for which it will be. Let $\theta \overset{1-1}{\longleftrightarrow} \phi(\theta)$. Then

$$\begin{aligned} & I(\phi) \\ &= E_\phi\left[-\frac{d^2}{d\phi^2}\log f(Y|\phi)\right] \\ &= E_\phi\left[-\frac{d}{d\phi}\left\{\frac{d}{d\theta}\log f(Y|\theta)\frac{d\theta}{d\phi}\right\}\right] \\ &= -E_\phi\left[\frac{d^2}{d\theta^2}\log f(Y|\theta)\left(\frac{d\theta}{d\phi}\right)^2 + \frac{d\theta}{d\phi}\frac{d}{d\theta}\log f(Y|\theta)\right] \\ &= -\left(\frac{d\theta}{d\phi}\right)^2 E_\theta\left[\frac{d^2}{d\theta^2}\log f(Y|\theta)\right] \end{aligned}$$

The last equality follows because $E_\theta\left[\frac{d}{d\theta}\log f(Y|\theta)\right] = \int_{-\infty}^{\infty}\frac{d}{d\theta}f(y|\theta)\,dy = \frac{d}{d\theta}\int_{-\infty}^{\infty}f(y|\theta)\,dy = \frac{d1}{d\theta} = 0$. Also note that in the above we have freely switched from ϕ to θ in $f(Y|\phi)$ to $f(Y|\theta)$ or $E_\theta[\cdot]$ from $E_\phi[\cdot]$, because when viewed as an argument of a function as in these, it does not matter which one one writes as $\phi \leftrightarrow \theta$ is 1-1. Thus we get the result that

$$I(\phi) = I(\theta)\left(\frac{d\theta}{d\phi}\right)^2. \tag{19}$$

Now if one chooses a $\phi(\theta)$ such that $\left|\frac{d\theta}{d\phi}\right| \propto I^{-1/2}(\theta)$ or $\left|\frac{d\phi}{d\theta}\right| \propto I^{1/2}(\theta)$ or $\phi(\theta) = \int I^{1/2}(\theta)\,d\theta$, then by (19), $I(\phi)$ becomes a constant free of ϕ. Then in terms of the transformed parameter $\phi(\theta)$, by (18), $L(\phi|\boldsymbol{y})$, the likelihood as a function of ϕ, is approximately $\propto \exp\left\{-(c/2)n(\phi-\hat{\phi})^2\right\}$, where $\hat{\phi}$ is the MLE of ϕ. Now since this approximate likelihood of ϕ, $\exp\left\{-(c/2)n(\phi-\hat{\phi})^2\right\}$ is completely known *apriori* before observing the data, except its exact location, which is determined by $\hat{\phi}$, it is data translated w.r.t. ϕ. Hence following the arguments provided in 7.2.3, a uniform prior appears to be a reasonable choice as a non-informative prior for ϕ *i.e.* a non-informative prior for ϕ is given by $\pi(\phi) \propto 1$. Now since ϕ is such that $\left|\frac{d\phi}{d\theta}\right| \propto I^{1/2}(\theta)$ a simple change of variable yields a non-informative for θ as

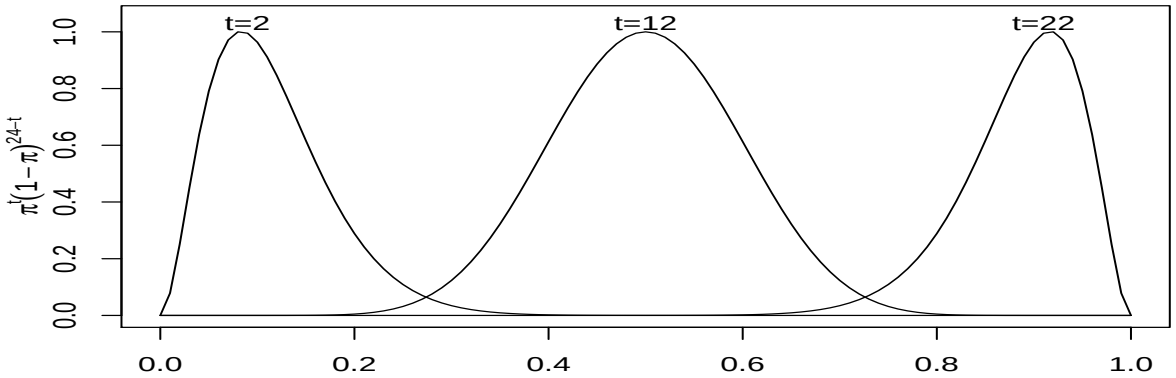
$$\pi_J(\theta) \propto I^{1/2}(\theta) \tag{20}$$

The original argument of Jeffreys which led him to the prior in (20) was a little different, which is as follows. An issue that plagues the problem of choosing a reasonable non-informative prior is that of reparameterization. That is if one imposes a prior $\pi_\theta(\theta)$ on θ and if $\theta \overset{1-1}{\longleftrightarrow} \phi$ then by the change of variable formula, the assumed prior for ϕ induced by $\pi_\theta(\theta)$ is given by $\pi_\phi(\phi) = \pi_\theta(\theta)\left|\frac{d\theta}{d\phi}\right|$. Thus if some principle of choice led one to $\pi_\theta(\theta)$ as a non-informative prior for θ, then the same principle applied to ϕ should now yield a $\pi_\phi(\phi)$ satisfying $\pi_\phi(\phi) = \pi_\theta(\theta)\left|\frac{d\theta}{d\phi}\right|$. Otherwise the principle is inconsistent. For the Jeffreys' prior in (20), $\pi_\phi(\phi) \propto I^{1/2}(\phi)$, which by (19) equals, $I^{1/2}(\theta)\left|\frac{d\theta}{d\phi}\right| \propto \pi_\theta(\theta)\left|\frac{d\theta}{d\phi}\right|$. A prior enjoying this property is called invariant under parameter transformation, which is a minimal requirement for any non-informative prior to be qualified to be called "reasonable".

A uniform prior, as a principle of choice for instance, does not have such a property. $\pi(\theta) \propto 1 \Rightarrow \pi(\phi) \propto \phi^{-1/2}$ for $\phi = \theta^2$, or $\pi(\phi) \propto e^{\phi}$ for $\phi = \log\theta$, or $\pi(\phi) \propto 1/\phi^2$ $\phi = 1/\theta$, none of which is again uniform. But if θ is such that $I^{1/2}(\theta) \propto 1$, then in terms of any reparameterization like $\phi = \theta^2$ or $\phi = \log\theta$ or $\phi = 1/\theta$, the corresponding Jeffreys' prior for these transformed parameters will respectively be proportional to $\phi^{-1/2}$, e^{ϕ} or $1/\phi^2$.

