A general Bayes theory of nested model comparisons

Thomas Jonathan Chadwick

A thesis submitted for the degree of Doctor of Philosophy

at the University of Newcastle upon Tyne

September 18, 2002

"Men are more apt to be mistaken in their generalisations than in their particular observations."

Niccolo Machiavelli (1469-1527).

Abstract

We propose a general Bayes analysis for nested model comparisons which does not suffer from Lindley's paradox. It does not use Bayes factors, but uses the posterior distribution of the likelihood ratio between the models evaluated at the true values of the nuisance parameters. This is obtained directly from the posterior distribution of the full model parameters. The analysis requires only conventional uninformative or flat priors, and prior odds on the models.

The conclusions from the posterior distribution of the likelihood ratio are in general in conflict with Bayes factor conclusions, but are in agreement with frequentist likelihood ratio test conclusions. Bayes factor conclusions and those from the BIC are, even in simple cases, in conflict with conclusions from HPD intervals for the same parameters, and appear untenable in general.

Examples of the new analysis are given, with comparisons to classical P-values and Bayes factors.

Acknowledgements

I would like to thank both Professor Murray Aitkin and Dr. Richard Boys for their support and encouragement during the preparation of this thesis.

Financial support during the first three years of study was provided by the Engineering and Physical Sciences Research Council.

The School of Mathematics and Statistics at the University of Newcastle upon Tyne provided invaluable support throughout. In particular, I would like to thank everyone there for making the completion of this thesis possible by arranging teaching duties once my initial funding had been exhausted.

Thanks should also be given to my friends in Newcastle, who have not only provided valuable assistance but have made my time here greatly enjoyable.

To my family and Alaa for their support - thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In this thesis we address the point null hypothesis testing problem from a Bayesian viewpoint. An important issue in the evaluation of Bayes and frequentist theories is the difference in the conclusions from Bayes factors and the likelihood ratio (LR) tests in large samples, due to the "Lindley paradox" (Lindley 1957, Bartlett 1957). One important aspect of this difference is the ability of Bayes factors to strongly support a point null hypothesis, where a frequentist analysis can only "fail to reject" with a large P-value; it can never support it strongly.

We illustrate this feature with an example due to Stone (1997) from the discussion of Aitkin (1997). A physicist running a particle-counting experiment wishes to identify the proportion θ of a certain type of particle. He has a well-defined scientific hypothesis that $\theta = 0.2$ precisely. There is no specific alternative. He counts $n = 527135$ particles and finds $r = 106298$ of the specified type.

The problem is to make inferences concerning the probability of success, θ, in a series of n trials with r successes. The likelihood has the following form from the binomial distribution

$$L(\theta) = \binom{n}{r}(1-\theta)^{n-r}\theta^r, \quad 0 < \theta < 1.$$

The frequentist test of $\theta = \theta_0 = 0.2$ uses the maximum likelihood estimator of θ, $\hat{\theta} = 0.201652$, together with its standard error $SE(\hat{\theta}) = 0.0005526$, giving $Z_0 = \left(\hat{\theta} - \theta_0\right)/SE(\hat{\theta}) = 2.990$ and so the null hypothesis is firmly rejected with two-sided $P = 0.0028$. At this stage we can also calculate the

maximised likelihood ratio to be

$$\frac{L(\theta_0)}{L(\hat{\theta})} = 0.01124,$$

and note that the likelihood at $\theta = 0.2$ is very small relative to the likelihood at $\hat{\theta}$.

The physicist now takes the proper uniform prior $\pi(\theta) = 1,\ 0 < \theta < 1$ under the alternative hypothesis, and computes the Bayes factor

$$B = \frac{L(\theta_0)}{\int L(\theta)\pi(\theta)\mathrm{d}\theta},$$

in the following manner. Firstly recall that the beta distribution, $\beta(a, b)$, has the form

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1.$$

In this case, by comparison of $f(x)$ above to our form for $L(\theta)$ we can see that the term in the denominator of the Bayes factor reduces to the integral of a beta likelihood between 0 and 1. We see that

$$\int_0^1 L(\theta)\pi(\theta)\mathrm{d}\theta = \int_0^1 L(\theta)\mathrm{d}\theta = \binom{527135}{106298} \frac{\Gamma(106299)\Gamma(420838)}{\Gamma(527137)}.$$

The posterior distribution of θ, using this proper uniform prior is

$$\theta|r \sim \beta(106299, 420838).$$

As we are working with a large sample we can approximate this distribution with the normal distribution

$$\theta|r \dot{\sim} N(\hat{\theta}, SE(\hat{\theta})^2).$$

We calculate the Bayes factor

$$\begin{aligned} B &= \frac{L(\theta_0)}{\int L(\theta)\pi(\theta)\mathrm{d}\theta} \\ &= \frac{\binom{527135}{106298} 0.8^{420837} 0.2^{106298}}{\binom{527135}{106298}\Gamma(106299)\Gamma(420838)/\Gamma(527137)} \\ &= \frac{\Gamma(527137)}{\Gamma(106299)\Gamma(420838)} 0.8^{420837} 0.2^{106298}, \end{aligned}$$

which is the density of $\theta|r$ evaluated at $\theta = \theta_0 = 0.2$. Under the normal approximation this is

$$\begin{aligned} B &= \frac{1}{\sqrt{2\pi}SE(\hat{\theta})} \exp\left(-\frac{(\theta_0 - \hat{\theta})^2}{2SE(\hat{\theta})^2}\right) \\ &= \frac{\phi(Z_0)}{SE(\hat{\theta})}, \end{aligned}$$

where $\phi(\cdot)$ is the standard normal density function and Z_0 is as defined earlier.Therefore we can see that

$$B = \frac{1}{SE}\phi(Z_0) = 8.26.$$

Assuming equal prior weight on the null and (general) alternative hypotheses, the Bayes factor equals the posterior odds on the null hypothesis: we seem to have quite strong posterior evidence in favour of H_0, despite the apparently strong frequentist evidence against H_0.

A similar conclusion is reached from the closely related Bayesian Information Criterion, Schwartz (1978). This is interpreted in the same way as -2 times the logarithm of a standard Bayes factor and is defined here as

$$BIC = -2\log\left\{\frac{L(\theta_0)}{L(\hat{\theta})}\right\} - \nu \log n,$$

where ν is the number of unknown parameters under the null hypothesis subtracted from the number under the alternative hypothesis. Here $\nu = 1$ because θ is the only unknown parameter under the alternative and it is known under the null. Therefore

$$\begin{aligned} BIC &= -2\log\left\{\left(\frac{\theta_0}{\hat{\theta}}\right)^{106298}\left(\frac{1-\theta_0}{1-\hat{\theta}}\right)^{420837}\right\} - 13.17521 \\ &= -2(106298[\log\theta_0 - \log\hat{\theta}] + 420837[\log(1-\theta_0) - \log(1-\hat{\theta})]) \\ &\quad - 13.17521 \\ &= 8.976366 - 13.17521 \\ &= -4.198844. \end{aligned}$$

The likelihood ratio test statistic of 8.98 for the null hypothesis is outweighed by the penalty function $\log n = 13.18$ on the alternative model, giving a BIC of 4.20 in favour of the null model. The BIC is a special case of the Bayes

factor when the information in a proper normal prior is proportional (in sample size) to that of the sample (Smith and Spiegelhalter 1980).

This inconsistency between frequentist and Bayes conclusions is not, however, a sign of conflict between the theories. It is, instead, a conflict within Bayes theory. Here the posterior distribution of θ is the Beta distribution with parameters 106299 and 420838, which is essentially normal with mean $\hat{\theta}$ and standard deviation $SE(\theta)$. The posterior probability that $\theta > 0.2$ is $\Phi(2.990) = 0.9986 = 1 - P/2$. Any Bayesian using this prior will have very strong posterior belief that θ does not have the value specified under the null hypothesis, but a larger value. Equivalently, the 99% highest posterior density interval for θ is

$$\hat{\theta} \pm 2.576 SE(\hat{\theta}) = (0.20023, 0.20308),$$

which excludes θ_0.

This inconsistency between Bayes factor and posterior density conclusions results from the integration of the likelihood over the prior. It has been frequently pointed out (e.g. Aitkin 1991, 1997), that for a fixed prior the increasing concentration of the likelihood as $n \to \infty$ results in a decreasing integrated likelihood. This has the consequence that the Bayes factor can become arbitrarily large for any specified value θ_0, however small its likelihood relative to that at $\hat{\theta}$.

Dempster (1974, 1997) and Aitkin (1997) addressed this difficulty by considering the posterior distribution of the likelihood ratio itself. Aitkin (1991) had also considered the ratio of the posterior mean of the likelihood under each model. Their aim was to make an inferential statement directly about the ratio $L(\theta_0)/L(\theta)$, where θ is unknown under the alternative, but has a known posterior distribution

$$\pi(\theta|y) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)\mathrm{d}\theta}.$$

Since the likelihood ratio

$$LR(\theta) = L(\theta_0)/L(\theta)$$

is a parametric function of θ, it has a posterior distribution $\pi(LR|y)$ which can be obtained from that of θ.

Given a specified value of θ under the alternative hypothesis, a likelihood ratio $LR < k$ would constitute strong sample evidence against H_0 for

sufficiently small k. Motivated by this observation, the posterior probability

$$\pi_k = Pr(LR(\theta) < k \mid y)$$

for a given value of k can then be computed, and if this probability is sufficiently large, the sample evidence against H_0 would be persuasive.

We now return to the example considered earlier. Without loss of generality we can discard the constant of proportionality as we are interested in a ratio of two such likelihoods. The likelihood, under a normal approximation as seen earlier, is then

$$L(\theta) = \exp\left\{\frac{(\theta - \hat{\theta})^2}{2SE^2}\right\}$$

and therefore

$$-2\log L(\theta) = \frac{(\theta - \hat{\theta})^2}{SE^2} = Z^2.$$

Then the likelihood ratio LR for testing θ_0 against θ satisfies

$$-2\log LR = Z_0^2 - Z^2$$

where Z has a posterior $N(0,1)$ distribution. It follows that

$$\begin{aligned} \pi_k &= Pr(LR < k), \\ &= Pr(-2\log LR > -2\log k), \\ &= Pr(Z^2 < Z_0^2 + 2\log k), \end{aligned}$$

where Z^2 has a posterior χ_1^2 distribution under H_0. Taking $k = 0.1$ with $Z_0 = 2.990$ as before, gives

$$\pi_{0.1} = Pr(Z^2 < 2.990^2 - 4.605) = 0.963.$$

We note that taking $k = 1$ instead of 0.1 gives $\pi_1 = 0.9972 = 1 - P$.

Thus this form of Bayes analysis leads to the same conclusion as the HPD interval and the frequentist analysis, and contradicts the Bayes factor and BIC conclusions.

Aitkin (1997) extended Dempster's approach to the general nuisance parameter model: in assessing a null hypothesis $H_0 : \theta = \theta_0$ in a model with

nuisance parameter ϕ, he considered the "true likelihood ratio"

$$LR = L(\theta_0, \phi)/L(\theta, \phi)$$

evaluated at the true values of θ and ϕ. These are unknown, but the joint posterior distribution $\pi(\theta, \phi|y)$ from the "full model" provides, as for the sample model above, the posterior distribution of LR, and hence the posterior probability that $LR < k$. Aitkin also generalised Dempster's striking result for $k = 1$, given above, to the nuisance parameter case. This means that for a normal likelihood $L(\theta, \phi)$ and flat prior distributions for the unknown parameters, the P-value under the null hypothesis equals the posterior probability that the true likelihood ratio is greater than 1. This can be read as meaning that the null hypothesis is better supported than the alternative. This result provides a unification of Bayes, likelihood and frequentist conclusions in the point null hypothesis testing problem. Note also that here we assume that the nuisance parameter takes the same value under both the null and the alternative hypotheses. This differs from the maximised likelihood ratio approach in which we estimate it separately under the two models.

We now illustrate this approach further with a particularly simple example.

1.2 A simple example

We observe a single realisation (x) of a normal random variable with known variance, σ^2. Without loss of generality we may set $\sigma = 1$ so that our model is

$$X \sim N(\mu, 1)$$

with μ unknown. We wish to test the point null hypothesis $H_0 : \mu = \mu_0$ against the general alternative $H_1 : \mu \neq \mu_0$ with $\mu, \mu_0 \in \mathbb{R}$. We set $\mu_0 = 0$ without any loss of generality. We firstly consider the likelihood ratio, LR, which we define as

$$LR = \frac{L(0)}{L(\mu)}.$$

For this model the likelihood function itself is given by

$$L(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x - \mu)^2\right\}$$

so that

$$LR = \exp\left\{\frac{1}{2}\left[(x-\mu)^2 - x^2\right]\right\}.$$

Therefore

$$2\log LR = Y - x^2$$

where

$$Y = (x-\mu)^2 = (\mu - x)^2.$$

The prior distribution for μ is taken to be diffuse so that we may express the posterior as

$$\mu \sim N(x, 1).$$

We can now see that the posterior distribution of Y is χ^2_1 and if we consider that $LR < k$ is equivalent to $2\log LR < 2\log k$ for real, positive k then we find

$$\begin{aligned} \pi_k &= Pr(LR < k|x) \\ &= Pr(Y - x^2 < 2\log k) \\ &= Pr(Y < 2\log k + x^2). \end{aligned}$$

Therefore the posterior distribution of the likelihood ratio is simply a shifted χ^2_1 distribution. When $k = 1$ we obtain the following:

$$\begin{aligned} \pi_1 &= Pr(Y < x^2) \\ &= Pr(-x < Z < x) \\ &= 1 - P, \end{aligned}$$

where $Z \sim N(0,1)$ and P is the two-sided P-value of the observation x. Hence, once more $\pi_1 = 1 - P$.

In addition to the form given above for π_k we can also obtain the maximised likelihood ratio (MLR) by calculating the ratio of the likelihoods maximised under the two hypotheses. Under the null hypothesis this is simply

$$L(0) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\},$$

while under the alternative we make use of $\hat{\mu}$, the maximum likelihood estimator of μ. Here this is $\hat{\mu} = x$ and we obtain the maximised likelihood

$$\begin{aligned} L(\hat{\mu}) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\hat{\mu})^2\right\} \\ &= \frac{1}{\sqrt{2\pi}}. \end{aligned}$$

Therefore

$$\begin{aligned} MLR &= \frac{L(0)}{L(\hat{\mu})} \\ &= \exp\left\{-\frac{x^2}{2}\right\}. \end{aligned}$$

We now illustrate the difficulties involved in calculating a Bayes factor using diffuse priors over an infinite range. Under the alternative hypothesis we take the proper flat prior for μ as

$$\pi(\mu) = \frac{1}{2C} \text{ on } -C < \mu < +C.$$

The integrated likelihood under this alternative is

$$\begin{aligned} L^B &= \int_{-\infty}^{\infty} L(\mu)\pi(\mu)\,\mathrm{d}\mu \\ &= \frac{1}{2C}\int_{-C}^{C} \phi(x-\mu)\,\mathrm{d}\mu \\ &= \frac{1}{2C}\left[\Phi(x+C) - \Phi(x-C)\right]. \end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions respectively. As $C \to \infty$ with increasing diffuseness $L^B \to 0$, and so the Bayes factor

$$B = \frac{L(0)}{L^B} \to \infty.$$

This, once more, is the Lindley (or Barlett) paradox. Whatever the value of x, if the prior is sufficiently diffuse the Bayes factor will appear to strongly support the null hypothesis. Here and for similar examples later in the thesis we are unable to calculate a Bayes factor for an improper flat prior over an infinite range.

1.3 Structure of the thesis

Aitkin (1997) considered the large-sample properties of the π_k approach and the re-calibration needed for standard frequentist and likelihood methods. He considered only one small-sample example. In the rest of this thesis we apply this approach to several difficult model comparison problems. Some are nested model comparisons while others are not, but all can be treated by the same general method, since we require only that the models being compared are both nested in a larger family. This approach applies directly to the comparison of any two models which are themselves nested in a higher family, irrespective of whether one is nested in the other.

We firstly consider, in Chapter 2, the familar problem of testing a normal mean in the situation where the variance is unknown. This is the problem whose standard analysis is performed by the t-test. In Chapter 3 we examine a problem that was first presented by Pitman (1937). We are given a two-variable normal regression and wish to select which of the two possible single-variable regression models is best supported by the data. Chapter 4 concerns the choice between a general AR(1) model for time series data and the special case of a (non-stationary) random walk. This is adapted from a problem studied by Marriott and Newbold (1998). Finally, in Chapter 5, we consider the well-known example due to Cox (1962). In this case having observed a sample from a discrete distribution we wish to determine whether a Poisson or a geometric distribution is better supported by the data.

In Chapters 3 and 5 we are able to use (improper) prior distributions in order to calculate Bayes factors for comparison purposes. We are able to do this in a straightforward manner as these examples are non-nested. We also obtain the maximised likelihood ratio in these cases.

We are unfortunately unable to calculate the actual Bayes factor for the problems that we consider in Chapters 2 and 4. The Bayes factor involves a ratio of integrated likelihoods under the two hypotheses being considered and for the cases given it is not possible to evaluate the required integrals over the infinite parameter spaces. We discuss this issue further in Chapter 6, where we suggest other possible comparisons as further work. This actually indicates a further strength of our approach as we are able to consider any choice of prior distribution whereas it is impossible to use the flat priors on infinite parameter spaces in order to calculate a Bayes factor.

We now note that it is possible to express all four of these applications in the same general form. For the calculation of π_k, we require the following elements, given a model parameter δ and data Y_i:

(a) Hypotheses H_0, H_1 (or H_1, H_2) and (possibly) a third encompassing

hypothesis H_e. This refers to a parent hypothesis that both H_0 and H_1 are nested within and in certain cases is simply H_1.

(b) A prior distribution over $\delta \in H_e$.

(c) A parameterisation of δ such that

$$\delta = (\theta, \gamma),$$

where θ is the parameter of interest and is specified under the null hypothesis and specified *or* a nuisance parameter under the alternative. γ is a nuisance parameter under both models.

For each example we then calculate (if possible in the case of the Bayes factor) the following quantities, denoting the likelihood function by $L(\cdot)$ and the (improper and uninformative) prior distribution used in the calculation of the Bayes factor by $\pi(\cdot)$. We here assume that we are working with H_0 and H_1.

(i) The maximised likelihood ratio which depends only on the data and is defined as

$$MLR = \frac{\max_\delta L_0(\delta)}{\max_\delta L_1(\delta)},$$

where $L_0(\cdot)$ and $L_1(\cdot)$ are the likelihood functions under the two models.

(ii) The posterior distribution of the "true" likelihood ratio. This is defined as $\pi_k = Pr(LR(\delta) < k|y)$ where

$$LR(\delta) = \frac{L_0(\delta)}{L_1(\delta)}.$$

Note that this depends on both the data and the parameter δ.

(iii) The Bayes factor, defined here as

$$B = \frac{\int L_0(\delta)\pi(\delta)\ \mathrm{d}\delta}{\int L_1(\delta))\pi(\delta)\ \mathrm{d}\delta}.$$

It is not always possible to calculate the Bayes factor. In fact, as discussed earlier, we are only able to obtain this for two of the four examples discussed.

We should note here that all but the first of these quantities depend on the choice of the nuisance parameter γ. The MLR however is invariant under reparameterisation of either the parameter of interest or the nuisance parameter. The dependence of the posterior distribution of LR, and indeed the Bayes factor, on the choice of parameterisation is to be expected if we are comparing different models. The dependence on the nuisance parameter γ may be reduced by the use of the orthogonal parameterisation for two parameters, should this exist. Alternatively the parameterisation which gives a diagonal expected information matrix could be chosen, if one exists.

We are now able to cast the examples discussed in the remainder of the thesis in the terms given above. For full details of the models the reader should refer to the relevant chapter.

1.3.1 Testing normal mean, σ unknown

In this example our parameter is $\delta = (\mu, \sigma)$, corresponding to a normal mean and variance. The required elements are as follows:

(a) Hypotheses:

$$H_0 : Y_i \sim N(0, \sigma^2),$$
$$H_1 \equiv H_e : Y_i \sim N(\mu, \sigma^2).$$

(b) We use a flat (diffuse) prior on $(\mu, \log\sigma)$ under H_e.

(c) Parameters:

$$\theta = \mu, \ \gamma = \sigma.$$

This is one of the examples where, due to the problems of integrating an improper prior over an infinite parameter space, we are unable to obtain the Bayes factor.

1.3.2 The choice between two single-variable regressions

Our parameter here is $\delta = (\beta_0, \beta_1, \beta_2, \sigma)$ and we are also given two covariate vectors x_1 and x_2 which we are interested in choosing between. We have the following required elements:

(a) Hypotheses:

$$H_1 : Y_i \sim N(\beta_0 + \beta_1 x_{1i}, \sigma^2),$$
$$H_2 : Y_i \sim N(\beta_0 + \beta_2 x_{2i}, \sigma^2),$$
$$H_e : Y_i \sim N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \sigma^2).$$

(b) We use a flat (diffuse) prior on $(\beta_0, \beta_1, \beta_2, \log \sigma)$ under H_e.

(c) Parameters:

$$\theta = (\beta_1, \beta_2),\ \gamma = (\beta_0, \sigma).$$

Here our two hypotheses are not nested within each other but within an encompassing hypothesis and so we are able to obtain a Bayes factor for model comparison using a diffuse prior on $(\beta_0, \beta_1, \beta_2, \sigma)$.

1.3.3 The choice between random walk and AR(1) time series

Our parameter here is $\delta = (\mu, \sigma, \phi)$ and we are interested in choosing between a random walk ($\phi = 1$) and a more general AR(1) model for our data. The elements for calculation of π_k are as follows:

(a) Hypotheses:

$$H_0 : Y_i | Y_{i-1} \sim N(Y_{i-1}, \sigma^2),$$
$$H_1 \equiv H_e : Y_i | Y_{i-1} \sim N(\phi Y_{i-1} + (1-\phi)\mu, \sigma^2).$$

(b) We use a flat (diffuse) prior on $((1-\phi)\mu, \log \sigma, \phi)$ under H_e.