Example 1 (Continued): Here $p(y|\pi) = \pi^y(1-\pi)^{1-y}$. Thus $\log p(y|\pi) = y\log\pi + (1-y)\log(1-\pi)$. Hence $\frac{d^2}{d\pi^2}\log p(Y|\pi) = \frac{d}{d\pi}\left(\frac{Y}{\pi} + \frac{1-Y}{1-\pi}\right) = -\left[\frac{Y}{\pi^2} + \frac{1-Y}{(1-\pi)^2}\right] = -\frac{\pi^2 - 2Y\pi + Y}{\pi^2(1-\pi)^2}$. Therefore since $E_\pi[Y] = \pi$, $I(\pi) = E_\pi\left[-\frac{d^2}{d\pi^2}\log p(Y|\pi)\right] = \frac{\pi(1-\pi)}{\pi^2(1-\pi)^2} = \pi^{-1}(1-\pi)^{-1}$. Thus by (20), Jeffreys' prior for π is given by $\pi_J(\pi) \propto \pi^{-1/2}(1-\pi)^{-1/2}$, $0 < \pi < 1$. Note that $\pi_J(\pi)$ is improper. In order to see why this is a reasonable non-informative prior, let us consider the likelihood function $L(\pi|\boldsymbol{y}) = \pi^t(1-\pi)^{n-t}$, where $t = \sum_{i=1}^n y_i$. Standardized $L(\pi|\boldsymbol{y})$, as function of π for $n = 24$ is plotted for $t = 2$, 12 and 22 in Figure 8 below.

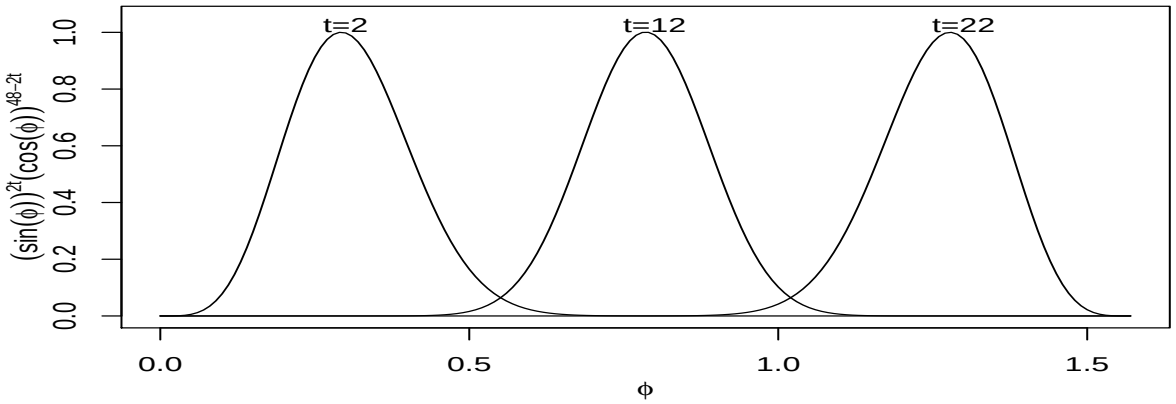
Figure 8: Likelihood Function of π



Obviously the likelihood is not data translated w.r.t. π. The three different likelihoods have three different shapes. The likelihoods for $t = 2$ is positively skewed, for $t = 12$ is symmetric while for $t = 22$ is negatively skewed.

In the derivation of Jeffreys' prior it was seen that the parameter w.r.t. which the likelihood is approximately data translated is given by $\phi = \int I^{1/2}(\theta)\, d\theta$. Thus in this case consider the transformed parameter $\phi = \int \pi^{-1/2}(1-\pi)^{-1/2}\, d\pi \propto \sin^{-1}\sqrt{\pi}$. Writing the likelihood function in terms of ϕ we get $L(\phi|\boldsymbol{y}) = \sin^{2t}\phi\cos^{2(n-t)}\phi$ for $\phi \in [0, \pi/2]$. Now the three standardized likelihoods in terms of ϕ is plotted for $t = 2$, 12 and 22 in Figure 9 below.

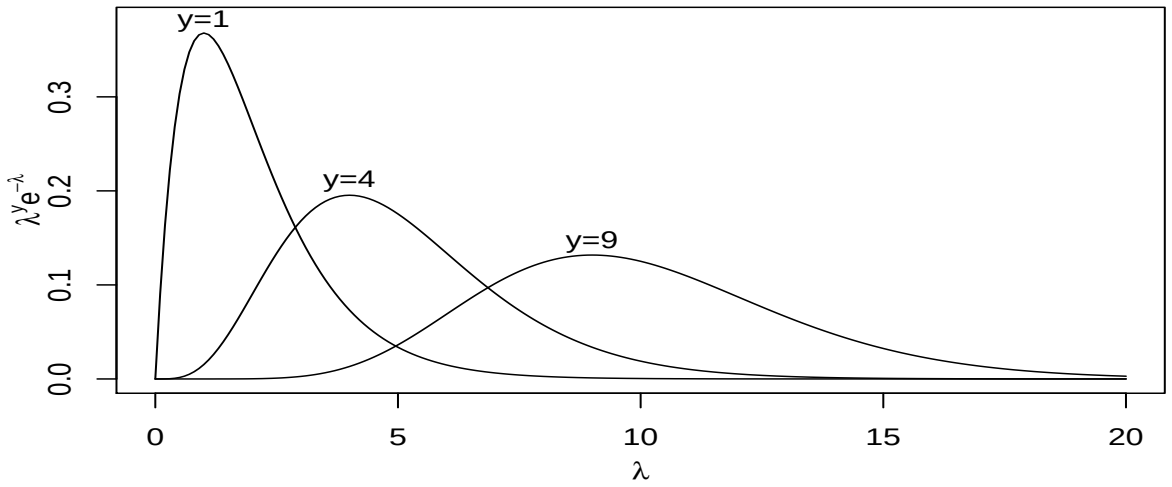
Figure 9: Likelihood Function of ϕ



The likelihoods in Figure 9 appear to be data translated and thus a uniform prior for $\sin^{-1}\sqrt{\pi}$ deem to be quite appropriate. This uniform prior on $\sin^{-1}\sqrt{\pi}$ implies that the Jeffreys' prior $\pi_J(\pi) \propto \pi^{-1/2}(1-\pi)^{-1/2}$ on π. $\bigtriangledown$

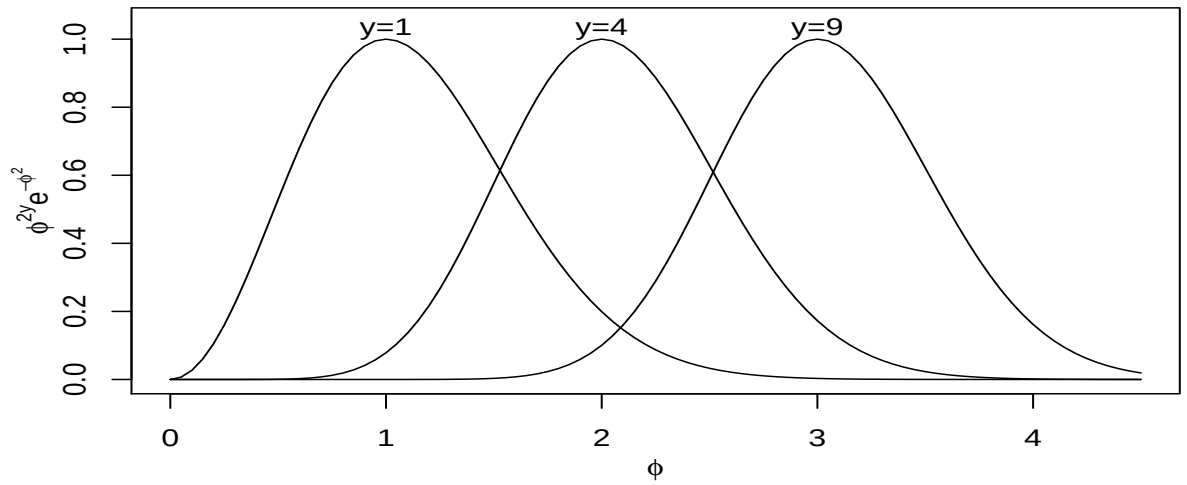
Example 10 (Continued): If $Y \sim$Poisson(λ), $\log p(y|\lambda) = c + y\log\lambda$ - λ. Hence $\frac{d^2}{d\lambda^2}\log p(Y|\lambda)$ $= \frac{d}{d\lambda}(Y/\lambda - 1) = -Y/\lambda^2$. Therefore since $E_\lambda[Y] = \lambda$, $I(\lambda) = E_\lambda\left[-\frac{d^2}{d\lambda^2}\log p(Y|\lambda)\right] = \lambda^{-1}$. Thus by (20), Jeffreys' prior for λ is given by $\pi_J(\lambda) \propto \lambda^{-1/2}$, $\lambda > 0$. Note that again $\pi_J(\lambda)$ is improper. The likelihood, as a function of λ is plotted in Figure 10 below for $n = 1$ and $y = 1$, 4 and 9.

Figure 10: Likelihood Function of λ



As expected the likelihood is not data translated w.r.t. λ. But now when the likelihood as a function of $\phi = \int I^{1/2}(\lambda)\, d\lambda = \int \lambda^{-1/2}\, d\lambda \propto \sqrt{\lambda}$, written as $L(\phi|\boldsymbol{y}) \propto \phi^{2y}e^{-\phi^2}$ is plotted for the same 3 values of y as in Figure 11 below:

Figure 11: Likelihood Function of ϕ



it suddenly appears to be data translated. Thus a uniform prior on $\sqrt{\lambda}$ appears to be reasonable, leading to the Jeffreys' prior $\pi_J(\lambda) \propto \lambda^{-1/2}$ on λ. $\bigtriangledown$

Thus in an essence Jeffreys' prior catches hold of a parametric function $\phi(\theta)$ w.r.t. which the likelihood becomes approximately data translated and thus imposes a uniform prior on ϕ which in turn induces a prior on θ. This whole process is automatically contained in formula

(20), and for every problem one need not explicitly work out the nature of ϕ for which the uniform prior is a good choice. The derivations of ϕ such as $\sin^{-1}\sqrt{\pi}$ and $\sqrt{\lambda}$ in the above examples are only for illustrative purpose.

For the multi-parameter case, $\boldsymbol{\theta}_{p\times 1} = (\theta_1, \theta_2, \ldots, \theta_p)'$. In this case the Fisher information is not a scalar but a $p \times p$ matrix $\boldsymbol{I}(\boldsymbol{\theta}) = ((I_{ij}(\boldsymbol{\theta})))_{p\times p}$, called the Fisher Information Matrix, with $I_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(Y|\boldsymbol{\theta})\right]$. For the multi-parameter case Jeffreys' prior is given by

$$\pi_J(\boldsymbol{\theta}) \propto |\boldsymbol{I}(\boldsymbol{\theta})|^{1/2} \tag{21}$$

While for the single parameter case, Jeffreys' prior is well-accepted as a reasonable non-informative prior, it is not so for the case of presence of nuisance parameters. However discussions for this case is well beyond the scope of these notes and we shall finish our discussion by showing an illustration of computation of Jeffreys' prior in the multi-parameter case.

Example 6 (Continued): For $Y \sim N(\mu, \sigma^2)$, $\log f(Y|\mu,\sigma^2) = c - \log\sigma - \frac{1}{2\sigma^2}(Y-\mu)^2$. Thus $\frac{\partial}{\partial\mu}\log f(Y|\mu,\sigma^2) = (Y-\mu)/\sigma^2$ and $\frac{\partial}{\partial\sigma}\log f(Y|\mu,\sigma^2) = -1/\sigma + (Y-\mu)^2/\sigma^3$, so that $\frac{\partial^2}{\partial\mu^2}\log f(Y|\mu,\sigma^2) = -1/\sigma^2$, $\frac{\partial^2}{\partial\mu\partial\sigma}\log f(Y|\mu,\sigma^2) = \frac{\partial^2}{\partial\sigma\partial\mu}\log f(Y|\mu,\sigma^2) = -2(Y-\mu)/\sigma^3$ and $\frac{\partial^2}{\partial\sigma^2}\log f(Y|\mu,\sigma^2) = 1/\sigma^2 - 3(Y-\mu)^2/\sigma^4$. Thus since $E_{\mu,\sigma}[Y-\mu] = 0$ and $E_{\mu,\sigma}[(Y-\mu)^2] = \sigma^2$,

$$\boldsymbol{I}(\mu,\sigma) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{bmatrix}$$