(c) Parameters:

$$\theta = \phi,\ \gamma = (\mu, \sigma).$$

Once more, in this nested case, we are unable to obtain a Bayes factor when using improper priors.

1.3.4 The choice between Poisson and geometric distributions

We nest both the Poisson (H_1) and geometric (H_2) distributions within the encompassing negative binomial distribution and here use the parameter

$\delta = (\mu, r)$. The elements we require follow:

(a) Hypotheses:

$$H_1 : \ Y_i \sim P(\mu) \equiv NB(\mu, \infty),$$
$$H_2 : \ Y_i \sim G\left(\frac{\mu}{\mu+1}\right) \equiv NB(\mu, 1),$$
$$H_e : \ Y_i \sim NB(\mu, r).$$

(b) We use both a flat (diffuse) prior on $(\mu/(\mu+r), r)$ and a flat prior on $(\mu/(\mu+r), \log r)$ under H_e.

(c) Parameters:

$$\theta = r, \ \gamma = \mu.$$

Again, as we have nested the Poisson and geometric hypotheses in the negative binomial encompassing hypothesis, we are able to calculate a Bayes factor both for a diffuse prior on μ and a diffuse prior on $\log \mu$. In this case we do not need to specify a prior for r as it is given under both H_1 and H_2 so needs no consideration.

Chapter 2

Testing normal mean, σ unknown

2.1 The t-test problem

In Aitkin (1997) the problem of testing a normal mean with unknown variance was considered. The standard method of analysis for this problem is the t-test. We present the analysis of this problem from the Aitkin paper and extend it by completing the numerical integration mentioned in the paper and displaying the resultant values for π_k.

We consider the following model, where we have n observations from a general normal distribution:

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, 2, \ldots, n$$

with both μ and σ unknown. We wish to test the point null hypothesis $H_0 : \mu = \mu_0$ against the general alternative $H_1 : \mu \neq \mu_0$ with $\mu, \mu_0 \in \mathbb{R}$.

2.2 Likelihood ratio

We first consider the likelihood ratio, LR, which we define to be

$$LR = \frac{L(\mu_0, \sigma)}{L(\mu, \sigma)}$$

which uses a *section* through the likelihood at the true but unknown σ. Here

$$\begin{aligned} L(\mu,\sigma) &= \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i-\bar{x})^2+(\bar{x}-\mu)^2\right]\right\} \end{aligned}$$

so that

$$\begin{aligned} LR &= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i-\bar{x})^2+(\bar{x}-\mu_0)^2-(x_i-\bar{x})^2-(\bar{x}-\mu)^2\right]\right\} \\ &= \exp\left\{-\frac{n}{2\sigma^2}\left[(\bar{x}-\mu_0)^2-(\bar{x}-\mu)^2\right]\right\}. \end{aligned}$$

Therefore

$$\begin{aligned} -2\log LR &= \frac{n}{\sigma^2}\left[(\bar{x}-\mu_0)^2-(\bar{x}-\mu)^2\right] \\ &= \frac{t^2}{n-1}\times\frac{(n-1)s^2}{\sigma^2}-\frac{n(\bar{x}-\mu)^2}{\sigma^2}. \end{aligned}$$

where

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \text{ and } t=\frac{\sqrt{n}(\bar{x}-\mu_0)}{s}.$$

The joint prior distribution for $(\mu,\log\sigma)$ is taken to be diffuse so that we can express the joint posterior as

$$\mu|\sigma\sim N\left(\bar{x},\frac{\sigma^2}{n}\right),\quad \frac{(n-1)s^2}{\sigma^2}\sim\chi^2_{n-1}.$$

We are interested in expressions of the form $LR<k$, or equivalently $-2\log LR>-2\log k$, for some real k. If we define

$$Y_1=\frac{n(\bar{x}-\mu)^2}{\sigma^2} \text{ and } Y_2=\frac{(n-1)s^2}{\sigma^2}$$

then the posterior distributions of $Y_1|Y_2$ and Y_2 are independent conditional on σ, and are respectively χ^2_1 and χ^2_{n-1}. Since these distributions do not involve σ, Y_1 and Y_2 are unconditionally independent.

We now turn our attention to $\pi_k = Pr(LR < k|x)$. In this case we have

$$\pi_k = Pr\left(Y_1 < \frac{t^2 Y_2}{n-1} + 2\log k\right).$$

Note that when $k = 1$, this simplifies to

$$\begin{aligned}
\pi_1 &= Pr\left(Y_1 < \frac{t^2}{n-1} Y_2\right) \\
&= Pr\left(\frac{Y_1/1}{Y_2/(n-1)} < t^2\right) \\
&= Pr\left(F_{1,n-1} < t^2\right) \\
&= 1 - P
\end{aligned}$$

where P is the frequentist P-value of the hypothesis being tested. In general, however, we consider $k \neq 1$ and here numerical evaluation of the posterior probability is required.

We now define

$$\eta = -2\log k$$

then

$$\pi_k = Pr\left(\frac{t^2 Y_2}{(n-1)} - Y_1 > \eta\right).$$

Now we know that the χ^2_ν density is

$$\frac{\exp\left\{-\frac{z}{2}\right\} z^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}$$

and Y_1 and Y_2 are independent so their joint distribution is

$$f(y_1, y_2) = \frac{y_1^{-\frac{1}{2}} y_2^{\frac{n-3}{2}} \exp\left\{-\frac{y_1}{2}\right\} \exp\left\{-\frac{y_2}{2}\right\}}{2^{\frac{n}{2}}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}.$$

Now $\Gamma(1/2) = \sqrt{\pi}$ and

$$\Gamma\left(\frac{n-1}{2}\right) = \begin{cases} \left(\frac{n-3}{2}\right)! & \text{if } n > 2 \text{ is odd} \\ \sqrt{\pi} \prod_{i=1}^{n/2-1} \left(\frac{n-(2i+1)}{2}\right) & \text{if } n > 2 \text{ is even.} \end{cases}$$

Define

$$\kappa = \left[\sqrt{\pi 2^n}\Gamma\left(\frac{n-1}{2}\right)\right]^{-1},$$

so that

$$f(y_1, y_2) = \kappa y_1^{-\frac{1}{2}} y_2^{\frac{n-3}{2}} \exp\left\{-\frac{y_1}{2}\right\} \exp\left\{-\frac{y_2}{2}\right\}.$$

Looking at the region $t^2 Y_2/(n-1) - Y_1 > \eta$ where, in addition, both Y_1 and $Y_2 > 0$ we see that if we set

$$Y_1 = y_1,\ 0 < y_1 < \infty,$$

then we require that

$$\frac{t^2}{n-1} Y_2 - y_1 > \eta,$$

which is equivalent to

$$Y_2 > (\eta + y_1)\left(\frac{n-1}{t^2}\right).$$

We can now use this result to express π_k in terms of an integral over the joint distribution of $Y_1,\ Y_2$. So we have

$$\pi_k = \kappa \int_{y_1=0}^{\infty} y_1^{-1/2} \exp\left\{-\frac{y_1}{2}\right\} \left(\int_{y_2=\frac{(y_1+\eta)(n-1)}{t^2}}^{\infty} y_2^{\frac{n-3}{2}} \exp\left\{-\frac{y_2}{2}\right\}\, \mathrm{d}y_2 \right)\, \mathrm{d}y_1.$$

2.3 Results

Unfortunately this integral is not possible to evaluate analytically so the integral is evaluated by numerical integration. This was carried out using Maple for sample sizes n from 3 to 21 (so that the degrees of freedom on

Table 2.1: π_k & P for $n = 3$

	k						
t	.01	.05	.1	.15	.2	1	P-value
1.0	.00006	.00144	.00577	.01299	.02309	.57736	.42265
1.1	.00030	.00434	.01365	.02669	.04294	.61396	.38604
1.2	.00079	.01009	.02642	.04641	.06920	.64700	.35300
1.3	.00291	.01953	.04436	.07168	.10075	.67675	.32325
1.4	.00640	.03309	.06712	.10152	.13616	.70353	.29647
1.5	.01214	.05075	.09397	.13475	.17402	.72761	.27239
1.6	.02052	.07215	.12399	.17020	.21309	.74927	.25073
1.7	.03175	.09670	.15623	.20683	.25239	.76877	.23123
1.8	.04582	.12374	.18981	.24379	.29117	.78633	.21367
1.9	.06255	.15257	.22400	.28042	.32888	.80218	.19782
2.0	.08165	.18257	.25820	.31623	.36515	.81650	.18350
2.1	.10275	.21318	.29193	.35086	.39976	.82945	.17055
2.2	.12544	.24394	.32484	.38409	.43258	.84119	.15881
2.3	.14936	.27446	.35669	.41578	.46356	.85185	.14815
2.4	.17411	.30446	.38731	.44586	.49270	.86155	.13845
2.5	.19939	.33372	.41659	.47431	.52005	.87039	.12961
2.6	.22491	.36208	.44449	.50114	.54566	.87846	.12154
2.7	.25041	.38942	.47099	.52640	.56963	.88584	.11416
2.8	.27571	.41566	.49609	.55015	.59204	.89261	.10739
2.9	.30064	.44083	.51983	.57245	.61298	.89882	.10118
3.0	.32507	.46484	.54225	.59338	.63255	.90453	.09547

Y_2 ran from 2 to 20) and for given values of t, which is the only sample quantity which affects the integral for π_k. Thus it is possible to calculate π_k for this range of values of t, for various values of k. We should note that as a result of evaluating the integral numerically there is a known error of at most 5×10^{-10}. In order to save space but still display results across the full range of n considered, Tables 2.1 to 2.3 present only the cases $n = 3, 11, 21$. Full tables are available from the author.

Table 2.2: π_k & P for $n = 11$

	k						
t	.01	.05	.1	.15	.2	1	P-value
1.0	.00000	.00000	.00000	.00001	.00013	.65911	.34089
1.1	.00000	.00000	.00001	.00019	.00113	.70289	.29711
1.2	.00000	.00000	.00016	.00133	.00554	.74220	.25780
1.3	.00000	.00004	.00102	.00572	.01795	.77723	.22277
1.4	.00000	.00031	.00421	.01719	.04334	.80823	.19177
1.5	.00001	.00144	.01257	.03992	.08465	.83549	.16451
1.6	.00004	.00482	.02949	.07662	.14154	.85932	.14068
1.7	.00022	.01261	.05776	.12747	.21079	.88003	.11997
1.8	.00085	.02729	.09851	.19028	.28767	.89795	.10205
1.9	.00257	.05098	.15094	.26135	.36734	.91338	.08662
2.0	.00642	.08478	.21274	.33652	.44569	.92661	.07339
2.1	.01376	.12851	.28071	.41202	.51972	.93792	.06208
2.2	.02605	.18088	.35153	.48433	.58758	.94756	.05244
2.3	.04453	.23979	.42221	.55288	.64834	.95575	.04425
2.4	.06997	.30279	.49037	.61492	.70179	.96268	.03732
2.5	.10257	.36746	.55432	.67039	.74816	.96855	.03145
2.6	.14191	.43165	.61302	.71925	.78797	.97351	.02649
2.7	.18704	.49364	.66595	.76176	.82186	.97769	.02231
2.8	.23669	.55216	.71301	.79841	.85056	.98121	.01879
2.9	.28943	.46440	.75437	.82977	.87473	.98417	.01583
3.0	.34379	.65589	.79039	.85645	.89504	.98666	.01334

2.4 Discussion

We now examine Tables 2.1 to 2.3 and compare them to the standard t-test for this problem. The rejection criteria differ considerably in that, as we have already stated, we can take different values for both k and π_k when using our (k, π_k) test while the t-test rejects the null hypothesis when the P-value is less than a certain value, 0.05, say. This P-value can be read off from the tables in the last column for each value of t. The (k, π_k) formulation requires that we reject H_0 when π_k is considered to be large for sufficiently small k, eg if $\pi_{0.1} > 0.7$.

From the tables we see that the P-value decreases roughly exponentially as t increases for a given value of n and the same effect occurs for a given value of t as we increase n. Now since $\pi_1 = 1 - P$ we see an increase in π_1

Table 2.3: π_k & P for $n = 21$

	k						
t	.01	.05	.1	.15	.2	1	P-value
1.0	.00000	.00000	.00000	.00000	.00000	.67074	.32926
1.1	.00000	.00000	.00000	.00000	.00002	.71560	.28440
1.2	.00000	.00000	.00000	.00003	.00038	.75584	.24416
1.3	.00000	.00000	.00002	.00039	.00307	.79162	.20838
1.4	.00000	.00000	.00021	.00271	.01386	.82316	.17684
1.5	.00000	.00003	.00150	.01165	.04181	.85076	.14924
1.6	.00000	.00025	.00666	.03453	.09424	.87472	.12528
1.7	.00000	.00142	.02089	.07821	.17174	.89537	.10463
1.8	.00001	.00552	.05038	.14495	.26782	.91304	.08696
1.9	.00007	.01618	.09940	.23105	.37235	.92805	.07195
2.0	.00039	.03795	.16811	.32876	.47563	.94073	.05927
2.1	.00152	.07474	.25241	.42934	.57071	.95138	.04862
2.2	.00473	.12819	.34567	.52549	.65385	.96027	.03973
2.3	.01203	.19695	.44078	.61242	.72392	.96765	.03235
2.4	.02607	.27717	.53181	.68776	.78146	.97375	.02625
2.5	.04947	.36366	.61471	.75105	.82787	.97877	.02123
2.6	.08415	.45112	.68733	.80301	.86485	.98287	.01713
2.7	.13073	.53511	.74907	.84495	.89408	.98622	.01378
2.8	.18834	.61244	.80036	.87842	.91708	.98894	.01106
2.9	.25483	.68124	.84220	.90488	.93511	.99115	.00885
3.0	.32721	.74079	.87589	.92570	.94923	.99292	.00708

in both these cases. It is very interesting to note that as we increase n for values of k other than 1, we can observe different behaviour depending on the value of k that we are considering. For example, for the value $k = 0.2$, $\pi_{0.2}$ decreases with n for $t \leq 1.8$, but increases with n for $t \geq 1.9$. For $k = 0.05$, $\pi_{0.05}$ decreases with n for $t \leq 2.4$, but first increases and then decreases as n increases for $t = 2.5$, and increases with n for $t \geq 2.6$. The cut-off points for this decreasing behaviour for the different values of k are given in Table 2.4.

It is notable that for large P-values, the posterior probabilities change dramatically with n, while for small P-values they are relatively stable with n. Table 2.5 shows this for $\pi_{0.2}$.

Table 2.4: Approximate t cut-off for values of k

k	0.2	0.15	0.1	0.05	0.01
Cut-off	1.8	2.0	2.2	2.5	3.0

Table 2.5: P-value, t and $\pi_{0.2}$ for given n

n	P	t	$\pi_{0.2}$	P	t	$\pi_{0.2}$	P	t	$\pi_{0.2}$	P	t	$\pi_{0.2}$
3	.101	2.9	.613	-	-	-	-	-	-	-	-	-
11	.102	1.8	.288	.052	2.2	.588	.026	2.6	.788	.016	2.9	.875
21	.105	1.7	.172	.049	2.1	.571	.026	2.4	.781	.017	2.6	.865

2.5 Simulation

The preceding work, while providing us with the values of π_k, is computationally time-consuming with each of the preceeding tables taking upwards of a day to compile. We therefore consider finding a more efficient method to evaluate π_k.

Recall that

$$\pi_k = Pr\left(\frac{t^2}{(n-1)}Y_2 - Y_1 > -2\log k\right).$$

The simulation approach can now be expressed directly in terms of finding the distribution of the random variable $t^2Y_2/(n-1) - Y_1$ by simulating Y_1 and Y_2. The tail area probabilities follow directly.

For any given dataset both t and n are known and since η is a constant which we set at a specific numerical value, the only random elements in this expression are the (independent) random variables Y_1 and Y_2. It follows that we can obtain an approximation to π_k in the following manner:

First, given t and n, we simulate a large number (N) of pairs (y_1, y_2) and for each pair we then evaluate

$$h(y_1, y_2) = \frac{t^2}{(n-1)}y_2 - y_1.$$

We now ascertain the number of pairs for which $h(y_1, y_2) > \eta$ for our chosen η and divide this by N to give an approximate value for π_k. This method

Table 2.6: Estimated π_k & P for $n = 3$

	k						
t	.01	.05	.1	.15	.2	1	P-value
1.0	.000	.001	.005	.012	.022	.578	.422
1.1	.000	.004	.013	.025	.042	.612	.388
1.2	.001	.009	.025	.045	.067	.647	.353
1.3	.002	.018	.042	.069	.097	.677	.323
1.4	.006	.031	.065	.099	.130	.704	.296
1.5	.012	.048	.090	.129	.168	.729	.271
1.6	.019	.071	.118	.164	.208	.752	.248
1.7	.030	.092	.150	.202	.247	.769	.231
1.8	.044	.119	.184	.238	.285	.787	.213
1.9	.061	.146	.220	.274	.327	.804	.196
2.0	.078	.176	.253	.313	.366	.817	.183
2.1	.099	.208	.287	.351	.403	.830	.170
2.2	.121	.238	.323	.388	.435	.843	.157
2.3	.143	.269	.356	.419	.467	.853	.147
2.4	.167	.301	.389	.448	.493	.862	.138
2.5	.192	.332	.419	.475	.520	.872	.128
2.6	.220	.361	.447	.502	.544	.880	.120
2.7	.246	.391	.471	.525	.566	.888	.112
2.8	.267	.417	.497	.548	.587	.894	.106
2.9	.295	.442	.518	.570	.607	.900	.100
3.0	.322	.465	.541	.588	.625	.906	.094

speeds up the evaluation of π_k very considerably with each table now taking no longer than a few minutes to compile. As can be seen in the following tables, the values from the simulation compare well with the values calculated earlier using numerical integration. Note that here we are taking $N = 10000$ and that once more we display, in Tables 2.6 to 2.8, only the results relating to $n = 3, 11, 21$.

2.6 Further discussion

Examination of the tables and comparison with the results given earlier shows that the estimated π_k are reasonably close to the values obtained through numerical integration and that the same patterns are exhibited. Therefore, the

Table 2.7: Estimated π_k & P for $n = 11$

	k						
t	.01	.05	.1	.15	.2	1	P-value
1.0	.000	.000	.000	.000	.000	.669	.331
1.1	.000	.000	.000	.000	.001	.710	.290
1.2	.000	.000	.000	.001	.004	.742	.258
1.3	.000	.000	.001	.004	.016	.774	.226
1.4	.000	.000	.003	.016	.039	.807	.193
1.5	.000	.001	.012	.035	.083	.836	.164
1.6	.000	.003	.026	.075	.144	.858	.142
1.7	.000	.012	.052	.129	.213	.880	.120
1.8	.001	.024	.099	.195	.296	.896	.104
1.9	.002	.046	.155	.270	.375	.912	.088
2.0	.005	.082	.216	.344	.447	.925	.075
2.1	.012	.131	.289	.417	.520	.936	.064
2.2	.023	.186	.356	.484	.591	.946	.054
2.3	.040	.246	.425	.552	.649	.954	.046
2.4	.065	.309	.490	.616	.702	.961	.039
2.5	.103	.370	.554	.673	.749	.967	.033
2.6	.146	.435	.615	.719	.789	.971	.029
2.7	.192	.496	.667	.764	.826	.976	.024
2.8	.243	.551	.712	.800	.854	.980	.020
2.9	.295	.607	.754	.834	.875	.983	.017
3.0	.348	.655	.792	.859	.892	.987	.013

comments made in the earlier discussion (section 2.4) concerning the values of π_k remain relevant here with reference to the tables obtained through simulation. We therefore consider issues arising specifically from the simulation method.

As this method produces only an approximation to π_k, we consider the possible error built into the procedure. We are, in effect, simulating a value of the random variable $N\hat{\pi}_k$ which has a Binomial distribution:

$$N\hat{\pi}_k \sim Bin\left(N, \pi_k\right).$$

In order to obtain a feel for the potential error in our approximation to π_k

Table 2.8: Estimated π_k & P for $n = 21$

	k						
t	.01	.05	.1	.15	.2	1	P-value
1.0	.000	.000	.000	.000	.000	.667	.333
1.1	.000	.000	.000	.000	.000	.711	.289
1.2	.000	.000	.000	.000	.000	.753	.247
1.3	.000	.000	.000	.000	.002	.790	.210
1.4	.000	.000	.000	.002	.013	.820	.180
1.5	.000	.000	.001	.011	.042	.848	.152
1.6	.000	.000	.005	.034	.096	.873	.127
1.7	.000	.001	.020	.080	.173	.894	.106
1.8	.000	.004	.052	.144	.265	.913	.087
1.9	.000	.016	.101	.231	.366	.927	.073
2.0	.000	.038	.169	.325	.471	.940	.060
2.1	.001	.077	.250	.424	.566	.951	.049
2.2	.004	.130	.342	.522	.651	.961	.039
2.3	.012	.198	.436	.607	.722	.968	.032
2.4	.027	.274	.525	.686	.779	.974	.026
2.5	.051	.358	.609	.752	.827	.978	.022
2.6	.086	.445	.686	.800	.863	.982	.018
2.7	.131	.528	.750	.843	.894	.987	.013
2.8	.188	.608	.798	.878	.915	.989	.011
2.9	.253	.680	.840	.905	.933	.990	.010
3.0	.323	.739	.877	.923	.948	.992	.008

we require the variance of our estimator, $\hat{\pi}_k$. Now we know that

$$\begin{aligned} E\left(N\hat{\pi}_k\right) &= N\pi_k \Rightarrow E\left(\hat{\pi}_k\right) = \pi_k, \\ Var\left(N\hat{\pi}_k\right) &= N\pi_k\left(1-\pi_k\right) \Rightarrow Var\left(\hat{\pi}_k\right) = \frac{\pi_k\left(1-\pi_k\right)}{N}. \end{aligned}$$

As $Var(\hat{\pi}_k)$ depends on the (unknown) exact value of π_k we must instead use the estimated variance of the estimator

$$\widehat{Var}\left(\hat{\pi}_k\right) = \frac{\hat{\pi}_k\left(1-\hat{\pi}_k\right)}{N}.$$

This is a maximum, for fixed N, when $\hat{\pi}_k = 1/2$. Hence the maximum value that this estimated variance can attain is $1/(4N)$, which for our earlier choice

of $N = 10000$ is 0.000025; the corresponding sampling standard error is at most 0.005. Thus the true proportion lies with 95% confidence within 2 standard errors (at most 0.01) of the observed proportion reported in Tables 2.6 to 2.8. The mimimum value is zero and is attained at $\hat{\pi}_k = 0, 1$. A comparison of the tables obtained through numerical integration and simulation establishes that the maximum difference is approximately 0.01. We can therefore be confident that our approximate values obtained through simulation are accurate to within 0.01 of the true values.

It is straightforward to obtain the (approximate) cumulative distribution function of the likelihood ratio by simply plotting π_k against k. We undertake this procedure for our selected values of n and t and alongside these plots we also display the corresponding density estimates. The densities are estimated using the "density" function in the software package R which provides kernel density estimates. The software disperses the mass of the empirical distribution function over a grid and then convolves this approximation with a discretised version of a normal kernel before using linear approximation to evaluate the density. The distribution and density functions for $n = 3, 11, 21$ and $t = 1, 2, 3$ are shown in Figures 2.1 to 2.3. On the density plots we also display, as a vertical line, the maximised likelihood ratio (MLR). This is obtained using the maximum likelihood estimates $(\hat{\mu}, \hat{\sigma})$ of (μ, σ) under H_1 in addition to the maximum likelihood estimate, $\tilde{\sigma}$ of σ under H_0. We define

$$MLR = \frac{L(\mu_0, \tilde{\sigma})}{L(\hat{\mu}, \hat{\sigma})}.$$

Standard maximisation techniques give:

$$\begin{aligned} \hat{\mu} &= \bar{x}, \\ \hat{\sigma}^2 &= \frac{s^2(n-1)}{n}, \\ \tilde{\sigma}^2 &= \frac{s^2(n-1+t^2)}{n}. \end{aligned}$$

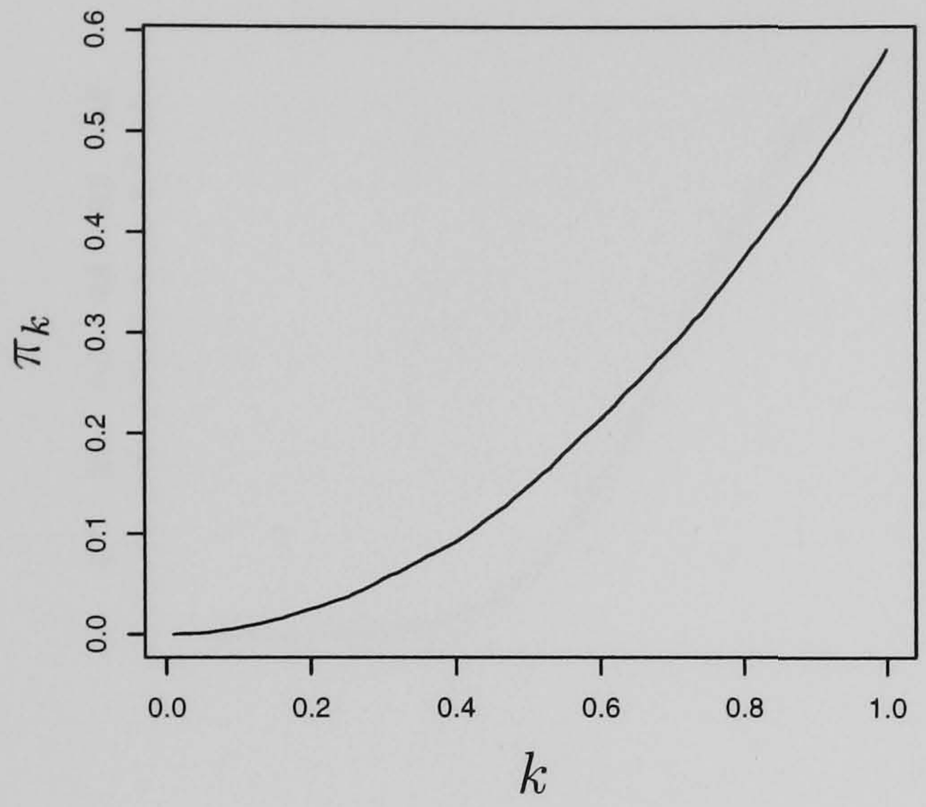
It is straightforward to show that the MLR (in favour of the null model) is given by

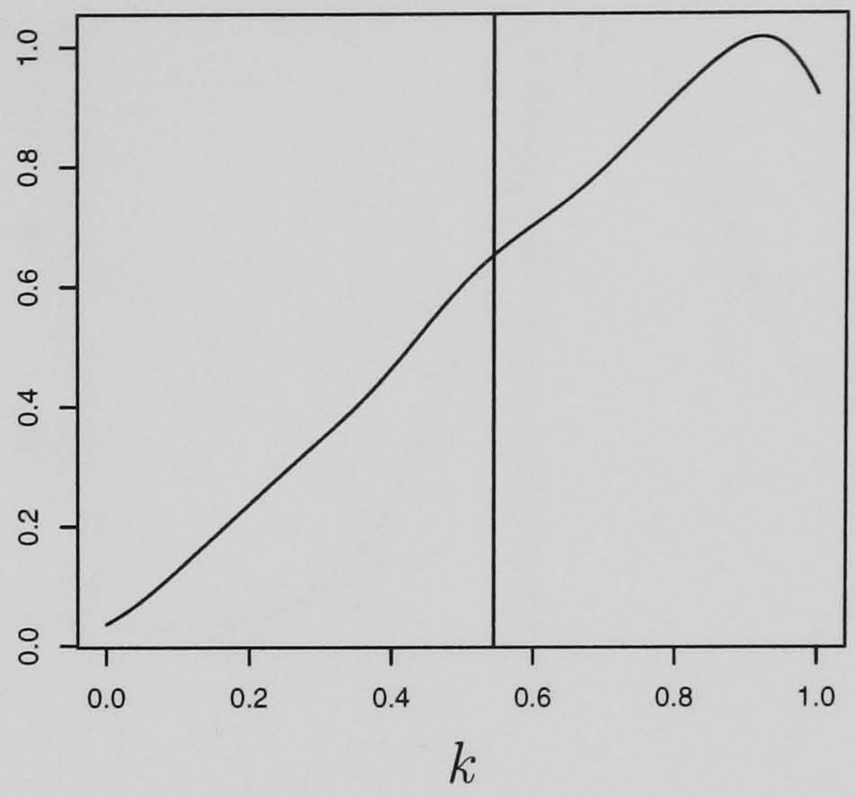
$$MLR = \left(\frac{n-1}{n-1+t^2}\right)^{n/2}.$$

It is immediately apparent from the density plots that, for $t = 2$ and 3, the MLR overstates the evidence in favour of the null hypothesis as its value lies to the right of the peak of our distribution. This overstatement

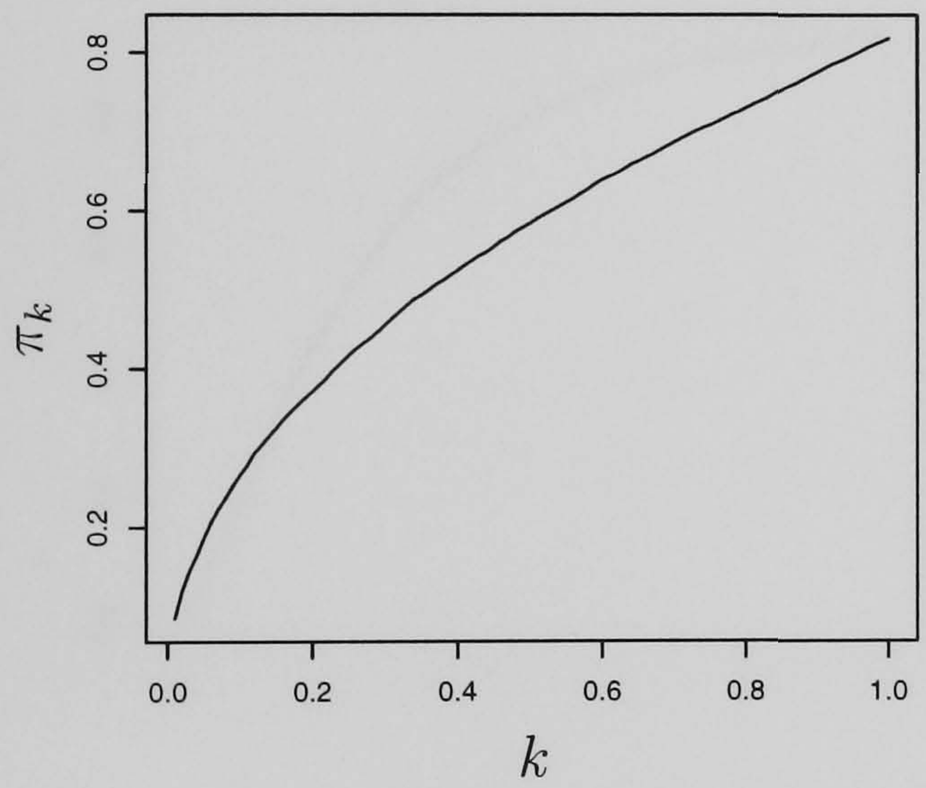
Figure 2.1: Distribution (left) & density (right) functions for $n = 3, t = 1, 2, 3$. Vertical line shows $k = MLR$.

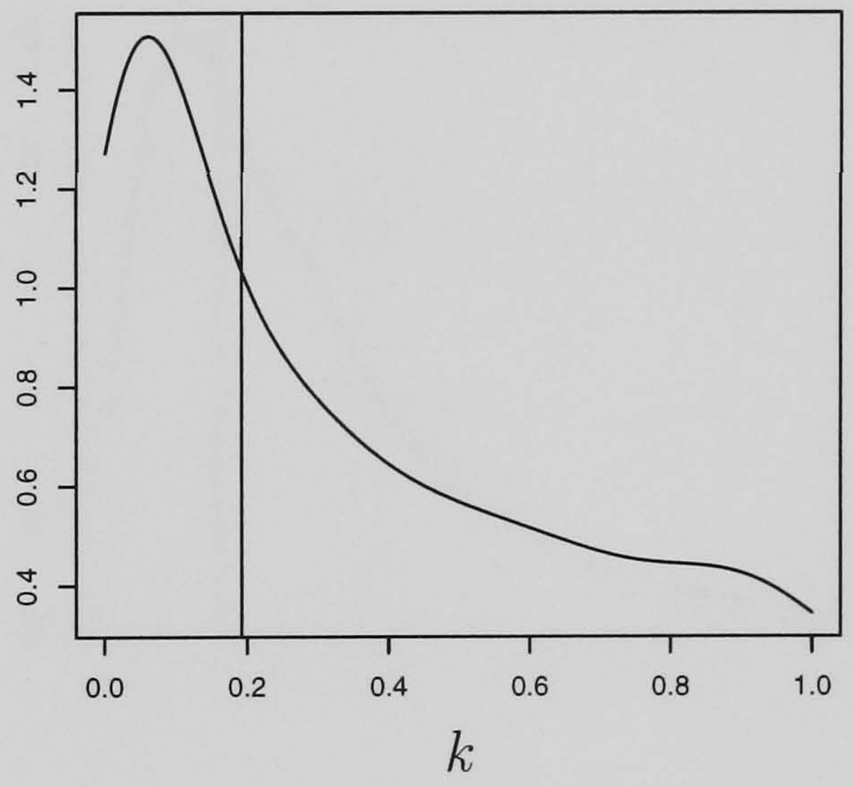
$MLR = 0.54433$



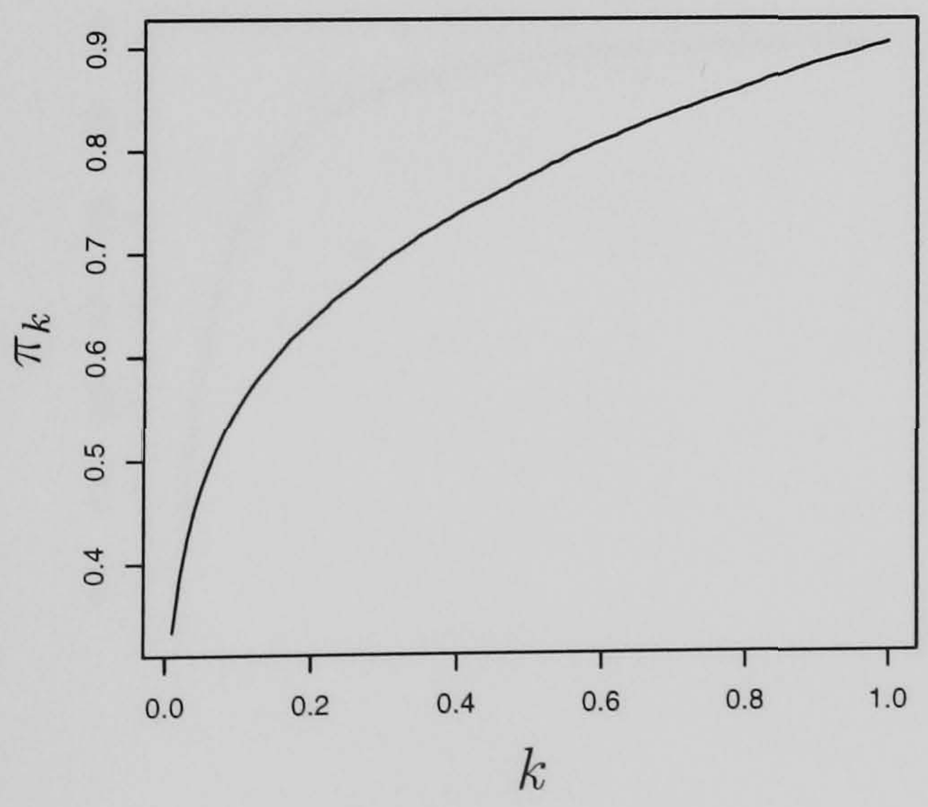


$MLR = 0.19245$





$MLR = 0.07753$



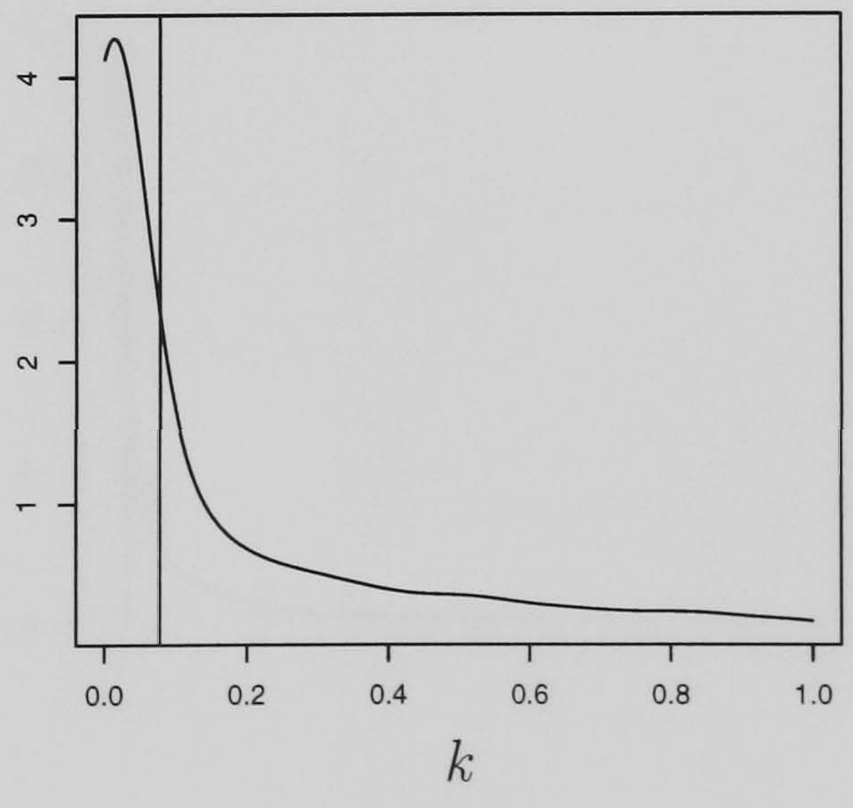
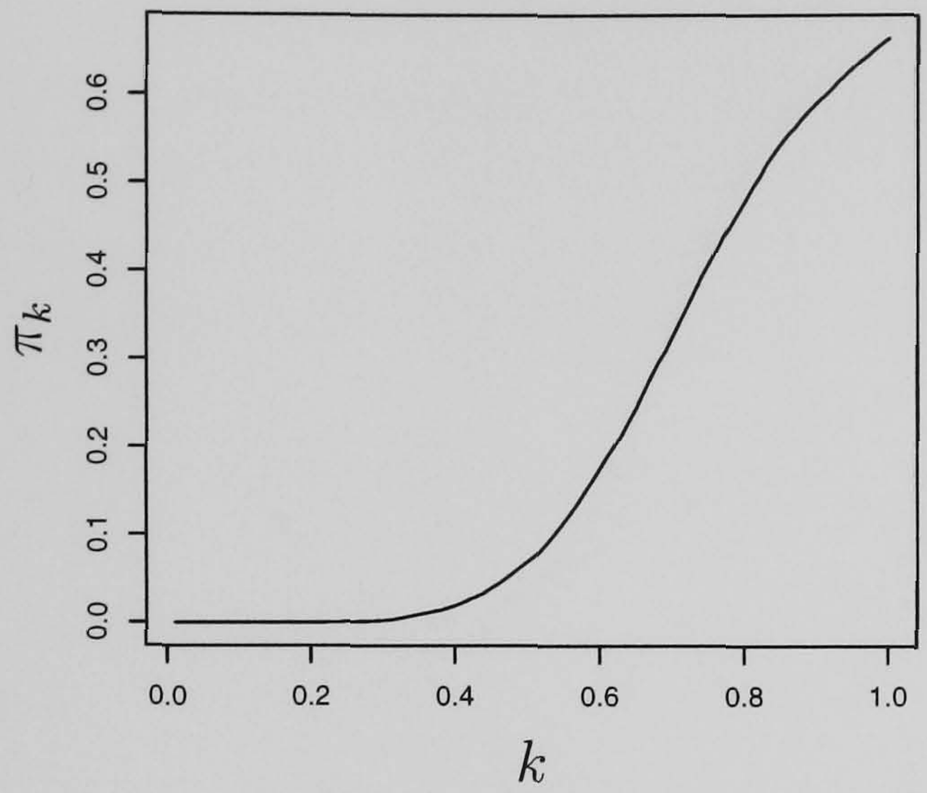
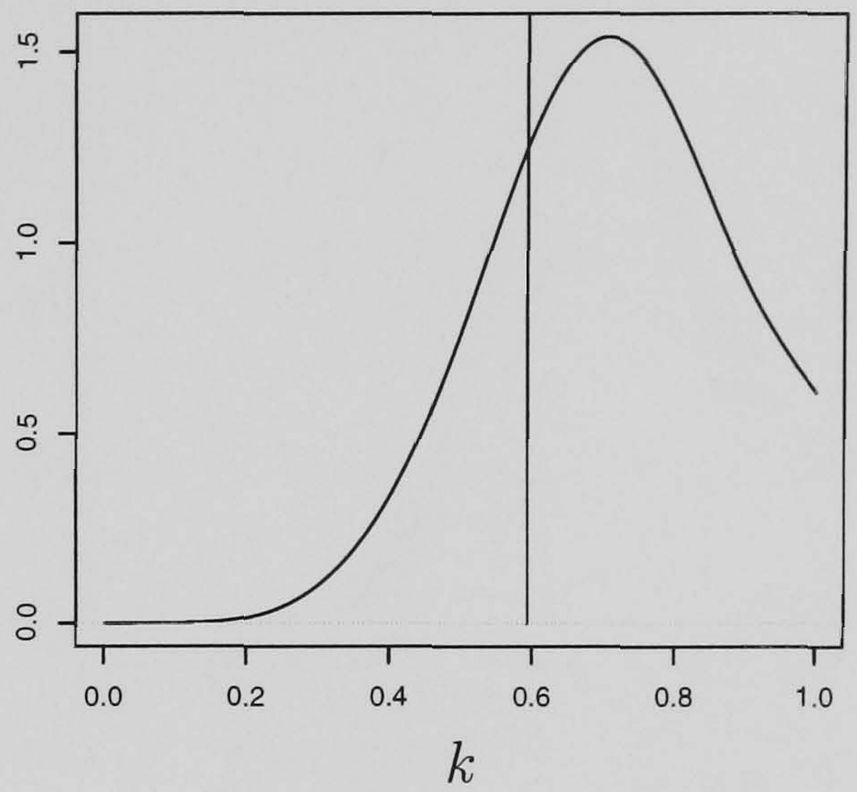


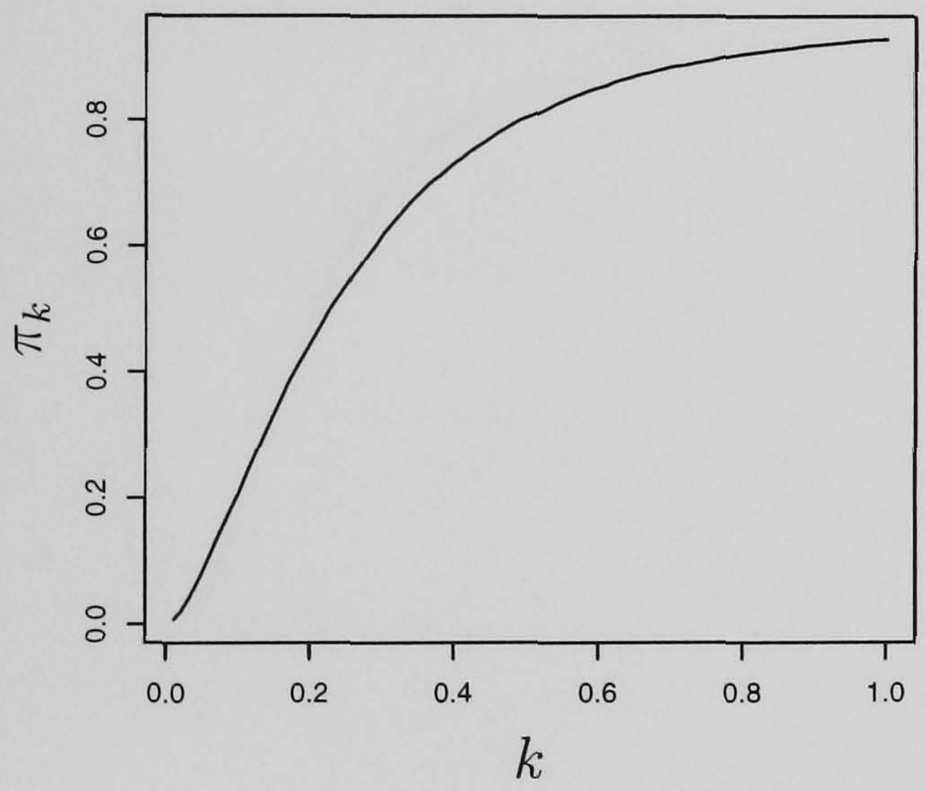
Figure 2.2: Distribution (left) & density (right) functions for $n = 11, t = 1, 2, 3$. Vertical line shows $k = MLR$.