Thus the Jeffreys' prior in this case is given by $\pi(\mu,\sigma) \propto |\boldsymbol{I}(mu,\sigma)|^{1/2} \propto 1/\sigma^2$. Note that it is slightly different from the standard non-informative prior for this case given in (7). The prior in (7) can be obtained by assuming that μ and σ are independent *apriori* and then multiplying their respective non-informative priors obtained in §7.2.1 and §7.2.2. This just goes on to show the problems associated with Jeffreys' prior in the multi-parameter case.▽

Appendix A: Sufficient Statistics

Definition A1: A (possibly vector valued) statistic $\boldsymbol{T}(Y_1, Y_2, \ldots, Y_n)$ is said to be **sufficient** for $\boldsymbol{\theta}$ if the conditional distribution of the original observations $Y_1, Y_2, \ldots, Y_n$ given $\boldsymbol{T}(Y_1, Y_2, \ldots, Y_n) = t$ does not depend on $\boldsymbol{\theta}$.

Sufficiency plays a central role in mathematical statistics. Intuitively, sufficient statistics provide a way of reducing the data without loosing any information about the unknown parameter $\boldsymbol{\theta}$. This is because if one has the value of the sufficient statistic $\boldsymbol{T}$ but the original data set $Y_1, Y_2, \ldots, Y_n$ is lost, one can still reconstruct a set of $Y_1, Y_2, \ldots, Y_n$ (using for example a random number generator) as it does not require knowledge of the unknown $\boldsymbol{\theta}$ (by definition), which is *equivalent* to the original data set in the sense that its probability distribution remains the same as the original data set. Thus if sufficient statistics exist, one need not carry around the entire original raw data set for drawing inference about the model parameters. Just having the values of the sufficient statistics is good enough or sufficient,

as these statistics carry all the relevant information about $\boldsymbol{\theta}$ contained in the observations $Y_1, Y_2, \ldots, Y_n$.

Example A1: Suppose Y_1 and Y_2 are i.i.d. Poisson(λ) *i.e.* we have a sample of size 2 from a Poisson population. Consider the statistic $T = Y_1 + Y_2$.

$$\begin{aligned}
& P(Y_1 = y | T = t) \\
&= \frac{P(Y_1 = y, Y_2 = t - y)}{P(T = t)} \\
&= \frac{P(Y_1 = y) P(Y_2 = t - y)}{e^{-2\lambda}(2\lambda)^t / t!} \\
&\quad \text{(because } Y_1 \text{ and } Y_2 \text{ are independent and } T \sim \text{Poisson}(2\lambda)\text{)} \\
&= \frac{e^{-2\lambda}\lambda^{y+t-y}/(y!(t-y)!)}{e^{-2\lambda}(2\lambda)^t/t!} \\
&= \binom{t}{y} \left(\frac{1}{2}\right)^y \left(\frac{1}{2}\right)^{t-y},
\end{aligned}$$

which does not depend on the unknown population parameter λ. It should now be easy to see that if we had $Y_1, Y_2, \ldots, Y_n$ a sample of size n from a Poisson(λ) population and $T = Y_1 + Y_2 + \cdots Y_n$,

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | T = t) = \frac{t!}{y_1! y_2! \cdots y_n!} \left(\frac{1}{n}\right)^{y_1} \left(\frac{1}{n}\right)^{y_2} \cdots \left(\frac{1}{n}\right)^{y_n},$$

which does not depend on λ. Thus according to the definition $T = \sum_{i=1}^n Y_i$ is a sufficient statistic for a Poisson sample. This is because if one has the value of T as t, one can reconstruct a version of the original sample by generating a set of values from a Multinomial($t; \frac{1}{n}, \ldots, \frac{1}{n}$) distribution, without bothering to carry around all the n $Y_1, Y_2, \ldots, Y_n$ values. $\bigtriangledown$

Now trying to intuitively guess, obtain and then show a statistic is sufficient from definition, as has been done in example A1 above, is an arduous if not an impossible task. Fortunately there is a theorem, called the **Factorization Theorem**, which helps one obtain a sufficient statistics in a routine manner from the expression of the p.m.f./p.d.f. of a probability model. Before presenting the Factorization Theorem theorem let us have a re-look at the definition of the **likelihood function**. Though the concept of likelihood function has already been introduced in the paragraph preceding equation (8) in page 18 of the text, a formal definition and its interpretation are as follows.

Definition A2: If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with p.d.f. $f(y|\boldsymbol{\theta})$ (or p.m.f. $p(y|\boldsymbol{\theta})$) with realized values $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$ the **likelihood function** of $\boldsymbol{\theta}$ is given by $L(\boldsymbol{\theta}|y_1, y_2, \ldots, y_n) = \prod_{i=1}^n f(y_i|\theta)$ (or $\prod_{i=1}^n p(y_i|\theta)$ in the discrete case).

Very loosely speaking the likelihood function sort of gives the probability of observing the data at hand given a value of the model parameter $\boldsymbol{\theta}$. But since $\boldsymbol{\theta}$ is unknown, we try to view this quantity in its totality as a function of the unknown $\boldsymbol{\theta}$ as it varies over its domain $\boldsymbol{\Theta}$. It is important to realize that in the expression of the likelihood function, the variable of

interest is $\boldsymbol{\theta}$ and not the observed data $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$. It is something akin to a probability only when viewed as a function of $y_1, y_2, \ldots, y_n$, but since the likelihood must be viewed as a function of $\boldsymbol{\theta}$ it is not a probability.