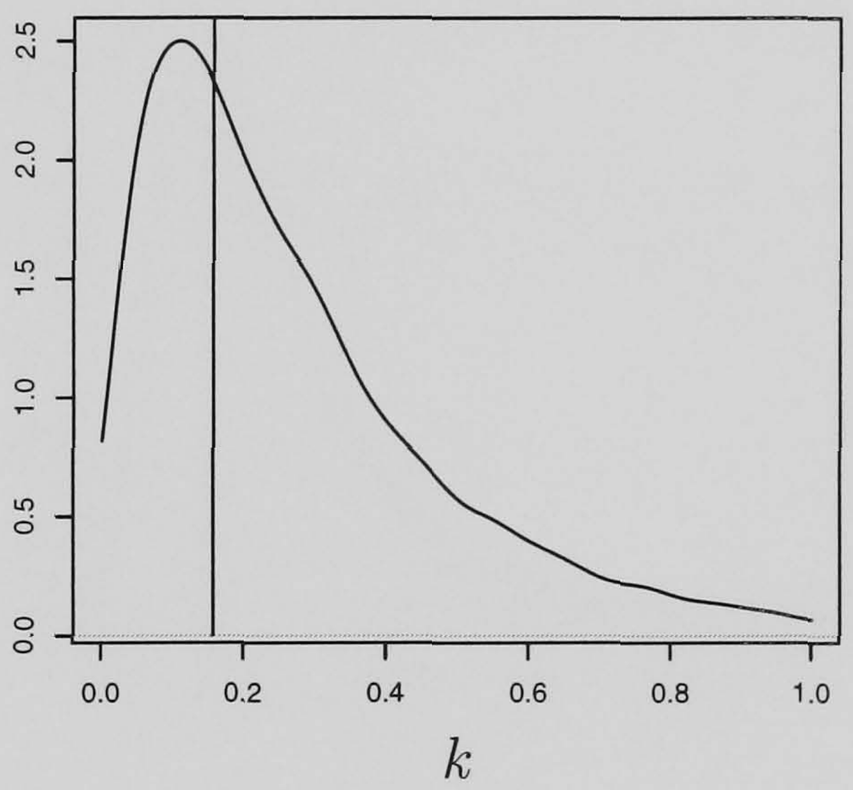
$MLR = 0.59203$



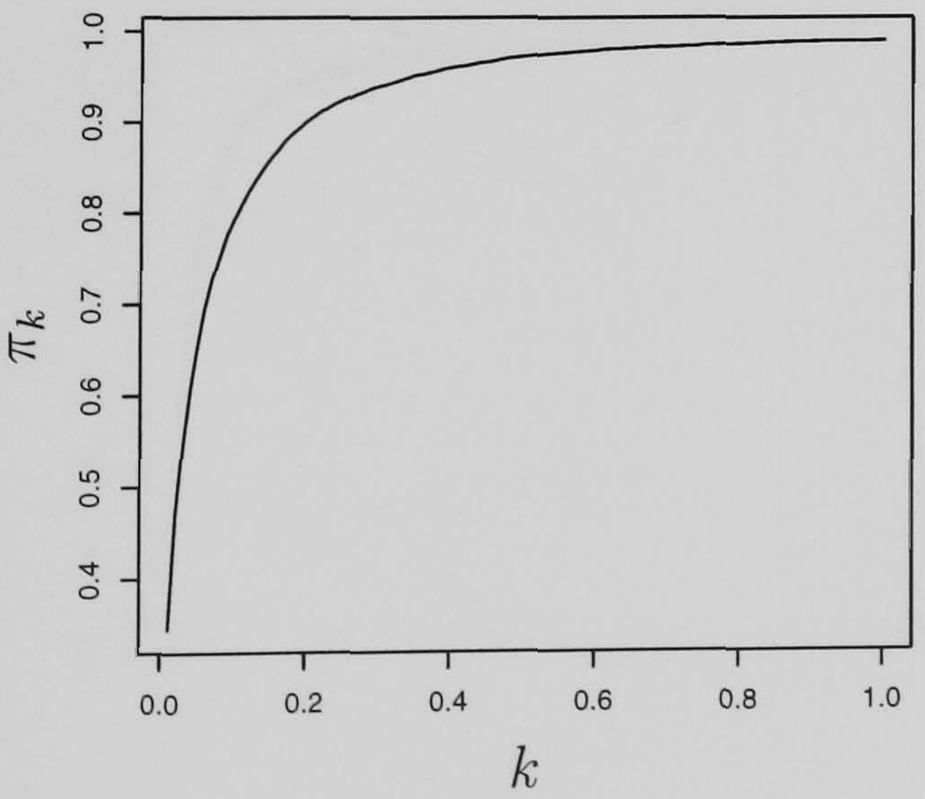


$MLR = 0.15714$





$MLR = 0.02930$



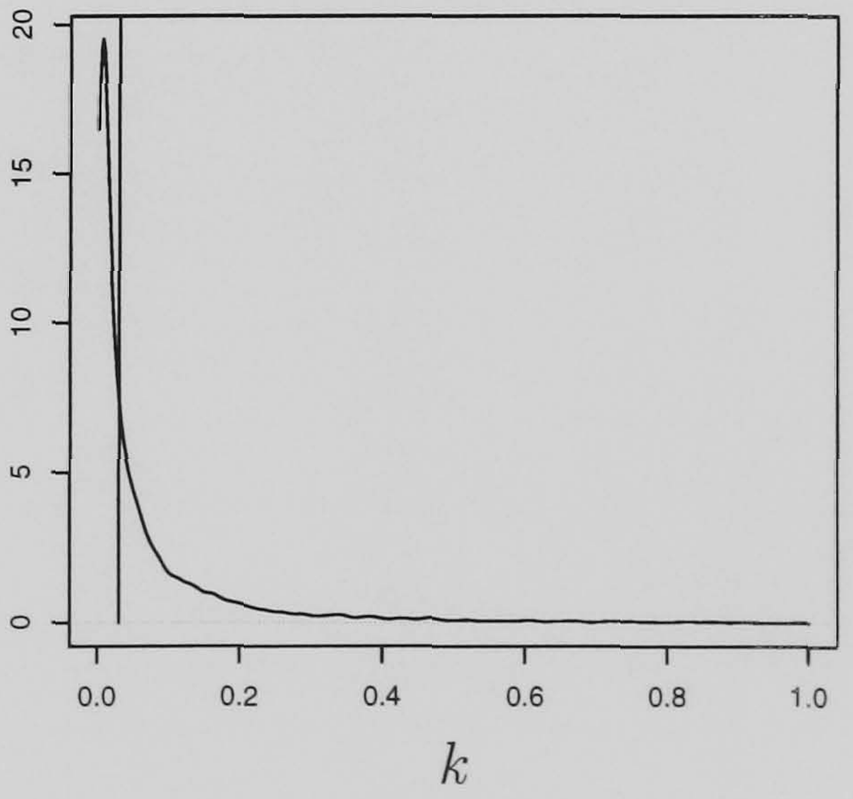
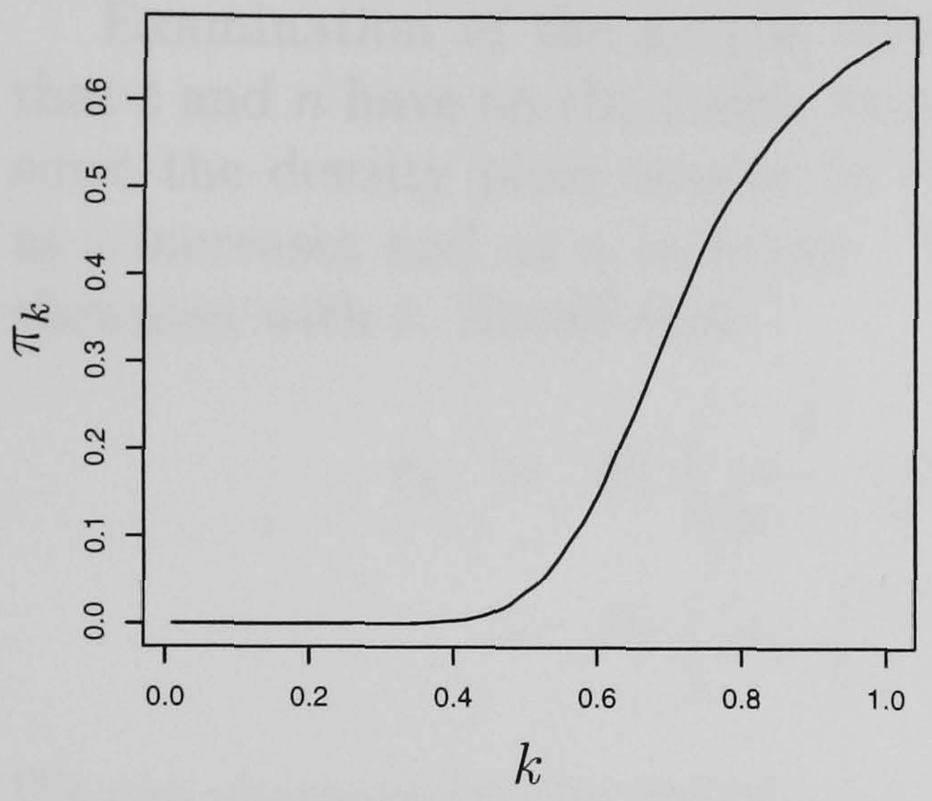
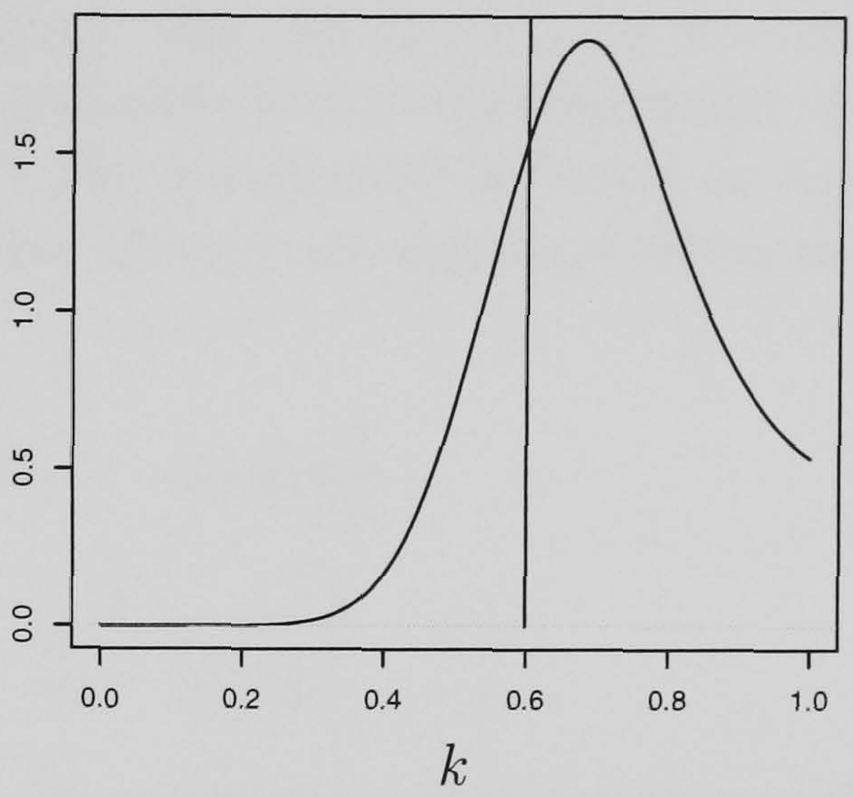


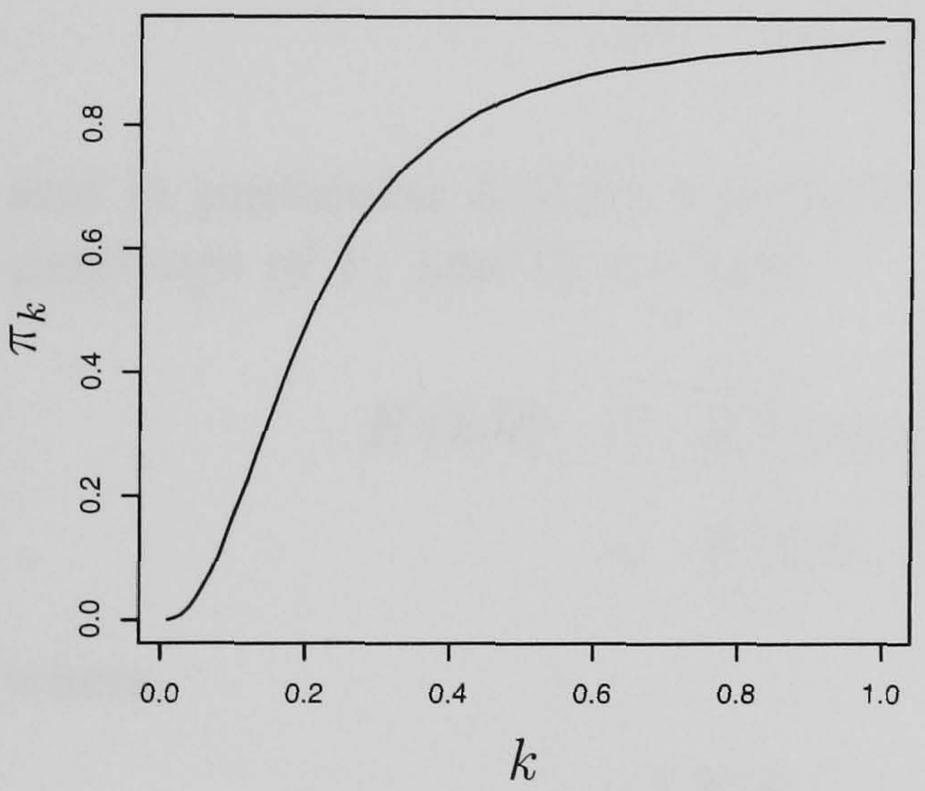
Figure 2.3: Distribution (left) & density (right) functions for $n = 21, t = 1, 2, 3$. Vertical line shows $k = MLR$.

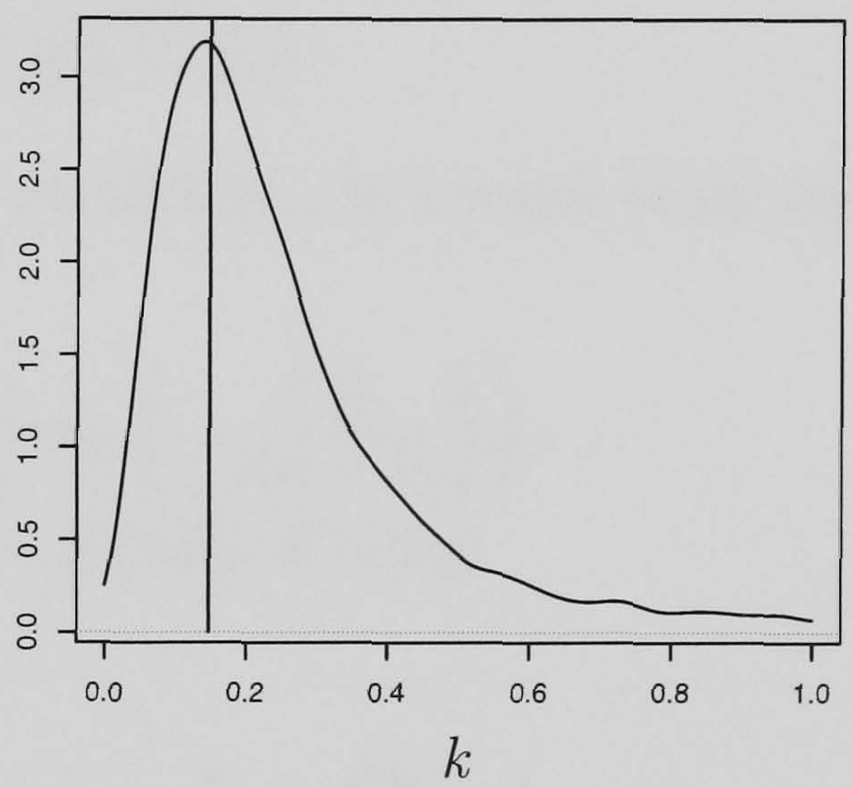
$MLR = 0.59912$



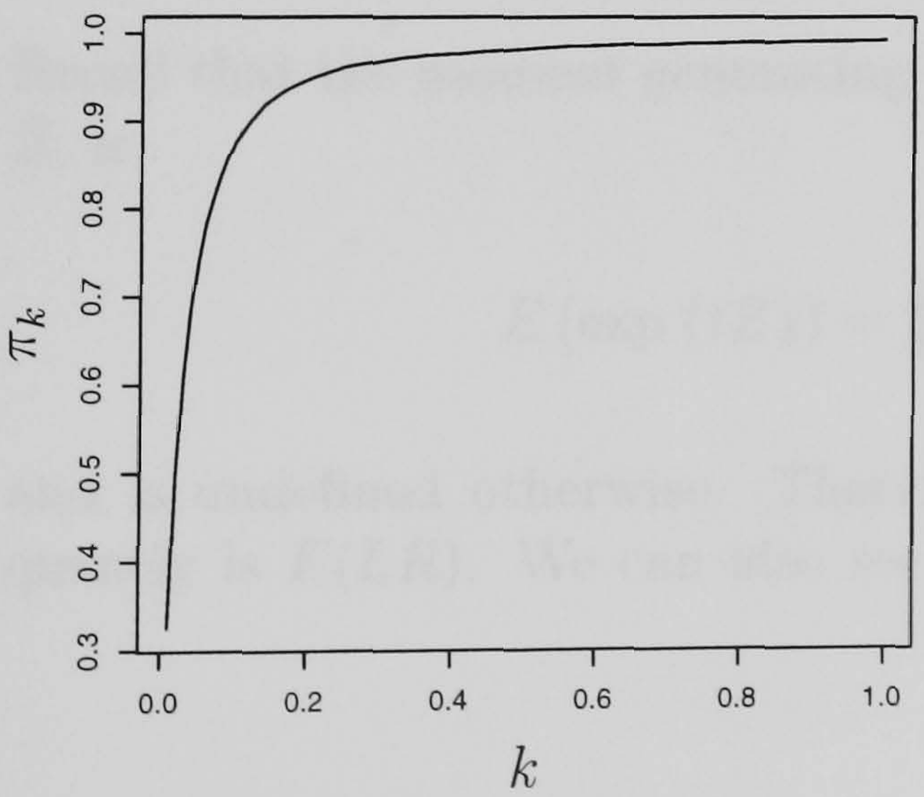


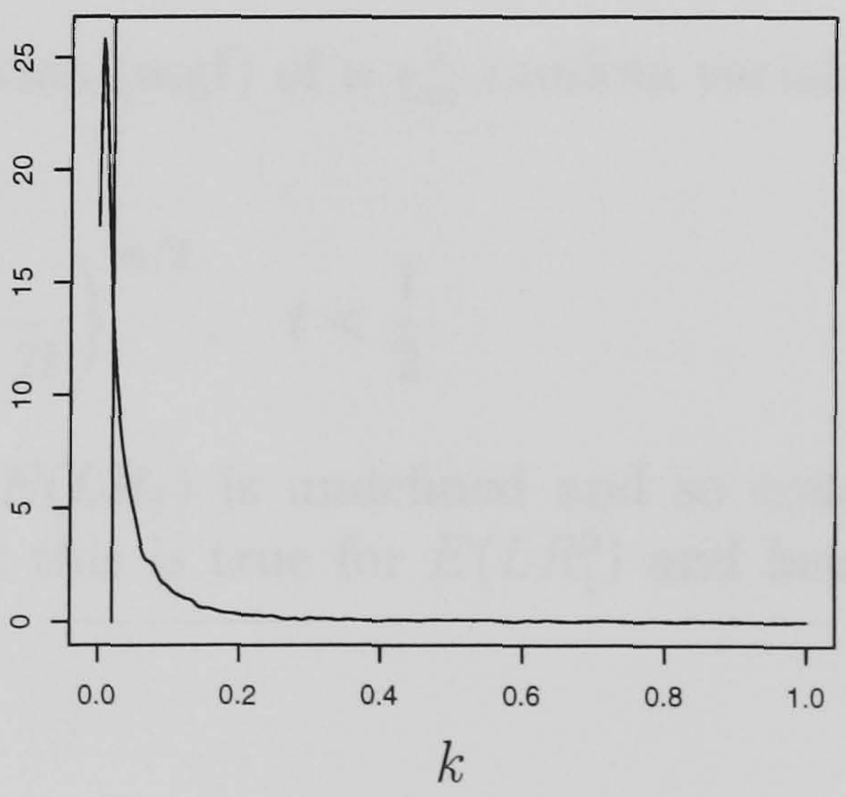
$MLR = 0.14743$





$MLR = 0.02021$





appears to lessen in severity as n increases. In the case of $t = 1$, the MLR is actually to the left of the peak, overstating the evidence against the null hypothesis. Again this overstatement lessens as n increases, so that we see that the MLR improves as a statement of evidence for larger sample sizes.

Examination of the graphs also suggests that we investigate the effect that t and n have on the mean, variance and skewness of the likelihood ratio since the density plots appear to exhibit the anticipated decrease in mean as t increases and as n increases. We also observe an apparent increase in skewness with t. Recall that

$$\begin{aligned} \pi_k &= Pr\left(\frac{t^2}{(n-1)}Y_2 - Y_1 > -2\log k\right) \\ &= Pr\left(\exp\left\{\frac{1}{2}\left[Y_1 - \frac{t^2}{(n-1)}Y_2\right]\right\} < k\right). \end{aligned}$$

We can characterise the distribution of π_k by considering the random variable

$$LR = \exp\left\{\frac{Y_1}{2} - \frac{t^2 Y_2}{2(n-1)}\right\},$$

and in particular $E(LR), Var(LR)$ and $Skew(LR)$. As a result of the independence of Y_1 and Y_2 we have

$$\begin{aligned} E\left(LR\right) &= E\left(\exp\left\{\frac{Y_1}{2}\right\}\exp\left\{-\frac{t^2 Y_2}{2(n-1)}\right\}\right), \\ &= E\left(LR_1 LR_2\right) = E\left(LR_1\right)E\left(LR_2\right) \end{aligned}$$

where

$$LR_1 = \exp\left\{\frac{Y_1}{2}\right\} \text{ and } LR_2 = \exp\left\{-\frac{t^2 Y_2}{2(n-1)}\right\}.$$

Recall that the moment generating function (mgf) of a χ^2_m random variable, Z, is

$$E\left(\exp\left\{tZ\right\}\right) = \left(\frac{1}{1-2t}\right)^{m/2}, \quad t < \frac{1}{2}$$

and is undefined otherwise. Therefore $E(LR_1)$ is undefined and so consequently is $E(LR)$. We can also see that this is true for $E(LR_1^2)$ and hence

for $Var(LR)$. Further

$$Skew\,(LR) = \frac{E\left([LR - E\,(LR)]^3\right)}{Var\,(LR)^{3/2}}$$

is also undefined as we can see from the mgf that this is the case for $E(LR_1^3)$. This behaviour is a result of the (very) heavy tail of the distribution of the likelihood ratio.

While we are not able to obtain summary statistics for LR itself, we can for $\log LR$, the log of the likelihood ratio. Here

$$\log LR = \frac{1}{2}Y_1 - \frac{t^2}{2(n-1)}Y_2.$$

We shall use the following results for a χ^2_m random variable Z:

$$\begin{aligned} E\,(Z) = m, \quad Var\,(Z) &= 2m, \\ E\,(Z^2) = m(m+2), \quad E\,(Z^3) &= m(m+2)(m+4), \\ Skew\,(Z) &= \sqrt{\frac{8}{m}}. \end{aligned}$$

Using these results we obtain

$$\begin{aligned} E\,(\log LR) &= \frac{1}{2} - \frac{t^2}{2}, \\ Var\,(\log LR) &= \frac{1}{2} + \frac{t^4}{2\,(n-1)}. \end{aligned} \qquad (*)$$

Note that it is possible for $E(\log LR)$ to be positive if $|t| < 1$. In this case, the null hypothesis is better supported (in expectation) than the alternative.

In order to evaluate the skewness

$$Skew\,(\log LR) = \frac{E\left([\log LR - E(\log LR)]^3\right)}{\{Var\,(\log LR)\}^{3/2}}$$

we require some preliminary calculations. Expanding $(\log LR)^3$, we obtain

$$\begin{aligned} E\left([\log LR]^3\right) &= \frac{1}{8}E\left(Y_1^3\right) - \frac{3t^2}{8(n-1)}E\left(Y_1^2\right)E\left(Y_2\right) \\ &\quad + \frac{3t^4}{8(n-1)^2}E\left(Y_1\right)E\left(Y_2^2\right) - \frac{t^6}{8(n-1)^3}E\left(Y_2^3\right), \\ &= \frac{15}{8} - \frac{9t^2}{8} + \frac{3t^4(n+1)}{8(n-1)} - \frac{t^6(n+1)(n+3)}{8(n-1)^2}. \end{aligned}$$

Also

$$\begin{aligned} E([\log LR &- E(\log LR)]^3) \\ &= E\left([\log LR]^3\right) - 3E\left([\log LR]^2\right)E\left(\log LR\right) + 2E\left(\log LR\right)^3. \end{aligned}$$

Using $E([\log LR]^2) = Var(\log LR) + E(\log LR)^2$ and the formulae given earlier we obtain

$$E([\log LR - E(\log LR)]^3) = 1 - \frac{t^6}{(n-1)^2}.$$

Combining this result with $(*)$ we can see that

$$Skew\left(\log LR\right) = \frac{1 - \dfrac{t^6}{(n-1)^2}}{\left(\dfrac{1}{2} + \dfrac{t^4}{2(n-1)}\right)^{3/2}}. \qquad (\dagger)$$

We now consider how these summary statistics behave as we let our sample size, n, tend to infinity. First, $E(\log LR)$ is actually independent of n and so remains fixed irrespective of the sample size. We also see that $Var(\log LR) \to 1/2$, a value independent of t. A similar effect is seen with the skewness, which, by considering $(\dagger)$ can be seen to tend to $2\sqrt{2}$. This is also independent of t and is the value of the skewness of a χ_1^2 random variable. This can be explained as we may observe that $\log LR$ becomes dominated by the term in Y_1, a χ_1^2 random variable, as $n \to \infty$.

We now relate our work to the case where the variance is known which we examined in Section 1.2 where we had $n = 1$ observation, x. We found that

$$\pi_k = Pr(Y_1 < 2\log k + x^2)$$

where $Y_1 \sim \chi^2_1$ as before. This means that the distribution of $2\log LR$ is concentrated above $-x^2$.

If we now consider the plots in Figures 2.1 to 2.3 we see that generally the density peaks show the pattern given in Table 2.9. The case for $n = 3$ does show a slightly different peak for $t = 1$ however, this being at around $k = 0.9$, $2\log k = -0.211$. We can observe from the table that the maximum

Table 2.9: Location of approximate peaks in density

t	k	$2\log k$
1	0.7	-0.713
2	0.1	-4.605
3	0.01	-9.210

density (of $2\log LR$) occurs around the value of $-t^2$ in these cases and we can also see from the plots that this distribution is similarly concentrated above this value as in the known variance case.

Chapter 3

The choice between two single-variable regressions

3.1 Introduction

We now consider a problem which was first presented by Pitman (1937). We are given the two-variable normal regression

$$M : y_i \mid x_{1i}, x_{2i} \sim N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \sigma^2), \quad i = 1, 2, \ldots, n$$

and we are interested in selecting the better of the two single-variable models using x_1 only or x_2 only. Williams (1959) gives a motivating example with two measures predicting wood density; further details of this example are given in Section 3.4. We formulate the two hypotheses $H_1 : \beta_2 = 0$, where only x_1 is needed, and $H_2 : \beta_1 = 0$, where only x_2 is required. We denote the models under H_1 and H_2 by M_1 and M_2 respectively and we wish to discover which of these hypotheses is better-supported by the data.

3.2 Model likelihoods

Let the vectors of observations be denoted by $\boldsymbol{x}_1 = (x_{1i})$, $\boldsymbol{x}_2 = (x_{2i})$, and define the design matrices $X = [\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2]$, $X_1 = [\mathbf{1}, \boldsymbol{x}_1]$, $X_2 = [\mathbf{1}, \boldsymbol{x}_2]$, with corresponding parameter vectors $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2)$, $\boldsymbol{\beta}_1^T = (\beta_0, \beta_1)$, $\boldsymbol{\beta}_2^T = (\beta_0, \beta_2)$.

The full model likelihood under M is

$$\begin{aligned} L(\beta_0, \beta_1, \beta_2, \sigma) &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta})\right\}, \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}[RSS + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]\right\}, \end{aligned}$$

where $RSS = (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$ is the residual sum of squares evaluated at the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$.

We have

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(0, \sigma^2(X^T X)^{-1}), \quad \frac{RSS}{\sigma^2} \sim \chi^2_{n-3}$$

using either frequentist or Bayes assumptions. Under the second, the distribution of $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is conditional on σ.

For the two sub-models, the likelihood for M_1 can be written as

$$L_1 = L(\beta_0, \beta_1, 0, \sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}[RSS_1 + (\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)^T X_1^T X_1(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)]\right\},$$

while for M_2

$$L_2 = L(\beta_0, 0, \beta_2, \sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}[RSS_2 + (\tilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)^T X_2^T X_2(\tilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)]\right\},$$

where for $j = 1, 2$

$$\tilde{\boldsymbol{\beta}}_j = (X_j^T X_j)^{-1} X_j^T \boldsymbol{y},$$

and

$$RSS_j = (\boldsymbol{y} - X_j\tilde{\boldsymbol{\beta}}_j)^T(\boldsymbol{y} - X_j\tilde{\boldsymbol{\beta}}_j).$$

The likelihood ratio between M_1 and M_2 is

$$LR = \frac{L_1}{L_2} = \exp\left\{-\frac{1}{2\sigma^2}(RSS_1 - RSS_2 + Q_1 - Q_2)\right\},$$

where Q_1 and Q_2 are quadratic forms in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, with

$$Q_j = \left(\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\right)^T X_j^T X_j \left(\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\right).$$

Therefore

$$-2\log LR = \frac{1}{\sigma^2}(RSS_1 - RSS_2 + Q_1 - Q_2).$$

We now assume, without any loss of generality (since the origin of x_1 and x_2 can be absorbed into β_0), that x_1 and x_2 are centred, so that $\sum x_{1i} = \sum x_{2i} = 0$. Then, for $j, k = 1, 2$, we define

$$S_{jk} = \sum x_{ji}x_{ki},\ S_{jy} = \sum x_{ji}y_i,\ S_y = \sum y_i,$$

and

$$X_j^T X_j = \begin{bmatrix} n & 0 \\ 0 & S_{jj} \end{bmatrix},\ X_j^T y = \begin{bmatrix} S_y \\ S_{jy} \end{bmatrix}.$$

Then for the full model

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix},\ S^{-1} = \begin{bmatrix} S^{11} & S^{12} \\ S^{12} & S^{22} \end{bmatrix},$$

$$X^T X = \begin{bmatrix} n & 0 \\ 0 & S \end{bmatrix},\ (X^T X)^{-1} = \begin{bmatrix} 1/n & 0 \\ 0 & S^{-1} \end{bmatrix},$$

$$X^T y = \begin{bmatrix} S_y \\ S_{1y} \\ S_{2y} \end{bmatrix}.$$

Now, using a diffuse prior distribution, we have

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \Big| \sigma \sim N\left[\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}, \sigma^2 S^{-1} \right],$$

Therefore, the posterior distribution of the likelihood ratio is

$$\pi_k = Pr(LR < k) = Pr\left(\frac{RSS_1 - RSS_2}{\sigma^2} + \frac{Q_1 - Q_2}{\sigma^2} > -2\log k \right).$$

Using our earlier results we can see that

$$\begin{aligned} Q_1 - Q_2 &= n(\bar{y} - \beta_0)^2 + S_{11}(\tilde{\beta}_1 - \beta_1)^2 - [n(\bar{y} - \beta_0)^2 + S_{22}(\tilde{\beta}_2 - \beta_2)^2] \\ &= S_{11}(\tilde{\beta}_1 - \beta_1)^2 - S_{22}(\tilde{\beta}_2 - \beta_2)^2. \end{aligned}$$

We note that this quadratic form in (β_1, β_2) is not centred at the full MLE $(\hat{\beta}_1, \hat{\beta}_2)$ but at the MLEs under the two sub-models.

Now

$$\frac{RSS_1 - RSS_2}{\sigma^2} = \frac{RSS}{\sigma^2} \times \frac{RSS_1 - RSS_2}{RSS},$$

where the second term is the model comparison of the two hypotheses relative to the full model. Let

$$t = \frac{RSS_1 - RSS_2}{RSS}$$

then

$$\pi_k = Pr\left(t\frac{RSS}{\sigma^2} + \frac{Q_1 - Q_2}{\sigma^2} > -2\log k\right).$$

If we now define,

$$\begin{aligned} \gamma_j &= \frac{\beta_j - \tilde{\beta}_j}{\sigma} \text{ and} \\ \hat{\gamma}_j &= \frac{\beta_j - \beta_j}{\sigma} \end{aligned}$$

then

$$\frac{Q_1 - Q_2}{\sigma^2} = S_{11}\gamma_1^2 - S_{22}\gamma_2^2.$$

We see that the form for π_k once more requires either numerical integration or simulation in order to obtain values for this probability. In light of our previous experience we shall use a simulation method, analagous to that of the previous chapter.

In order to simulate from $(Q_1 - Q_2)/\sigma^2$ we have to be able to simulate values of γ_j for $j = 1, 2$. We know that

$$\left(\begin{array}{c} \gamma_1 - \hat{\gamma}_1 \\ \gamma_2 - \hat{\gamma}_2 \end{array}\right) \Big| \sigma \sim N\left(\mathbf{0}, S^{-1}\right)$$

and if we now observe that we can write γ_j as

$$\begin{aligned} \gamma_j &= (\gamma_j - \hat{\gamma}_j) + \hat{\gamma}_j, \\ &= (\gamma_j - \hat{\gamma}_j) + \frac{\hat{\beta}_j - \tilde{\beta}_j}{\sigma}, \end{aligned}$$

we see that, as we can simulate $(\hat{\beta}_i - \tilde{\beta}_i)/\sigma$ using $RSS/\sigma^2 \sim \chi^2_{n-3}$ to obtain a value for the unknown (random) quantity σ, we are able to simulate from $(Q_1 - Q_2)/\sigma$ by obtaining realisations from $\gamma_i - \hat{\gamma}_i$. This quantity has a bivariate normal distribution with a (possibly) non-diagonal covariance matrix.

3.3 Simulation

Let $\eta = -2\log k$ and

$$-2\log LR = t\frac{RSS}{\sigma^2} + \frac{Q_1 - Q_2}{\sigma^2}.$$

Recall that $\pi_k = Pr[-2\log LR > \eta]$, $\gamma_j = (\beta_j - \tilde{\beta}_j)/\sigma$ and

$$\frac{Q_1 - Q_2}{\sigma^2} = S_{11}\gamma_1^2 - S_{22}\gamma_2^2. \qquad (*)$$

We know that, for a given dataset, we can calculate S_{11}, S_{12} and S_{22}. Standard linear regression techniques provide estimates of $\hat{\beta}_j$ and $\tilde{\beta}_j$, and, thereby values for S^{-1}, RSS, RSS_1 and RSS_2. We can therefore calculate

$$t = \frac{RSS_1 - RSS_2}{RSS}.$$

We can easily simulate from $RSS/\sigma^2 \sim \chi^2_{n-3}$ then obtain realisations from $(Q_1 - Q_2)/\sigma^2$ in the following manner, by simulating values of γ_j.

We can obtain an observation from $(Q_1 - Q_2)/\sigma^2$ by substituting realisations of γ_1 and γ_2 into $(*)$. We can obtain the sample values of γ_j via the standard method for multivariate normal sampling for realisations of $(\gamma_j - \hat{\gamma}_j)$. We know

$$\begin{pmatrix} \gamma_1 - \hat{\gamma}_1 \\ \gamma_2 - \hat{\gamma}_2 \end{pmatrix} \sim N\left(\mathbf{0}, S^{-1}\right).$$

We firstly simulate from

$$(\gamma_2 - \hat{\gamma}_2) \sim N(0, S^{22}),$$

and then from

$$(\gamma_1 - \hat{\gamma}_1)|(\gamma_2 - \hat{\gamma}_2) \sim N\left(\frac{S^{12}(\gamma_2 - \hat{\gamma}_2)}{S^{22}}, S^{11} - \frac{(S^{12})^2}{S^{22}}\right).$$

We then use our simulated value of RSS/σ^2 to obtain a value of σ in order to form

$$\frac{Q_1 - Q_2}{\sigma^2} = S_{11}\left((\gamma_1 - \hat{\gamma}_1) + \frac{\hat{\beta}_1 - \tilde{\beta}_1}{\sigma}\right)^2 - S_{22}\left((\gamma_2 - \hat{\gamma}_2) + \frac{\hat{\beta}_2 - \tilde{\beta}_2}{\sigma}\right)^2.$$

Hence, we can simulate the necessary value required for our simulation based procedure to estimate π_k. We generate a large number (N) of observations from

$$\left(\chi^2_{n-3}, \frac{Q_1 - Q_2}{\sigma^2}\right)$$

and for each set of values we then evaluate L. We then simply ascertain the number of pairs for which $L > \eta$ for our chosen η and divide this by N to give an approximate value for π_k. As in the previous chapter we have that the estimated variance of our estimator $\hat{\pi}_k$ is

$$\widehat{Var}\left(\hat{\pi}_k\right) = \frac{\hat{\pi}_k\left(1 - \hat{\pi}_k\right)}{N}.$$

Once more, we see that this has a maximum of $1/(4N)$ and hence we can choose N such that we obtain a desired level of minimum accuracy.

3.4 Bayes factor and maximised likelihood ratio

For comparison purposes we shall also calculate the Bayes factor and the maximised likelihood ratio for this example. We proceed in the following fashion.

3.4.1 Bayes factor calculation

We define the Bayes factor (in favour of M_1, the model using $\boldsymbol{x}_1$ only, over M_2) to be

$$B = \frac{\int_{\sigma=0}^{\infty}\int_{\beta_1=-\infty}^{\infty}\int_{\beta_0=-\infty}^{\infty} L(\beta_0,\beta_1,0,\sigma)\pi_{\beta_0}(\beta_0)\pi_{\beta_1}(\beta_1)\pi_{\sigma}(\sigma)\,\mathrm{d}\beta_0\,\mathrm{d}\beta_1\,\mathrm{d}\sigma}{\int_{\sigma=0}^{\infty}\int_{\beta_2=-\infty}^{\infty}\int_{\beta_0=-\infty}^{\infty} L(\beta_0,0,\beta_2,\sigma)\pi_{\beta_0}(\beta_0)\pi_{\beta_2}(\beta_2)\pi_{\sigma}(\sigma)\,\mathrm{d}\beta_0\,\mathrm{d}\beta_2\,\mathrm{d}\sigma},$$

where the (improper) prior distributions for $\beta_0, \beta_1, \beta_2$ and σ are taken to be diffuse and are given by

$$\pi_{\beta_0}(\beta) = \pi_{\beta_1}(\beta) = \pi_{\beta_2}(\beta) = k_1, \quad -\infty < \beta < \infty$$

and

$$\pi_\sigma(\sigma) = k_2, \quad 0 < \sigma < \infty.$$

Note that we may ignore the constants of proportionality k_1 and k_2 as they will cancel when we calculate the ratio of integrals. Using these priors the Bayes factor can be written as

$$B = \frac{\int L_1 \, \mathrm{d}\boldsymbol{\beta}_1 \, \mathrm{d}\sigma}{\int L_2 \, \mathrm{d}\boldsymbol{\beta}_2 \, \mathrm{d}\sigma}$$

where, as defined previously for $j = 1, 2$,

$$L_j = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}[RSS_j + (\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^T X_j^T X_j (\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)]\right\},$$

$\boldsymbol{\beta}_j^T = (\beta_0, \beta_j)$, RSS_j is the residual sum of squares and $\tilde{\boldsymbol{\beta}}_j$ is the maximum likelihood estimate of $\boldsymbol{\beta}_j$ for model M_j. Aitkin (1991, pp. 119) shows that

$$\int L_j \, \mathrm{d}\boldsymbol{\beta}_j \, \mathrm{d}\sigma = 2^{-3/2}\pi^{-(n-2)/2}\Gamma\left(\frac{n-3}{2}\right)|X_j^T X_j|^{-1/2} RSS_j^{-(n-3)/2}.$$

Earlier we saw that

$$X_j^T X_j = \begin{bmatrix} n & 0 \\ 0 & S_{jj} \end{bmatrix},$$

where $S_{jj} = \sum_i x_{ji}^2$, and so $|X_j^T X_j| = nS_{jj}$. Therefore the Bayes factor is

$$\begin{aligned} B &= \left(\frac{nS_{11}}{nS_{22}}\right)^{-1/2} \left[\frac{RSS_1}{RSS_2}\right]^{-(n-3)/2} \\ &= \sqrt{\frac{S_{22}}{S_{11}}} \left(\frac{RSS_2}{RSS_1}\right)^{(n-3)/2}. \end{aligned}$$

We note that the second term of this expression is a function of ratio of the residual sum of squares for the two sub-models.