Example A2: A. If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$, with realized values $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, then the likelihood function of (μ, σ^2) is given by

$$L(\mu, \sigma^2 | y_1, y_2, \ldots, y_n) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2} = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\left\{n(\overline{y}-\mu)^2 + \sum_{i=1}^n (y_i - \overline{y})^2\right\}} \tag{A1}$$

B. If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. Bernoulli(π), so that each Y_i is 0-1 valued with probability of assuming the value 1 is π and 0 is $1 - \pi$, which may be expressed as the p.m.f. $p(y|\pi) = \pi^y (1-\pi)^{1-y}$ for $y = 0, 1$, with realized values $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, then the likelihood function of π is given by

$$L(\pi | y_1, y_2, \ldots, y_n) = \pi^{\sum_{i=1}^n y_i} (1-\pi)^{n - \sum_{i=1}^n y_i} \tag{A2}$$

C. If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. Poisson(λ) with realized values $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, then the likelihood function of λ is given by

$$L(\lambda | y_1, y_2, \ldots, y_n) = e^{-n\lambda} \lambda^{\sum_{i=1}^n y_i} / \prod_{i=1}^n y_i! \tag{A3}$$

(Factorization Theorem): If the population random variable Y has p.d.f. $f(y|\boldsymbol{\theta})$ (or p.m.f. $p(y|\boldsymbol{\theta})$) then given the the observed data $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, statistics $\boldsymbol{T}(Y_1, Y_2, \ldots, Y_n)$ is sufficient for $\boldsymbol{\theta}$ if and only if the likelihood function can be factored as ($\boldsymbol{t}$ being the realized value of $\boldsymbol{T}$)

$$L(\boldsymbol{\theta} | y_1, y_2, \ldots, y_n) = g(\boldsymbol{t}(y_1, y_2, \ldots, y_n), \boldsymbol{\theta}) h(y_1, y_2, \ldots, y_n). \tag{A4}$$

That is $\boldsymbol{t}(y_1, y_2, \ldots, y_n)$ is sufficient for $\boldsymbol{\theta} \iff$ the likelihood function can be factored into two components, where the expression of the first component involves $\boldsymbol{\theta}$ and terms involving $y_1, y_2, \ldots, y_n$ appearing only through $\boldsymbol{t}(y_1, y_2, \ldots, y_n)$, and the expression of the second component involves only $y_1, y_2, \ldots, y_n$ without any term involving $\boldsymbol{\theta}$.

Proof: We shall present the proof for discrete Y, which is a little more intuitive and illustrative but less technical than the continuous case.
"only if" or $\Longrightarrow$ part: Suppose $\boldsymbol{T}(Y_1, Y_2, \ldots, Y_n)$ is sufficient for $\boldsymbol{\theta}$ and let $\boldsymbol{t}(y_1, y_2, \ldots, y_n) = \boldsymbol{t}$ denote the observed value of $\boldsymbol{T}$. Then

$$\begin{aligned}
& L(\boldsymbol{\theta} | y_1, y_2, \ldots, y_n) \\
&= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | \boldsymbol{\theta}) \\
&\quad \text{(by definition of likelihood function)} \\
&= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n, \boldsymbol{T} = \boldsymbol{t} | \boldsymbol{\theta}) \\
&\quad \text{(as the two events are same)} \\
&= P(\boldsymbol{T} = \boldsymbol{t} | \boldsymbol{\theta}) P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | \boldsymbol{T} = \boldsymbol{t}, \boldsymbol{\theta}) \\
&\quad \text{(by definition of conditional probability.)} \\
&= g(\boldsymbol{t}, \boldsymbol{\theta}) h(y_1, y_2, \ldots, y_n)
\end{aligned}$$

where $g(\boldsymbol{t},\theta) = P(\boldsymbol{T} = \boldsymbol{t}|\boldsymbol{\theta})$, which involves only $\boldsymbol{\theta}$ and $\boldsymbol{t}$ (without directly involving $y_1, y_2, \ldots, y_n$ - the only way $y_1, y_2, \ldots, y_n$ appear in the expression of $g(\boldsymbol{t},\boldsymbol{\theta})$ is through $\boldsymbol{t}$); and $h(y_1, y_2, \ldots, y_n\ ) = P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|\boldsymbol{T} = \boldsymbol{t},\boldsymbol{\theta})$, which does not involve $\boldsymbol{\theta}$, by definition of the sufficiency of $\boldsymbol{T}$.
"if" or $\Longleftarrow$ part: Suppose the probability model of $Y_1, Y_2, \ldots, Y_n$ is such that it admits the factorization (A4). Then

$$\begin{aligned}
&P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n|\boldsymbol{T} = \boldsymbol{t},\boldsymbol{\theta}) \\
&= \frac{P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n, \boldsymbol{T} = \boldsymbol{t}|\boldsymbol{\theta})}{P(\boldsymbol{T} = \boldsymbol{t}|\boldsymbol{\theta})} \\
&= \frac{g(\boldsymbol{t},\theta)h(y_1, y_2, \ldots, y_n\ )}{\sum_{\{y_1, y_2, \ldots, y_n\ :\boldsymbol{T}(y_1, y_2, \ldots, y_n\ )=\boldsymbol{t}\}} g(\boldsymbol{t},\boldsymbol{\theta})h(y_1, y_2, \ldots, y_n\ )} \\
&= \frac{g(\boldsymbol{t},\boldsymbol{\theta})h(y_1, y_2, \ldots, y_n\ )}{g(\boldsymbol{t},\boldsymbol{\theta})\sum_{\{y_1, y_2, \ldots, y_n\ :\boldsymbol{T}(y_1, y_2, \ldots, y_n\ )=\boldsymbol{t}\}} h(y_1, y_2, \ldots, y_n\ )} \\
&= \frac{h(y_1, y_2, \ldots, y_n\ )}{\sum_{y_1, y_2, \ldots, y_n\ :\boldsymbol{T}(y_1, y_2, \ldots, y_n\ )=\boldsymbol{t}} h(y_1, y_2, \ldots, y_n\ )}
\end{aligned}$$

which does not depend on $\boldsymbol{\theta}$. ▽

Example A2 (Continued): **A. i.** Consider the Normal likelihood given in (A1). Assume that σ^2 is known but not μ. Let $T(Y_1, Y_2, \ldots, Y_n\ ) = \overline{Y}$, so that $t = \overline{y}$. Then the likelihood function (A1) can be factorized into $g(t,\mu) = e^{-n(t-\mu)^2/(2\sigma^2)}$ and $h(y_1, y_2, \ldots, y_n\ ) = (2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\overline{y})^2}$, showing that $\overline{Y}$ is sufficient for μ in a $N(\mu,\sigma^2)$ model for known σ^2.

A. ii. Again consider the Normal likelihood given in (A1). This time assume that μ is known but not σ^2. Let $T(Y_1, Y_2, \ldots, Y_n\ ) = \sum_{i=1}^n (Y_i - \mu)^2$. Note that this T is a statistic because μ is known. In this case define $g(t,\sigma^2) = (2\pi\sigma^2)^{-n/2}e^{-t/(2\sigma^2)}$ and $h(y_1, y_2, \ldots, y_n\ ) = 1$ so that $L(\sigma^2|y_1, y_2, \ldots, y_n\ ) = g(t,\sigma^2)h(y_1, y_2, \ldots, y_n\ )$. Thus in this case $\sum_{i=1}^n (Y_i - \mu)^2$ is sufficient for σ^2.

A. iii. Finally consider the Normal likelihood in (A1) with both (μ,σ^2) unknown. Note that in this case the unknown parameter is vector valued with $\boldsymbol{\theta} = (\mu,\sigma^2)$. In this case we should have a vector valued sufficient statistics $\boldsymbol{T}$. Thus let $\boldsymbol{T} = (\overline{Y}, \sum_{i=1}^n (Y_i - \overline{Y})^2)$, $g(\boldsymbol{t},\boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}\left\{n(\overline{y}-\mu)^2+\sum_{i=1}^n (y_i-\overline{y})^2\right\}}$ and $h(y_1, y_2, \ldots, y_n\ ) = 1$, so that $L(\boldsymbol{\theta}|y_1, y_2, \ldots, y_n\ ) = g(\boldsymbol{t},\boldsymbol{\theta})h(y_1, y_2, \ldots, y_n\ )$. Thus in this case $(\overline{Y}, \sum_{i=1}^n (Y_i - \overline{Y})^2)$ is sufficient for $\boldsymbol{\theta} = (\mu,\sigma^2)$.

B. For the Bernoulli likelihood in (A2) let $T = \sum_{i=1}^n Y_i$, $g(t,\pi) = \pi^t(1-\pi)^{n-t}$ and $h(y_1, y_2, \ldots, y_n\ ) = 1$. Then $L(\pi|y_1, y_2, \ldots, y_n\ ) = g(t,\pi)h(y_1, y_2, \ldots, y_n\ )$ and thus $\sum_{i=1}^n Y_i$ is sufficient for π.

C. $\sum_{i=1}^n Y_i$ is sufficient for λ of the Poisson model, because for the Poisson model define $T = \sum_{i=1}^n Y_i$, $g(t,\lambda) = e^{-n\lambda}\lambda^t$, and $h(y_1, y_2, \ldots, y_n\ ) = 1/\prod_{i=1}^n y_i!$ so that the Poisson likelihood in (A3) equals $g(t,\lambda)h(y_1, y_2, \ldots, y_n\ )$. ▽

Now according to (A4), if $\boldsymbol{T}$ is sufficient for $\boldsymbol{Y}$, since the likelihood $L(\boldsymbol{\theta}|\boldsymbol{y}) = g(\boldsymbol{t},\boldsymbol{\theta})h(\boldsymbol{y})$ we can simplify (8) further by writing

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{t}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{A5}$$

because as $\boldsymbol{y}$ is given, we need not keep track of the terms involving $\boldsymbol{y}$, just as we had dropped the denominators of (3), (4), (5) and (6) in order to arrive at (8). Also note that in (A5) we have rewritten $g(\boldsymbol{t}, \boldsymbol{\theta})$ as $f(\boldsymbol{t}|\boldsymbol{\theta})$, where $f(\boldsymbol{t}|\boldsymbol{\theta})$ is the p.m.f./p.d.f. or sampling distribution of the sufficient statistics of $\boldsymbol{T}$, which actually follows from the proof of the factorization theorem.

Appendix B: Subjective Probability

Subjective probability of an event or a hypothesis A like "it will rain tomorrow" , "the value of Stock X will sky-rocket in the long run " , "the launch of this new product would be successful" , "I will get an S in Stats" , "we will get the movie ticket this afternoon" , "we will get a really good candidate to fill up the vacancy" , "the negotiation for the contract will go in favor of my organization", "managers with an engineering background tend to be more quality conscious" etc. is an individual's personal belief about the likelihood of the event happening. Nothing stops an individual from having a subjective judgment about events which are repeatable and thus too some extent verifiable by experimentation like "result of a toss of this coin would be Head" , "the amount of rice in 1 Kg. packets of brand X is less than a Kg." , "the light-bulbs manufactured by company Y last more than 1000 hours" , "annual profit of company Z lies somewhere between Rs.4 and 5 crores" etc. but the concept is more useful (and possibly the only tool available for analysis) in situations described in the first sentence, where there is no scope of repeatability of an experiment to verify one's subjective judgment directly, yet there is uncertainty in face of which decisions must be taken.

What we intend to do here is to provide a gentle non-rigorous i.e. non-axiomatic introduction to the concept of subjective probability and derive the probability laws which guide them from this basic understanding, without taking those probability laws to be granted as a part and parcel i.e. definitions and theorems of the mathematical treatment of the subject.

To begin with, and for the rest of these notes, it is easiest to understand subjective probability in terms of betting schemes. Suppose you are to assign a number between say 0 and 1, to your degree of personal belief about a possibly uncertain event A, with the understanding that 0 corresponds to impossibility of the event and 1 to complete certainty. How does one go about pinning down a number on the paper to express one's subjective belief about an uncertain event A? Consider a betting scheme in which you will gain Rs. 1 if the event A happens, and nothing (Rs. 0) if A does not happen. Since you are not allowed to gain anything for free (in case A happens that is) a certain amount say Rs.e would be charged as an entry-fee allowing you to enter the bet. That is the game is for you to buy a lottery ticket for Rs.e which would be worth Rs. 1 if A happens and becomes worthless if A does not happen. The question is, how much are you willing to pay for entering this bet or buying the lottery ticket? Depending upon *your* degree of (subjective) belief about the occurrence of the uncertain event A, there must be a certain threshold price p, below which you would reckon it to be profitable for you to enter the bet, and above which you are not prepared to pay. That is in your mind you try to evaluate the "fair" price of this bet (in which you stand to win Rs. 1 if A happens and 0 if it does not) in terms of its "expected" winning amount. This threshold "fair" price p of the lottery-ticket (or in other words the entry-fee of the bet)

is *your* subjective probability of occurrence of event A. That is p is your "expected" winning amount, such that you buy the lottery ticket (or enter the bet) if its price (entry-fee) e is less than (or even equal to, in which case you are really indifferent about buying this lottery-ticket) p, and do not enter the bet otherwise.