3.4.2 Maximised likelihood ratio calculation

Writing the full model likelihood from Section 3.2 as

$$\begin{aligned} L(\beta_0, \beta_1, \beta_2, \sigma) &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta})\right\}, \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2\right\}, \end{aligned}$$

we define the maximised likelihood ratio to be

$$MLR = \frac{L(\tilde{\beta}_0^{(1)}, \tilde{\beta}_1^{(1)}, 0, \tilde{\sigma}^{(1)})}{L(\tilde{\beta}_0^{(2)}, 0, \tilde{\beta}_2^{(2)}, \tilde{\sigma}^{(2)})}.$$

Here we consider the maximum likelihood estimates of the parameters under the two sub-models. We denote the estimate of β_j under model M_i by $\beta_j^{(i)}$ and use the same notation for the estimates of σ. The residual sum of squares for each sub-model are

$$RSS_1 = \sum_{i=1}^{n}\left(y_i - \tilde{\beta}_0^{(1)} - \tilde{\beta}_1^{(1)} x_{1i}\right)^2$$

and

$$RSS_2 = \sum_{i=1}^{n}\left(y_i - \tilde{\beta}_0^{(2)} - \tilde{\beta}_2^{(2)} x_{2i}\right)^2.$$

By differentiation of the likelihood function with respect to σ and setting the resultant form equal to zero we find that, under model M_i for $i = 1, 2$,

$$\tilde{\sigma}^{(i)2} = \frac{RSS_i}{n}.$$

Hence, we see that the maximised likelihood ratio is given by

$$\begin{aligned} MLR &= \frac{(n/(2\pi))^{n/2}\, RSS_1^{-n/2} e^{-n/2}}{(n/(2\pi))^{n/2}\, RSS_2^{-n/2} e^{-n/2}} \\ &= \left(\frac{RSS_2}{RSS_1}\right)^{n/2}. \end{aligned}$$

This is simply a function of the ratio of the residual sum of squares for the two sub-models. This fact, when considered in conjuction with the similar

Table 3.1: Strength of radiata pine with density and resin-adjusted density

Strength	Density	Adjusted Density	Strength	Density	Adjusted Density
3040	29.2	25.4	2470	24.7	22.2
3610	32.3	32.2	3480	31.3	31.0
3810	31.5	30.9	2330	24.5	23.9
1800	19.9	19.2	3110	27.3	27.2
3160	27.1	26.3	2310	24.0	23.9
4360	33.8	33.2	1880	21.5	21.0
3670	32.2	29.0	1740	22.5	22.0
2250	27.5	23.8	2650	25.6	25.3
4970	34.5	34.2	2620	26.2	25.7
2900	26.7	26.4	1670	21.1	20.0
2540	24.1	23.9	3840	30.7	30.7
3800	32.7	32.6	4600	32.6	32.5
1900	22.1	20.8	2530	25.3	23.1
2920	30.8	29.8	4990	38.9	38.1
1670	22.1	21.3	3310	29.2	28.5
3450	30.1	29.2	3600	31.4	31.4
2850	26.7	25.9	1590	22.1	21.4
3770	30.3	29.8	3850	32.0	30.6
2480	23.2	22.6	3570	30.3	30.3
2620	29.9	23.8	1890	20.8	18.4
3030	33.2	29.4	3030	28.2	28.2

form of the Bayes factor, makes the relative values of RSS_1 and RSS_2 a useful guide as to which model we should prefer.

3.5 Example

We consider the dataset used in the Williams (1959) example. We are given 42 observations of the strength of radiata pine, together with the corresponding density and density adjusted for resin content. The data are shown in Table 3.1. We are interested in making a choice between the two single-variable models, each using only one of the density measures. We make this choice as the two explanatory variables are essentially functions of one another as they are measuring the same quantity in two different ways.

In terms of our original model, M, we represent the strength by y_i, the density by x_{1i} and the resin-adjusted density by x_{2i} for $i = 1, 2, \ldots, 42$. Our hypotheses now take the following meanings:

H_1 selects the model with the non-adjusted density measure only and H_2 the model with the resin-adjusted density measure only.

In our derivation we assumed that both covariates summed to zero, and therefore we use new (centred) covariates

$$\tilde{\boldsymbol{x}}_j = \boldsymbol{x}_j - \bar{\boldsymbol{x}}_j, \quad j = 1, 2.$$

At this point we recall that RSS_j represents the residual sum of squares when fitting model M_j, and that M_1 is the model using only $\boldsymbol{x}_1$ as a regressor and M_2 is the model using only $\boldsymbol{x}_2$. The following key summaries can be calculated from the data:

$$\begin{aligned} \bar{x}_1 = 27.85952, \quad & \bar{x}_2 = 26.78810, \quad RSS = 2979320, \\ RSS_1 = 4602769, \quad & RSS_2 = 3066459, \quad t = 0.5156578, \\ \hat{\beta}_1 = 35.92847, \quad & \hat{\beta}_2 = 149.9736, \quad \tilde{\beta}_1 = 184.5528, \\ \tilde{\beta}_2 = 183.2733, \quad & S_{11} = 828.2412, \quad S_{22} = 885.5840. \end{aligned}$$

It is clear from the values of RSS, RSS_1 and RSS_2 that the data favour M_2, the resin-adjusted density model, as there is little difference in RSS between the full model and the model using $\boldsymbol{x}_2$ alone. This means that our method should favour using the resin-adjusted density.

We also obtain

$$S = \begin{pmatrix} 828.2412 & 820.7898 \\ 820.7898 & 885.5840 \end{pmatrix},$$

with inverse

$$S^{-1} = \begin{pmatrix} 0.01481 & -0.01373 \\ -0.01373 & 0.01385 \end{pmatrix}.$$

We now simulate $N = 100000$ observations from χ^2_{n-3} and the random quantities $(\gamma_2 - \hat{\gamma}_2)$ and $(\gamma_1 - \hat{\gamma}_1)|(\gamma_2 - \hat{\gamma}_2)$, using

$$(\gamma_2 - \hat{\gamma}_2) \sim N(0, 0.01385),$$

and

$$(\gamma_1 - \hat{\gamma}_1)|(\gamma_2 - \hat{\gamma}_2) \sim N(-0.99100(\gamma_2 - \hat{\gamma}_2), 0.00121).$$

Table 3.2 shows the the estimates obtained for π_k.

Table 3.2: Approximate π_k values of k

k	π_k	k	π_k
2.0	0.972	1.9	0.973
1.8	0.972	1.7	0.972
1.6	0.972	1.5	0.972
1.4	0.971	1.3	0.971
1.2	0.971	1.1	0.971
1.0	0.971	0.9	0.971
0.8	0.971	0.7	0.971
0.6	0.971	0.5	0.970
0.4	0.970	0.3	0.970
0.2	0.969	0.1	0.969

From inspection of Table 3.2, we can see that we have a very strong preference towards H_2. This is particularly shown by the large value of π_1. Hence we, once more, see a strong preference towards the resin-adjusted density model.

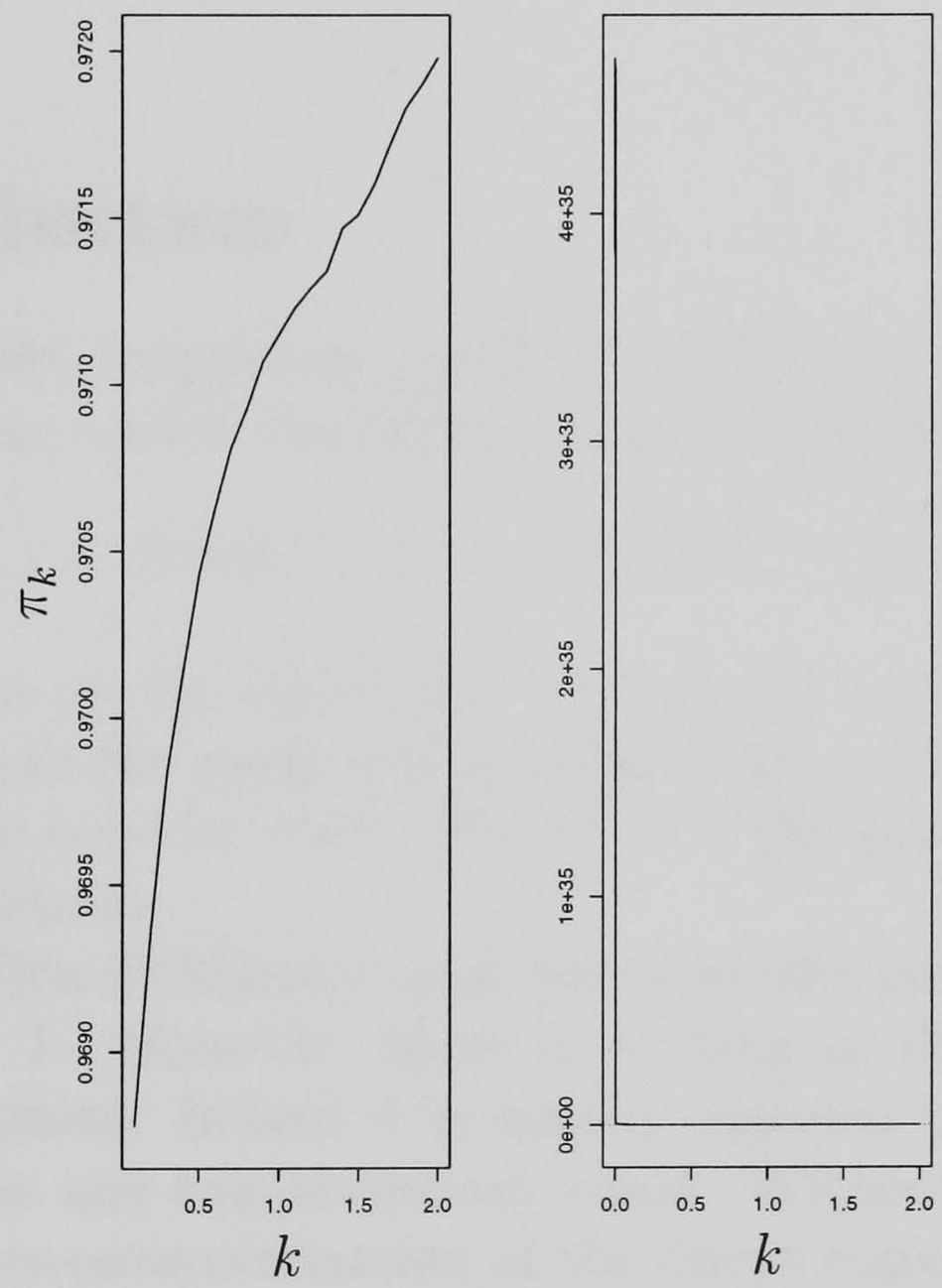
Figure 3.1 gives a plot of the (approximate) cumulative distribution function of the likelihood ratio by plotting π_k against k. Alongside this we show the corresponding density estimate, again obtained using the software package R. This figure shows that the density is concentrated near $k = 0$, once more indicating very strong support for H_2 for these data. We are also able to calculate both the Bayes factor and the maximised likelihood ratio for this example:

$$B = \sqrt{\frac{885.5840}{828.2412}} \left(\frac{3066459}{4602769}\right)^{39/2} = 0.000376,$$

$$MLR = \left(\frac{3066459}{4602769}\right)^{21} = 0.000198.$$

These results agree with the conclusion of *extremely* strong support for M_2 which was given by our π_k method.

Figure 3.1: Distribution (left) & density (right) functions for the Williams data



Chapter 4

The choice between random walk and AR(1) time series

4.1 Introduction

We adapt the model comparison problem of Marriott and Newbold (1998) to a slightly simpler model. Specifically consider the AR(1) model

$$X_i \mid X_{i-1} \sim N(\phi X_{i-1} + (1-\phi)\mu, \sigma^2), \quad i = 1, 2, \ldots, n$$

where we condition on the initial X_0. This model is stationary for $|\phi| < 1$, and our interest is in the point null hypothesis $H_0 : \phi = 1$, where the model is a non-stationary random walk. For $|\phi| > 1$ the model is non-stationary with increasing variance.

In the formulation of Marriott and Newbold, the parameter space is constrained to $|\phi| \leq 1$. However, there is nothing in the above model that requires this constraint. Indeed it is usually imposed (artificially) after differencing to remove any non-stationary trend. Without this constraint, the model is a simple re-parametrization of the linear regression model

$$X_i \mid X_{i-1} \sim N(\alpha + \beta X_{i-1}, \sigma^2)$$

with $\beta = \phi$ and $\alpha = (1-\phi)\mu$. Under the null hypothesis $H_0 : \phi = 1$, $(\alpha = 0, \beta = 1)$, the likelihood function is flat in μ. This is a rank deficiency in the model which causes no difficulty however in our analysis.

4.2 Likelihood ratio

The likelihood function (omitting multiplicative constants) is

$$L(\mu, \sigma, \phi) = \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (x_i - \phi x_{i-1} - (1-\phi)\mu)^2 \right] \right\}$$

and under $H_0 : \phi = 1$, this is

$$L(\mu, \sigma, 1) = \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - x_{i-1})^2 \right\}.$$

The likelihood ratio is

$$\begin{aligned} LR &= \frac{L(\mu, \sigma, 1)}{L(\mu, \sigma, \phi)} \\ &= \exp\left\{ \frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} \{x_i - \phi x_{i-1} - (1-\phi)\mu\}^2 - \sum_{i=1}^{n} (x_i - x_{i-1})^2 \right] \right\}. \end{aligned}$$

We now re-parameterize to α and β. Using diffuse priors, the posterior distributions of α, β and σ are given by standard results from regression analysis (in the absence of parameter constraints) as

$$\theta = (\alpha, \beta)^T \mid x, \sigma \sim N\left(\left(\hat{\alpha}, \hat{\beta}\right)^T, \sigma^2 (X^T X)^{-1} \right), \quad \frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$$

where $\hat{\alpha}$, $\hat{\beta}$ are the MLEs, X is the design matrix $[\mathbf{1}, \boldsymbol{x}]$, where

$$\boldsymbol{x}^T = (x_0, x_1, \ldots, x_{n-1}) \text{ and } \mathbf{1}^T = (1, \ldots, 1).$$

If we define

$$\boldsymbol{y}^T = (x_1, x_2, \ldots, x_n),$$

the residual sum of squares is

$$RSS = (\boldsymbol{y} - X\hat{\theta})^T (\boldsymbol{y} - X\hat{\theta}).$$

Then

$$LR = \exp\left\{ \frac{1}{2\sigma^2} \left\{ RSS + (\theta - \hat{\theta})^T X^T X (\theta - \hat{\theta}) - D \right\} \right\}$$

and

$$-2\log LR = -\frac{1}{\sigma^2}\left\{RSS + (\theta-\hat{\theta})^T X^T X(\theta-\hat{\theta}) - D\right\},$$

where D is the sum of squared lag 1 differences. Therefore the posterior probability $\pi_k = Pr(LR < k)$ is

$$\begin{aligned}
\pi_k &= Pr\left\{\frac{RSS-D}{\sigma^2} + \frac{1}{\sigma^2}(\theta-\hat{\theta})^T X^T X(\theta-\hat{\theta}) < 2\log k\right\} \\
&= Pr\left\{\frac{RSS}{\sigma^2} \times \frac{RSS-D}{RSS} + \frac{1}{\sigma^2}(\theta-\hat{\theta})^T X^T X(\theta-\hat{\theta}) < 2\log k\right\} \\
&= Pr\left\{-\frac{2\tilde{F}}{n-2}Y_2 + Y_1 < 2\log k\right\}
\end{aligned}$$

where $Y_1 \sim \chi^2_2$, $Y_2 \sim \chi^2_{n-2}$ and

$$\tilde{F} = \frac{(D-RSS)/2}{RSS/(n-2)}$$

is the F-statistic for testing $H_0 : \phi = 1 \quad (\alpha = 0, \beta = 1)$ against $H_1 : \phi \neq 1$ by the usual ANOVA. Again the posterior distribution is a (weighted) difference of central χ^2 random variables.

As before we first consider the case when $k = 1$. Here we obtain

$$\begin{aligned}
\pi_1 &= Pr\left[Y_1 < \frac{2\tilde{F}}{n-2}Y_2\right] \\
&= Pr\left[F_{2,n-2} < \tilde{F}\right].
\end{aligned}$$

This is easily found by reference to the appropriate F distribution table.

We now return to the general case where we firstly take $0 \leq k \leq 1$ so that $\log k \leq 0$. For this particular example we are able to evaluate π_k analytically in the following way.

4.3 Evaluation of π_k

If we now once more define

$$\eta = -2\log k$$

and note that $\eta \geq 0$ then we can see that we have the following general form for π_k:

$$\pi_k = Pr\left(aY_2 - Y_1 > \eta\right)$$

where

$$a = \frac{2\tilde{F}}{n-2}.$$

We now consider the various possibilities for a and we do this by considering various values that $\tilde{F}$ can assume. We should first note that, for indentifiability of the three parameters, α, β, σ, we restrict ourselves to $n > 2$. We can also see that if we allowed $n = 2$ we would, in this case, be faced with $\tilde{F} = 0$. This would leave us unable to evaluate a and hence unable to obtain π_k.

A slight rearrangement of the form for $\tilde{F}$ gives

$$\tilde{F} = \frac{n-2}{2}\left(\frac{D}{RSS} - 1\right).$$

Under H_0, the sampling distribution of $\tilde{F}$ has the form of an F distribution and hence $F \geq 0$. Now, if $F = 0$ then $a = 0$ and we see that in this case

$$\pi_k = Pr\left(-Y_1 > \eta\right),$$

the probability of a (random) negative number being larger than a positive number. Therefore, in this case, $\pi_k = 0$.

We now concern ourselves with the case $0 < \tilde{F} < \infty$. In this range we know that $a > 0$ and we also know that $\eta \geq 0$. We are interested in obtaining a value for

$$\pi_k = Pr\left(aY_2 - Y_1 > \eta\right).$$

Y_1 has density function

$$f_{Y_1}(y_1) = \frac{\exp\left\{-\frac{y_1}{2}\right\}}{2}$$

and Y_2

$$f_{Y_2}(y_2) = \frac{\exp\left\{-\frac{y_2}{2}\right\} y_2^{\frac{n-4}{2}}}{2^{\frac{n-2}{2}}\Gamma\left(\frac{n-2}{2}\right)}.$$

We also note that Y_1 and Y_2 are independent. We see that we may re-arrange the expression $ay_2 - y_1 > \eta \geq 0$ as $0 < y_1 < ay_2 - \eta$ and hence obtain $y_2 > \frac{\eta}{a}$. Using these results to give our ranges of integration we proceed as follows:

$$\begin{aligned}
\pi_k &= \int_{\eta/a}^{\infty} f_{Y_2}(y_2) \left(\int_0^{ay_2-\eta} f_{Y_1}(y_1)\, \mathrm{d}y_1 \right) \mathrm{d}y_2, \\
&= \int_{\eta/a}^{\infty} f_{Y_2}(y_2) \left(1 - \exp\left\{ -\frac{ay_2 - \eta}{2} \right\} \right) \mathrm{d}y_2, \\
&= \int_{\eta/a}^{\infty} f_{Y_2}(y_2)\, \mathrm{d}y_2 - \int_{\eta/a}^{\infty} f_{Y_2}(y_2) \exp\left\{ -\frac{ay_2 - \eta}{2} \right\} \mathrm{d}y_2.
\end{aligned}$$

Now let us define $G(x)$ to be the (cumulative) distribution function of a χ^2_{n-2} random variable and also let $g(x) = G'(x)$ be the corresponding density function. Then

$$g(x) = \frac{\exp\left\{\frac{x}{2}\right\} x^{\frac{n-4}{2}}}{2^{\frac{n-2}{2}} \Gamma\left(\frac{n-2}{2}\right)}, \quad x > 0.$$

Recall now that $Y_2 \sim \chi^2_{n-2}$, so that

$$\pi_k = \left\{ 1 - G\left(\frac{\eta}{a}\right) \right\} - \int_{\eta/a}^{\infty} f_{Y_2}(y_2) \exp\left\{ -\frac{ay_2 - \eta}{2} \right\} \mathrm{d}y_2.$$

If we now re-arrange the integrand we obtain

$$\pi_k = \left\{ 1 - G\left(\frac{\eta}{a}\right) \right\} - \frac{\exp\left\{\frac{\eta}{2}\right\}}{2^{\frac{n-2}{2}} \Gamma\left(\frac{n-2}{2}\right)} \int_{\eta/a}^{\infty} y_2^{\frac{n-4}{2}} \exp\left\{ -\frac{(a+1)y_2}{2} \right\} \mathrm{d}y_2.$$

Changing variables from y_2 to $y_3 = (a+1)y_2$ we obtain

$$\pi_k = \left\{ 1 - G\left(\frac{\eta}{a}\right) \right\} - \frac{\exp\left\{\frac{\eta}{2}\right\}}{2^{\frac{n-2}{2}} \Gamma\left(\frac{n-2}{2}\right) (a+1)^{\frac{n-2}{2}}} \int_{\eta+\eta/a}^{\infty} y_3^{\frac{n-4}{2}} \exp\left\{ -\frac{y_3}{2} \right\} \mathrm{d}y_3.$$

We now see that the integrand has the same general form as the density of a χ^2_{n-2} distribution so that we find

$$\pi_k = \left\{ 1 - G\left(\frac{\eta}{a}\right) \right\} - \frac{\exp\left\{\frac{\eta}{2}\right\}}{(a+1)^{\frac{n-2}{2}}} \left\{ 1 - G\left(\eta + \frac{\eta}{a}\right) \right\}.$$

If we now substitute our expressions for a and η we see that

$$\begin{aligned}\pi_k &= \left\{1 - G\left(-\frac{(n-2)\log k}{\tilde{F}}\right)\right\} \\ &\quad - \frac{(n-2)^{\frac{n-2}{2}}}{k\left(n-2+2\tilde{F}\right)^{\frac{n-2}{2}}}\left\{1 - G\left(-2\log k - \frac{(n-2)\log k}{\tilde{F}}\right)\right\}.\end{aligned}$$

We now note that if we substitute $\tilde{F} = 0$ into the above form we obtain $\pi_k = 0$ so that when $n > 2$ this form holds for $0 \leq \tilde{F} < \infty$.

Evaluation of π_k in the case $1 \leq k < \infty$ proceeds in a similar fashion. Here we have that $\eta \leq 0$ which implies that $y_2 > 0 \geq \eta/a$ and hence

$$\pi_k = \int_0^\infty f_{Y_2}(y_2)\left(\int_0^{ay_2-\eta} f_{Y_1}(y_1)\,\mathrm{d}y_1\right)\,\mathrm{d}y_2.$$

We obtain:

$$\begin{aligned}\pi_k &= \{1 - G(0)\} - \frac{\exp\left\{\frac{\eta}{2}\right\}}{(a+1)^{\frac{n-2}{2}}}\{1 - G(0)\}, \\ &= 1 - \frac{(n-2)^{\frac{n-2}{2}}}{k\left(n-2+2\tilde{F}\right)^{\frac{n-2}{2}}}, \quad 0 \leq \tilde{F} < \infty.\end{aligned}$$

4.4 Results

We now proceed as in the work on testing a normal mean with unknown σ by calculating π_k for a range of values of $\tilde{F}$ and $0 \leq k \leq 1$. We display results for $n = 10, 50, 100$ in Tables 4.1 to 4.3.

4.5 Discussion

If we recall that higher values of π_k indicate that we are more likely to reject the random walk model in favour of the AR(1) model with parameter $\phi \neq 1$, we can see that, as we would expect, π_k increases with both k and $\tilde{F}$. If we now consider a fixed $\tilde{F}$ and look at the behaviour of π_k when we increase n, we again see that here π_k is increasing, meaning that, as would be expected, for larger n we require a smaller value of $\tilde{F}$ to provide convincing evidence against the random walk model than for smaller n.

Table 4.1: π_k & P for $n = 10$

	k										
$\tilde{F}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	P
0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
0.2	0.000	0.000	0.000	0.000	0.000	0.001	0.005	0.028	0.091	0.177	0.823
0.4	0.000	0.000	0.000	0.002	0.011	0.036	0.085	0.158	0.242	0.317	0.683
0.6	0.000	0.001	0.007	0.027	0.067	0.128	0.206	0.289	0.365	0.428	0.572
0.8	0.001	0.009	0.036	0.085	0.155	0.237	0.321	0.399	0.464	0.518	0.482
1.0	0.005	0.032	0.088	0.165	0.252	0.340	0.420	0.489	0.545	0.590	0.410
1.2	0.016	0.072	0.156	0.251	0.345	0.430	0.503	0.563	0.611	0.650	0.350
1.4	0.037	0.125	0.230	0.334	0.427	0.507	0.572	0.624	0.665	0.699	0.301
1.6	0.068	0.185	0.304	0.410	0.499	0.572	0.629	0.675	0.711	0.740	0.260
1.8	0.108	0.249	0.375	0.478	0.561	0.627	0.678	0.717	0.749	0.774	0.226
2.0	0.154	0.312	0.439	0.539	0.615	0.673	0.718	0.753	0.781	0.802	0.198
2.2	0.203	0.372	0.498	0.591	0.661	0.713	0.753	0.783	0.808	0.827	0.173
2.4	0.254	0.429	0.551	0.637	0.700	0.747	0.782	0.809	0.830	0.847	0.153
2.6	0.304	0.481	0.597	0.677	0.734	0.776	0.807	0.831	0.850	0.865	0.135
2.8	0.353	0.529	0.639	0.712	0.764	0.801	0.829	0.850	0.867	0.880	0.120
3.0	0.401	0.573	0.675	0.743	0.789	0.823	0.848	0.867	0.882	0.893	0.107

We now plot both the cumulative distribution and the density functions of the likelihood ratio. The cumulative distribution is obtained by simply plotting π_k against k while the density is obtained by direct differentiation of the function obtained for π_k. This means that where $\pi_k = 0$ the density is identically zero. Elsewhere we proceed as follows, firstly considering $0 \leq k \leq 1$. We require the derivative of π_k with respect to k, denoting this by π' and applying the chain rule we find

$$\pi' = -\frac{2}{k}\left\{-\frac{1}{a}g\left(\frac{\eta}{a}\right) + \frac{\exp\{\frac{\eta}{2}\}}{a(a+1)^{\frac{n-4}{2}}}g\left(\eta+\frac{\eta}{a}\right) - \frac{\exp\{\frac{\eta}{2}\}}{2(a+1)^{\frac{n-2}{2}}}\left(1-G\left(\eta+\frac{\eta}{a}\right)\right)\right\}.$$

Substituting our expressions for a and η we obtain

$$\begin{aligned}\pi' &= \frac{n-2}{k\tilde{F}}\left[g\left(-\frac{(n-2)\log k}{\tilde{F}}\right) + \frac{(n-2)^{\frac{n-4}{2}}}{k(n-2+2\tilde{F})^{\frac{n-4}{2}}}g\left(-2\log k - \frac{(n-2)\log k}{\tilde{F}}\right)\right] \\ &\quad + \frac{(n-2)^{\frac{n-2}{2}}}{k^2(n-2+2\tilde{F})^{\frac{n-2}{2}}}\left[1-G\left(-2\log k - \frac{(n-2)\log k}{\tilde{F}}\right)\right].\end{aligned}$$

Table 4.2: π_k & P for $n = 50$

	k										
$\tilde{F}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	P
0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
0.2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.090	0.181	0.819
0.4	0.000	0.000	0.000	0.000	0.000	0.004	0.054	0.159	0.253	0.327	0.673
0.6	0.000	0.000	0.000	0.001	0.016	0.094	0.211	0.309	0.386	0.447	0.553
0.8	0.000	0.000	0.001	0.023	0.114	0.242	0.350	0.431	0.494	0.545	0.455
1.0	0.000	0.000	0.018	0.110	0.253	0.374	0.464	0.531	0.583	0.625	0.375
1.2	0.000	0.007	0.081	0.237	0.380	0.483	0.557	0.612	0.655	0.690	0.310
1.4	0.000	0.036	0.187	0.361	0.487	0.573	0.634	0.679	0.715	0.744	0.256
1.6	0.003	0.102	0.306	0.469	0.575	0.646	0.696	0.734	0.764	0.788	0.212
1.8	0.016	0.196	0.417	0.559	0.647	0.706	0.748	0.780	0.804	0.824	0.176
2.0	0.048	0.302	0.513	0.634	0.707	0.756	0.791	0.817	0.837	0.854	0.146
2.2	0.103	0.406	0.594	0.695	0.756	0.797	0.826	0.848	0.865	0.878	0.122
2.4	0.178	0.499	0.662	0.746	0.797	0.831	0.855	0.873	0.887	0.898	0.102
2.6	0.265	0.579	0.718	0.788	0.831	0.859	0.879	0.894	0.906	0.915	0.085
2.8	0.356	0.647	0.764	0.823	0.858	0.882	0.899	0.912	0.921	0.929	0.071
3.0	0.443	0.704	0.803	0.852	0.882	0.901	0.915	0.926	0.934	0.941	0.059

For $1 \leq k < \infty$ we find simply that

$$\pi' = \frac{(n-2)^{\frac{n-2}{2}}}{k^2(n-2+2\tilde{F})^{\frac{n-2}{2}}}.$$

We show these plots of the distribution and density functions in Figures 4.1 to 4.3 for our selected values of n with $\tilde{F} = 1, 2, 3$.

For the cases graphed we also calculate the maximised likelihood ratio (MLR). This is obtained using the maximum likelihood estimate $\tilde{\sigma}$ of σ) under H_0 in addition to the maximum likelihood estimates, $(\hat{\mu}, \hat{\sigma}, \hat{\phi})$ of (μ, σ, ϕ) under H_1. Note that the form of the likelihood is independent of μ under H_0 and that, in fact, under H_1 we only require the value of RSS in addition to $\hat{\sigma}$. This is because $\hat{\mu}, \hat{\phi}$ enter the maximised likelihood only through the residual sum of squares which we express here as

$$RSS = \sum_{i=1}^{n} \left(x_i - \hat{\phi} x_{i-1} - (1-\hat{\phi})\hat{\mu} \right)^2.$$

Figure 4.1: Distribution (left) & density (right) functions for $n = 10, \tilde{F} = 1, 2, 3$

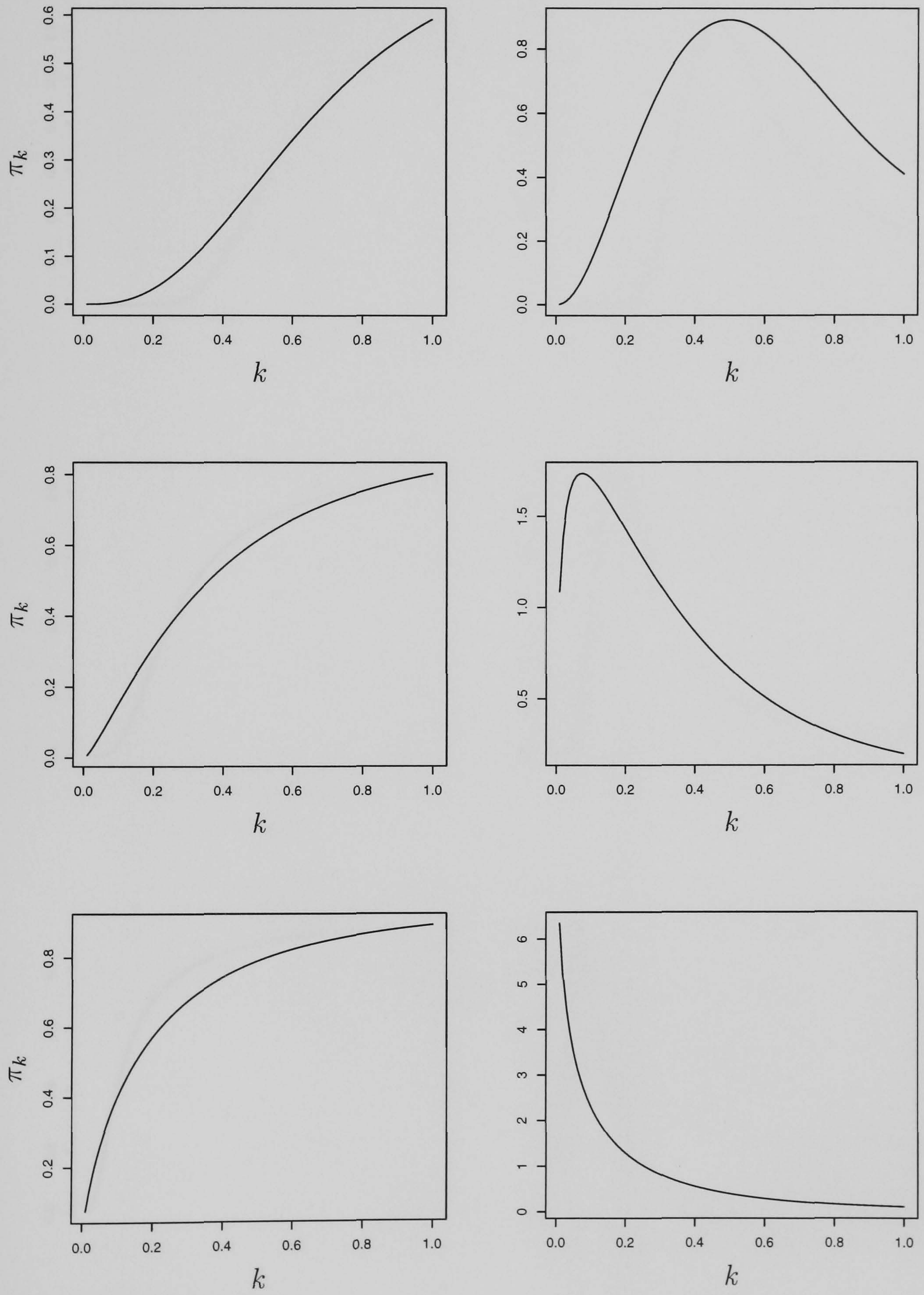


Figure 4.2: Distribution (left) & density (right) functions for $n = 50, \tilde{F} = 1, 2, 3$

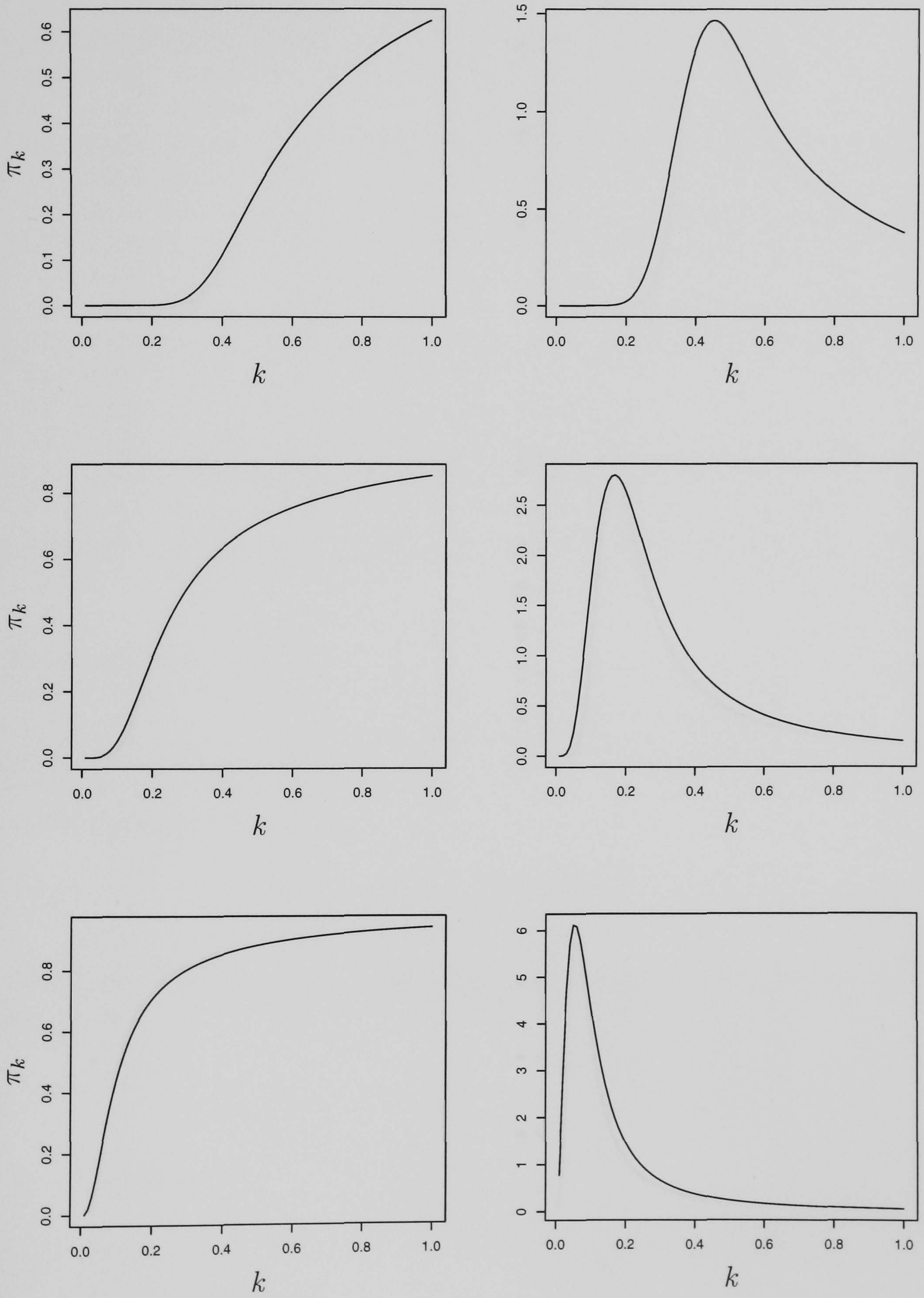
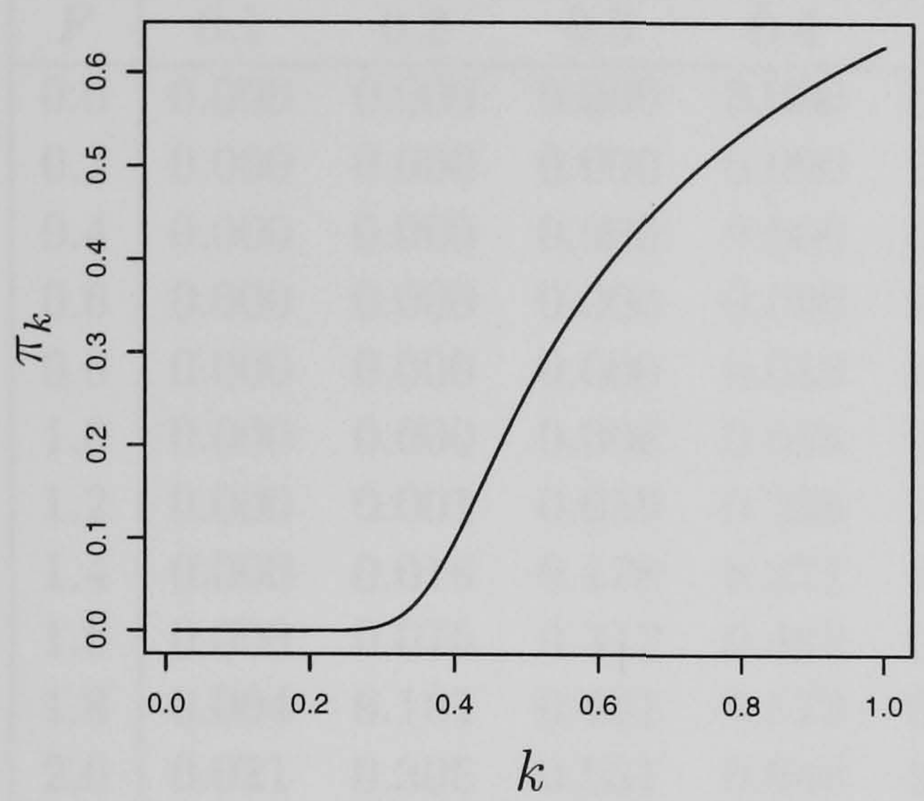
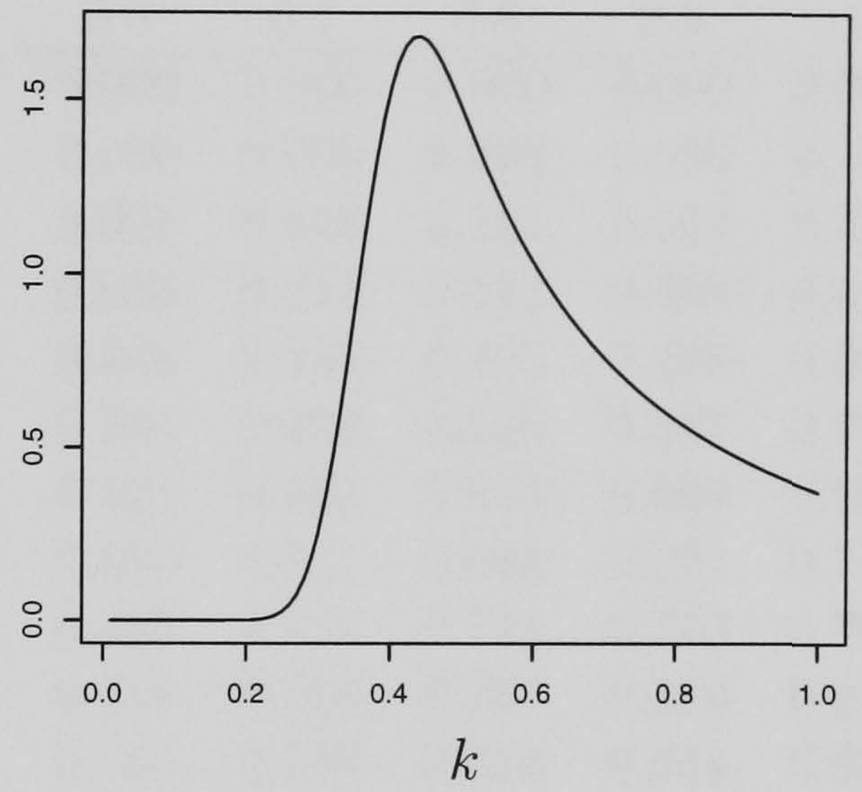
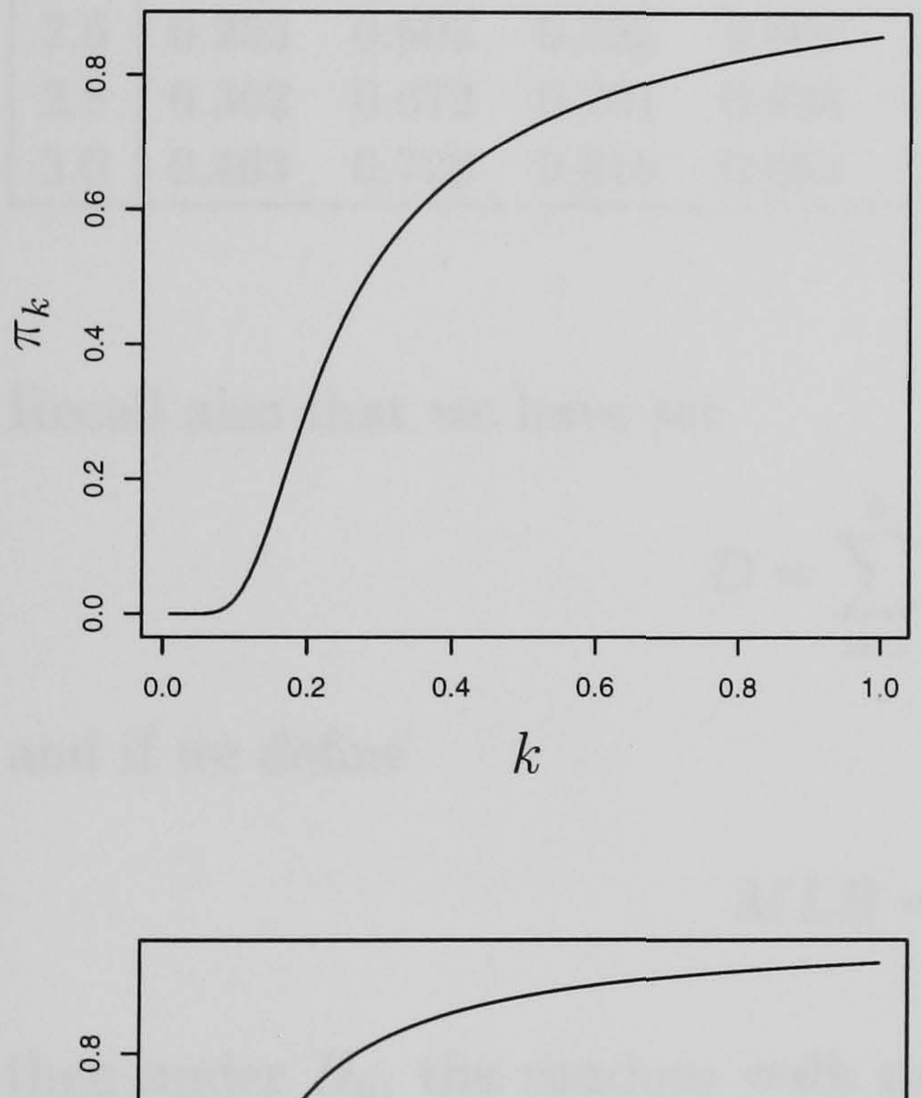
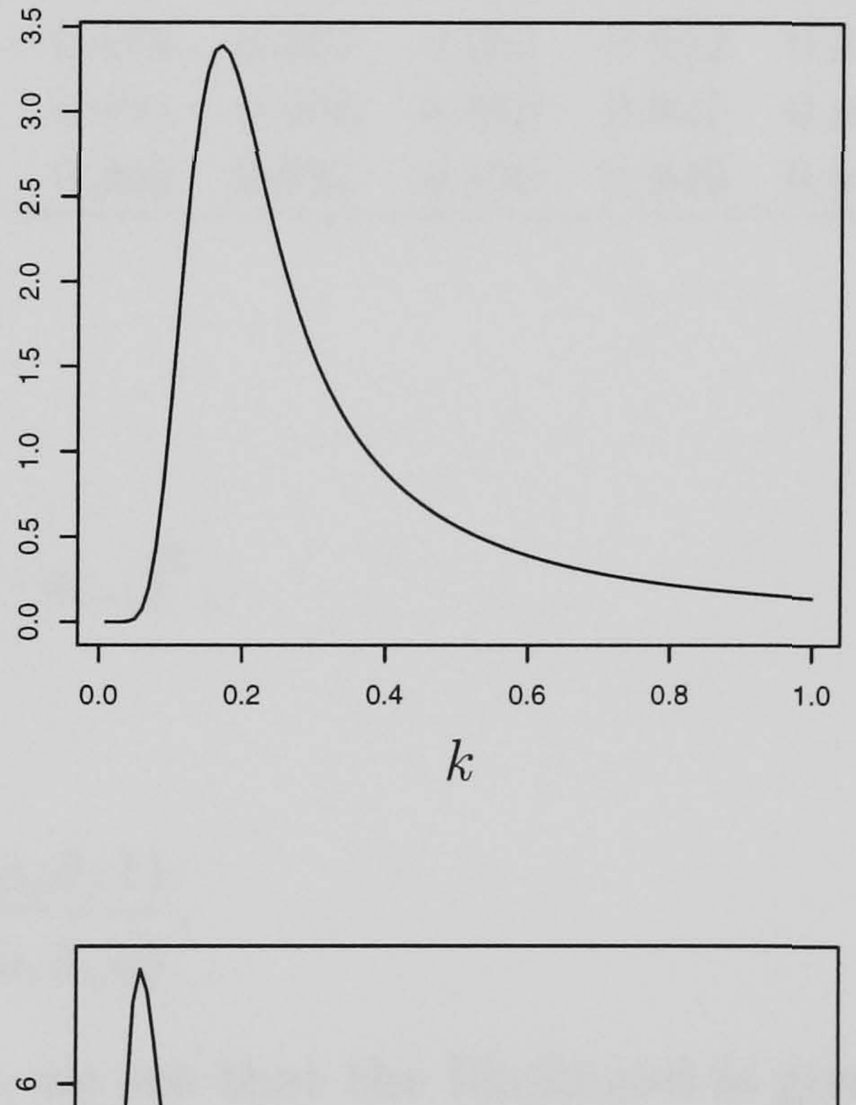