Though the above concept is fairly straight-forward, a few numerical example would help illustrating the concept. Obviously $e > 0$ and nobody would buy the lottery-ticket if it is priced above Rs. 1, the winning amount. So

$$0 \leq p \leq 1$$

(you will not enter the bet if you are absolutely certain that A would not happen no matter however small e is - this corresponds to $p = 0$; and on the other hand as long as $e < 1$ you would consider buying the lottery ticket to be in your advantage if you are absolutely certain of occurrence of A corresponding to $p = 1$). That is stronger your belief that "A would occur" more would you be willing1 to pay for the lottery-ticket. For example if you believe that there is an 80% chance of occurrence of A then your "expected" winning amount is Rs.O.80 ($1 \times 0.8 + 0\times$ whatever), and you would enter the bet as long as you are paying less than 80 paise as its entry-fee; if you believe that there is a 90% chance of occurrence of A you should be willing to pay up to 90 paise to enter the bet. Similarly if you believe that the chance of occurrence of A is only 0.25 you will not be prepared to pay more than 25 paise to enter the bet, or if you believe that the chance of occurrence of A is only 0.05 you will not enter the bet if the entry fee exceeds 5 paise.

So it can be summarized by saying that, subjective probability of an event A is what you consider to be a "fair" entry-fee for a bet, in which you stand to win Rs. 1 if A happens and nothing in case it does not. This fair price of the bet is mathematically same as how much *you* "expect" to win, but since this "expected" winning amount is harder to elucidate, for evaluating subjective probabilities always try to think of what *you* consider to be a "fair" price of the lottery ticket.

Now that we have a basic understanding of subjective probability in terms of the "fair" price of a bet, let's delve on this concept a little bit more for understanding the basic "Dutch Book Arguments[16]" used to prove the probability laws in later part of these notes. Since according to your judgment p is the "fair" price of the bet, now turning around the table, you should be equally willing to enter the same bet with someone else, who now pays you Rs. p as an entry-fee of a bet where you have to pay Rs. 1 if A happens and nothing if A does not happen. In fact in a situation where *your* subjective probability of A happening is p and where you have to pay Rs. 1 if A happens and nothing if A does not happen, it's just not "fairness" which dictates that you should accept Rs. p as the entry-fee, if you are asking for more than Rs. p as entry-fee, in your own ,assessment you would be loosing opportunity in making money. Let's see why. In previous paragraphs we have seen that if *your* subjective probability of A happening is p, then Rs. p is the maximum amount you were willing to pay to enter that bet. Now when someone else is entering the same bet with you for an entry-fee,

[16]In British racing jargon a *book* is the set of bets a book-maker has accepted, and a book *against* someone -a "Dutch book" -is one in which the book-maker stands to suffer a net loss no matter how the race turns out.

you would naturally not accept an entry-fee less than Rs. p, because you would "expect" to loose in that case. However now if you fix an entry-fee of $p' > p$, so that you would not entertain any bet which I pays you an entry-fee of less than Rs. p', then from someone who is offering an entry-fee of say $\frac{p+p'}{2}$ in your own assessment you would loose an opportunity of an "expected" amount of Rs.$\frac{p+p'}{2} - p > 0$. That is what we just proved is:

Basic Lemma: If my subjective probability of A happening is p, then a bet in which I win Rs.1 if A happens and nothing if it does not, for an entry fee of Rs.p (is "fair" by definition of subjective probability, and) is equivalent to me to the reverse bet in which I accept an entry fee of Rs.p and pay Rs.1 if A happens and nothing if it does not.

Now we are in a position to give the proofs or the basic probability laws. In the sequel we will use the notation $P(A)$ as the subjective probability of occurrence of event A.

Complementation Law: For any event A, $P(A^c) = 1 - P(A)$, where A^c is the non-occurrence of event A.

Proof: Let $P(A) = p$ and $P(A^c) = p'$ be *my* subjective probabilities of A and A^c respectively, but suppose $p \neq 1 - p'$.

Case1: $p < 1 - p'$
Consider you entering the reverse bets for both A and A^c with me for respective entry-fees of Rs. p and p'. That is you enter into two bets with me. In the first you pay Rs. p as an entry fee to me with the understanding that I will pay you Rs. 1 if A happens, and nothing if it noes not (A^c happens). In the second you pay Rs. p' as an entry fee to me with the understanding that I will pay you Rs. 1 if A^c happens, and nothing if it does not (A happens). Since *my* subjective probabilities for A and A^c respectively are p and p', by the basic lemma, I should be willing to enter into both of these reverse bets with you. Now no matter which one of the events happen, A or A^c, you are guaranteed a winning amount of exactly Rs.1 from me by paying me a total sum of $p + p'$, which is less than 1 by assumption and thus holding a "Dutch Book" against me! So if I am assigning a subjective probability of p to event A it would be irrational for me to assign a subjective probability of less than $1 - p$ to A^c, because then someone can hold a Dutch book against me compelling me to suffer a sure loss.

Case2: $p > 1 - p'$
Show that in this case also you can hold a Dutch book against me, and thus it would again be irrational for me to assign a subjective probability of more than $1 - p$ to A^c. $\bigtriangledown$

Thus what we have just shown is that if you assign a subjective probability of p to an event A, then you do not have any choice but to assign a subjective probability of $1 - p$ to A^c. For otherwise it would be irrational and someone else can hold a Dutch book against you, making sure that you loose money no matter what. Thus the first law of subjective probability is the complementary law.

Addition Law: If $A_1 and A_2$ are two mutually exclusive events, that is they cannot occur simultaneously or $A_1 \bigcap A_2 = \phi$, then $P(A_1 \bigcup A_2) = P(A_1) + P(A_2)$.

Proof: Let $A = A_l \bigcup A_2$ and $P(A_1) = p_1, P(A_2) = p_2$ and $P(A) = p$ be *my* subjective

probabilities of A_1, A_2 and A respectively, but suppose $p_1 + p_2 \neq p$.

Case 1: $p_1 + p_2 < p$
Since I believe that $P(A_1) = p_1$ and $P(A_2) = p_2$ I should be willing to enter the reverse bets on A_1 and A_2 with you. Or in other words, first you enter into two bets with me on A_1 and A_2, in which you pay me Rs.p_1 and p_2 respectively as entry-fees, with the understanding that for the first bet I would pay you Rs. 1 if A_1 happens (and nothing otherwise) and for the second bet I would pay you Rs. 1 if A_2 happens (and nothing otherwise). At the same time since I believe that $P(A) = p$, I should be willing to enter a forward bet with you on A in which I pay you Rs. p as an entry-fee and win Rs, 1 from you if A happens and nothing if it does not. Now since $A = A_1 \bigcup A_2$ and $A_1 \bigcap A_2 = \phi$, if A happens exactly one of A_1 and A_2 must happen and vice-versa, and if A does not happen none of A_1 or A_2 can happen either and vice-versa. So if A happens I pay you Rs, 1 for one and only one of the reverse bets on A_1 and A_2, while you also pay me Rs. 1 for the forward bet on A, resulting in you gaining a net amount of $p - p_1 - p_2 > 0$ in entry-fees. Similarly if A does not happen both of us loose all our bets (you two on A_1 and A_2 and me the one on A) but you still retain your gain of $p - p_1 - p_2 > O$ in entry-fees. Thus if I hold that $p_1 + p_2 < p$, then you can hold a Dutch book against me with the above betting scheme forcing me to incur a sure financial loss. So to be rational I must not hold $p_1 + p_2 < p$.

Case 2: $p_1 + p_2 > p$
Show that in this case also you can hold a Dutch book against me, and thus it would again be irrational for me to hold that $p_1 + p_2 > p$. $\bigtriangledown$

Thus in assigning subjective probabilities to events one must have regard for the above addition law along with the complementation law. To illustrate the point with an example, suppose that a contract, which would be awarded to only one party, is being bid by you and two other competitors A and B. If you assign (subjective) probabilities 0.4 and 0.3 respectively for A and B bagging the contract, then you must assign a probability of 0.7 to the event "your competitor bags the contract" as a consequence of the addition law and subsequently a probability of 0.3 to the event "you bag the contract" as a result of the complementation law. (You have to be a bit indifferent while thinking of the betting scheme for eliciting the probability of the event "you bag the contract" because, even if you honestly feel that you have 30% chance of bagging the contract you may not be willing to pay 30 paise as the entry fee to this bet which might compound your loss in case of loosing the bet by loosing the contract as well!).

Now we will discuss the last law of subjective probability, namely the product law. Here we are concerned with conditional probabilities where we express our uncertainty about an event depending on the state of our information. Actually all probabilities are conditional probabilities, because when we are appraising probabilities we are doing so at the *given* state of our knowledge. If the state of knowledge remains unchanged from event to event we usually suppress this explicit dependence for the sake of simplifying notation. But in general our state of knowledge changes as we proceed with a problem and calls for additional notations to distinguish between the two situations -one with and the other without the piece of information.

Let B denote an additional piece of information which is available to us for the problem of eliciting probability of the event of interest A. We denote this probability by $P(A|B)$ to distinguish it from $P(A)$ referring to the probability of event A when we do not have information B. To clarify the concept let A denote the event that, "Boss will like my idea" and B denote the event "Boss is in a terrible mood". On a fine morning after a night-out on your work-station you may have some appraisal for $P(A)$ but this would probably change into something different, call it $P(A|B)$, once you have had a chat with the boss's secretary in the coffee room.

The only difficulty in expressing $P(A|B)$ as a (subjectively) "fair" entry fee of a certain bet lies in the fact that it essentially has to be left as an undefined quantity in case of $P(B) = 0$. That it has to be left as an undefined quantity in this case is intuitively very clear though. For in this case, you are asked to give your belief about the occurrence of certain event A with the knowledge of all event B, which you consider to be impossible to happen!

To circumvent this difficulty in appraising $P(A|B)$, think of the conditional bet, where *you* consider Rs. $P(A|B)$ to be a "fair" entry fee for a bet where you get Rs. 1 if A happens and nothing if it does not, when you know that B has happened; and you do not entertain any bet (on A) at all if B does not happen. This is same as the following bet which is slightly more transparent. Consider a conditional bet in which you win Rs. 1 if both A and B happens, nothing if A does not happen but B happens, and the bet is called off in case B does not happen and you get your entry fee back. What would you consider to be a "fair" entry fee for this bet? Since it is essentially same as the aforementioned conditional bet, the "fair" entry-fee for this bet should also be Rs3. $P(A|B)$. Now we are in a position to state and prove the:

Multiplication Law: $P(A \bigcap B) = P(A|B)P(B)$

Proof: Consider two other bets apart from the ones introduced in the last paragraph. The first bet is for elucidating the subjective probability of $A \bigcap B$ in which you gain Rs. 1 if both A and B (or in other words $A \bigcap B$) happens and nothing otherwise for *your* "fair" entry-fee of $P(A \bigcap B)$. The second bet, though only involves the occurrence of event B, is a little peculiar in terms of its pay-off not being a whole Rs. 1. In this bet you gain Rs. $P(A|B)$ if B does not happen and nothing otherwise for an entry fee of Rs. $P(A|B)P(B^c)$ which is less than the winning amount (provided of course $P(B^c) < 1$ - otherwise in your assessment B is an impossible event with $P(B) = 0$ implying $P(B^c) = 1$ by the complementation law and your sure winning amount of $P(A|B)$ equals the entry fee which is only "fair"). Now it is all easy exercise to check that the total winning amounts of these two new bets together equals the winning amount of the conditional bet introduced in the last paragraph in all possible combinations of occurrence and non-occurrence of the events A and B. This is demonstrated in the following two tables:

Winning Amount of the Conditional Bet

$A \rightarrow$ $B \downarrow$	Happens	Does Not
Happens	1	0
Does Not	$P(A\|B)$	$P(A\|B)$

Sum of the Winning Amounts of the Two Bets

$A \rightarrow$ $B \downarrow$	Happens	Does Not
Happens	$1 + 0 = 0$	$0 + 0 = 0$
Does Not	$0 + P(A\|B)$ $= P(A\|B)$	$0 + P(A\|B)$ $= P(A\|B)$

Now since the winning amount of the conditional bet is exactly same as the sum of the winning amounts of the two bets, the entry fee of the conditional bet must equal the sum of the entry-fees of these two bets. For if it is less, then one can offer a reverse bet on the conditional bet with a smaller entry-fee and at the same time offer forward bets on the other two gaining more in entry fee, retaining this excess gain no matter what happens to A and B and thus holding a Dutch book against an individual who has the above as his/her subjective probabilities. Similarly if it is more, one can offer a forward bet on the conditional bet and reverse bets on the other two and thus again holding a Dutch book against such an irrational individual. So these two must coincide, or in other words,

$$P(A|B) = P(A \bigcap B) + P(A|B)P(B^c)$$

Now by the complementation law since $P(B^c) = 1 - P(B)$, the above equality yields the multiplication law.