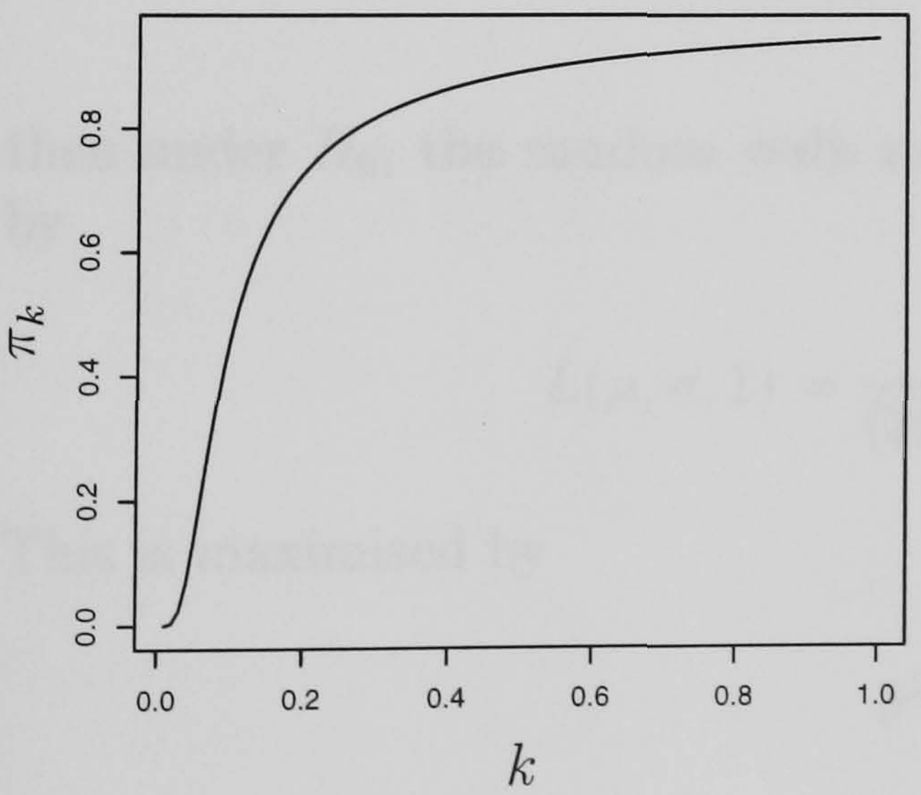
Figure 4.3: Distribution (left) & density (right) functions for $n = 100, \tilde{F} = 1, 2, 3$











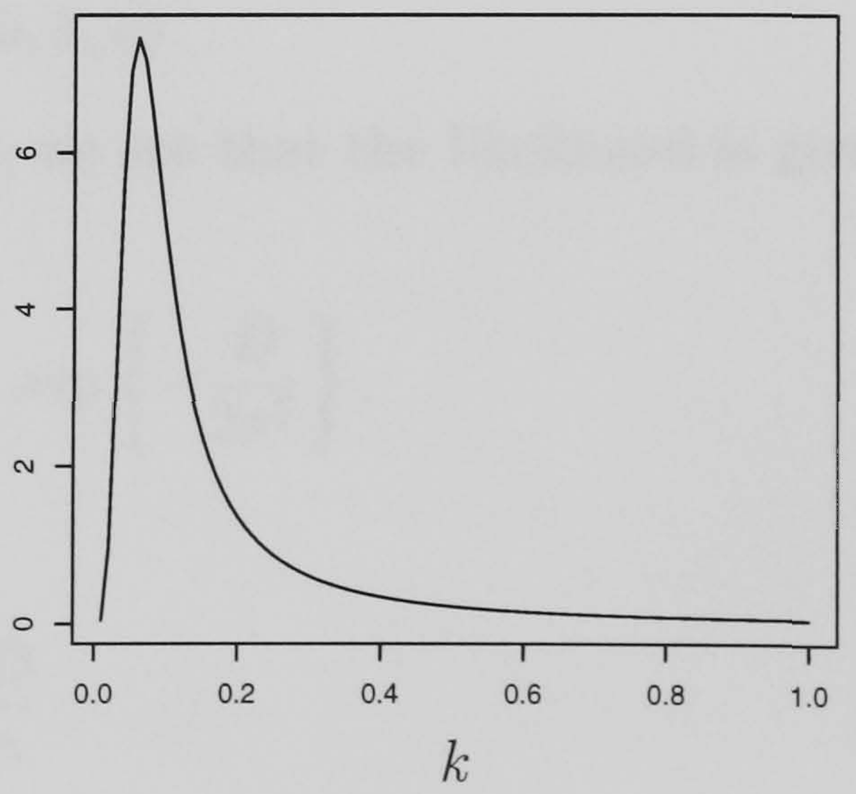


Table 4.3: π_k & P for $n = 100$

	k										
$\tilde{F}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	P
0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
0.2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.090	0.181	0.819
0.4	0.000	0.000	0.000	0.000	0.000	0.001	0.048	0.161	0.254	0.329	0.671
0.6	0.000	0.000	0.000	0.000	0.007	0.088	0.213	0.311	0.388	0.449	0.551
0.8	0.000	0.000	0.000	0.010	0.105	0.246	0.354	0.435	0.498	0.548	0.452
1.0	0.000	0.000	0.006	0.096	0.257	0.381	0.469	0.535	0.587	0.628	0.372
1.2	0.000	0.001	0.059	0.238	0.389	0.491	0.563	0.618	0.660	0.694	0.306
1.4	0.000	0.016	0.178	0.371	0.497	0.581	0.641	0.686	0.721	0.749	0.251
1.6	0.000	0.075	0.312	0.482	0.586	0.655	0.704	0.741	0.770	0.793	0.207
1.8	0.004	0.181	0.431	0.573	0.659	0.715	0.756	0.787	0.810	0.829	0.171
2.0	0.021	0.305	0.531	0.648	0.718	0.765	0.799	0.824	0.844	0.859	0.141
2.2	0.069	0.421	0.613	0.709	0.768	0.806	0.834	0.855	0.871	0.884	0.116
2.4	0.150	0.520	0.680	0.760	0.808	0.840	0.863	0.880	0.893	0.904	0.096
2.6	0.253	0.603	0.735	0.802	0.841	0.868	0.887	0.901	0.912	0.921	0.079
2.8	0.362	0.672	0.781	0.836	0.869	0.891	0.906	0.918	0.927	0.934	0.066
3.0	0.463	0.728	0.819	0.864	0.891	0.909	0.922	0.932	0.940	0.946	0.054

Recall also that we have set

$$D = \sum_{i=1}^{n} (x_i - x_{i-1})^2 ,$$

and if we define

$$MLR = \frac{L(\mu, \tilde{\sigma}, 1)}{L(\hat{\mu}, \hat{\sigma}, \hat{\phi})},$$

then under H_0, the random walk model, we see that the likelihood is given by

$$L(\mu, \sigma, 1) = \frac{1}{(2\pi\sigma)^n} \exp\left\{-\frac{D}{2\sigma^2}\right\}.$$

This is maximised by

$$\tilde{\sigma}^2 = \frac{D}{n}$$

and the corresponding maximised likelihood is

$$L(\mu, \tilde{\sigma}, 1) = \frac{n^{n/2}}{(2\pi)^n D^{n/2}} \exp\left\{-\frac{n}{2}\right\}.$$

We now evaluate the likelihood under H_1 at $\hat{\mu}$ and $\hat{\phi}$ for general σ to give

$$L(\hat{\mu}, \sigma, \hat{\phi}) = \frac{1}{(2\pi\sigma)^n} \exp\left\{-\frac{RSS}{2\sigma^2}\right\}.$$

This is at a maximum when

$$\sigma^2 = \hat{\sigma}^2 = \frac{RSS}{n}$$

and we obtain the maximised likelihood,

$$L(\hat{\mu}, \hat{\sigma}, \hat{\phi}) = \frac{n^{n/2}}{(2\pi)^n RSS^{n/2}} \exp\left\{-\frac{n}{2}\right\}.$$

We now see that the MLR (in favour of the null model) is given by

$$MLR = \left(\frac{RSS}{D}\right)^{n/2}.$$

If we recall that

$$\tilde{F} = \frac{(D - RSS)/2}{RSS/(n-2)},$$

then we can express the MLR as

$$MLR = \left(\frac{n-2}{2\tilde{F} + n - 2}\right)^{n/2}.$$

Table 4.4 shows the values of this statistic graphed in Figures 4.1 to 4.3. We see that these values broadly agree with the peaks of the density plots in the figures meaning that we have consistent conclusions from both the π_k method and the MLR.

We see from the density plots that as we move away from $\tilde{F} = 0$ (essentially moving away from the null model) the density becomes more concentrated towards $k = 0$. As we would hope, this shows stronger preference for the alternative model.

Further examination of the graphs, as in our work relating to the t-test,

Table 4.4: Maximised likelihood ratio values

	$\tilde{F}$		
n	1	2	3
10	0.328	0.132	0.061
50	0.360	0.135	0.053
100	0.364	0.135	0.051

suggests that we examine the effect that $\tilde{F}$ and n have on the mean, variance and skewness of the likelihood ratio. We now proceed as in our earlier work, recalling that $Y_1 \sim \chi^2_2$ and $Y_2 \sim \chi^2_{n-2}$. Now

$$\begin{aligned} \pi_k &= Pr\left(\frac{2\tilde{F}}{n-2}Y_2 - Y_1 > \eta\right), \\ &= Pr\left(\frac{1}{2}\left[Y_1 - \frac{2\tilde{F}}{n-2}Y_2\right] < \log k\right), \\ &= Pr\left(\exp\left\{\frac{1}{2}\left[Y_1 - \frac{2\tilde{F}}{n-2}Y_2\right]\right\} < k\right), \end{aligned}$$

and if we now let

$$LR = \exp\left\{\frac{1}{2}Y_1 - \frac{\tilde{F}}{n-2}Y_2\right\},$$

then we simply wish to evaluate $E(LR), Var(LR)$ and $Skew(LR)$. As a result of the independence of Y_1 and Y_2 we know that

$$\begin{aligned} E\left(LR\right) &= E\left(\exp\left\{\frac{1}{2}Y_1\right\}\exp\left\{\frac{-\tilde{F}}{n-2}Y_2\right\}\right), \\ &= E\left(LR_1 LR_2\right) = E\left(LR_1\right)E\left(LR_2\right) \end{aligned}$$

where

$$\begin{aligned} LR_1 &= \exp\left\{\frac{1}{2}Y_1\right\}, \\ LR_2 &= \exp\left\{\frac{-\tilde{F}}{n-2}Y_2\right\}. \end{aligned}$$

We know from our earlier work that $E(LR_1)$ is undefined which, once more, means that $E(LR), Var(LR)$ and $Skew(LR)$ are all undefined. As before, though, we are able to obtain the summary statistics for $M \equiv \log LR$. We know that

$$M = \frac{1}{2}Y_1 - \frac{\tilde{F}}{n-2}Y_2,$$

and it is then straightforward to obtain

$$E(M) = 1 - \tilde{F}$$

and

$$Var(M) = 1 + \frac{2\tilde{F}^2}{(n-2)^2}.$$

Now recall

$$a = \frac{2\tilde{F}}{n-2},$$

so that to calculate the skewness we proceed as follows

$$\begin{aligned} E(M^3) &= \frac{1}{8}E(Y_1^3) - \frac{3a}{8}E(Y_1^2)E(Y_2) + \frac{3a^2}{8}E(Y_1)E(Y_2^2) - \frac{a^3}{8}E(Y_2^3), \\ &= 6 - 3a(n-2) + \frac{3a^2}{4}n(n-2) - \frac{a^3}{8}n(n-2)(n+2). \end{aligned}$$

Again using our earlier work we know that

$$Skew(M) = \frac{E(M^3) - 3E(M^2)E(M) + 2E(M)^3}{Var(M)^{3/2}},$$

and using

$$E(M^2) = Var(M) + E(M)^2$$

we can see that

$$\begin{aligned} Skew(M) &= \frac{E(M^3) - 3Var(M)E(M) - E(M)^3}{Var(M)^{3/2}}, \\ &= \frac{2 - a^3(n-2)}{(1 + \frac{a^2}{2}(n-2))^{3/2}}. \end{aligned}$$

If we now consider how these summary statistics behave as we let our sample size n tend to infinity we observe that $E(M) = 1 - \tilde{F}$ is independent of sample size and that

$$Var(M) = 1 + \frac{2\tilde{F}^2}{n-2} \rightarrow 1 + 0 = 1.$$

It is notable that the limit of the variance is, as with our earlier work, independent of the observed data. If we now consider the skewness and use our result for the limit of the variance we can see that

$$Skew(M) = \frac{2 - \frac{8\tilde{F}^3}{(n-2)^2}}{\left(1 + \frac{2\tilde{F}^2}{(n-2)}\right)^{3/2}} \rightarrow \frac{2-0}{1-0} = 2.$$

Once more we have a result that is independent of the data observed.

Chapter 5

The choice between Poisson and geometric distributions

5.1 Introduction

In the famous example of Cox (1962), a sample $y_1, y_2, \ldots, y_n$ is observed from a discrete distribution which is either a Poisson (H_1) or a geometric (H_2) distribution. We shall use our likelihood method to investigate how to decide which of these two hypotheses is better-supported by the data. We parametrise both distributions to have mean μ. Under H_1 we have:

$$Pr(Y = y) = e^{-\mu}\mu^y/y!, \quad y = 0, 1, \ldots .$$

While under H_2 we have:

$$Pr(Y = y) = \mu^y/(1+\mu)^{y+1}, \quad y = 0, 1, \ldots .$$

5.2 Likelihood ratio

Let the model implied by hypothesis H_i be denoted by M_i for $i = 1, 2$. We shall embed both of these models in the negative binomial family, with probability function:

$$Pr(Y = y) = \binom{y+r-1}{r-1} r^r \mu^y/(r+\mu)^{y+r}, \quad \mu, r > 0, \quad y = 0, 1, \ldots .$$

When $r = 1$ this simplifies to the geometric form given above for M_2 and when $r \to \infty$ we obtain the form for the Poisson model (M_1). We shall denote $\sum_1^n y_i$ by y_+ and we then see that the likelihood from the negative

binomial model is

$$\begin{aligned} L(\mu, r) &= \prod_{i=1}^{n} \binom{y_i + r - 1}{r - 1} r^r \mu^{y_i} / (r + \mu)^{y_i + r}, \\ &= \frac{r^{nr} \mu^{y_+}}{(r + \mu)^{y_+ + nr}} \prod_{i=1}^{n} \binom{y_i + r - 1}{r - 1}. \end{aligned}$$

The likelihood ratio of M_1 relative to M_2 is

$$LR = \frac{L(\mu, \infty)}{L(\mu, 1)} = \frac{e^{-n\mu} \mu^{y_+} / \prod_1^n y_i!}{\mu^{y_+} / (1 + \mu)^{y_+ + n}} = \frac{e^{-n\mu} (1 + \mu)^{y_+ + n}}{\prod_1^n y_i!}.$$

Then

$$\begin{aligned} \pi_k &= Pr(LR < k) \\ &= Pr\left(e^{-n\mu} (1 + \mu)^{y_+ + n} < k \prod_1^n y_i! \right) \\ &= Pr\left(\left\{ \frac{y_+}{n} + 1 \right\} \log(1 + \mu) - \mu < \frac{1}{n} \left\{ \log k + \sum_1^n \log y_i! \right\} \right) \\ &= Pr\left(\left\{ \mu + \frac{1}{n} \left(\log k - y_+ \log(1 + \mu) + \sum_1^n \log y_i! \right) \right\} / \log(1 + \mu) > 1 \right), \end{aligned}$$

which is an integral over a region in the (μ, r) plane of the joint posterior distribution of (μ, r) derived from the likelihood.

5.3 Evaluation of π_k

In order to employ our simulation method to evaluate π_k it is clear from the above form that we must obtain a realisation of μ from the joint posterior and then test whether

$$g(\mu) = \left\{ \mu + \frac{1}{n} \left(\log k - y_+ \log(1 + \mu) + \sum_1^n \log y_i! \right) \right\} / \log(1 + \mu) > 1.$$

Unfortunately, due to the fact that there is no conjugate prior that we can place on r (the index of the negative binomial distribution), we are unable to obtain a straightforward joint posterior distribution of (μ, r) from which to simulate μ. We therefore require a different method of simulation, and

consider the following reparameterisation of the negative binomial, using $p = \mu/(\mu + r)$.

$$\begin{aligned} L(p,r) &= p^{y_+}(1-p)^{nr}\prod_{i=1}^{n}\binom{y_i+r-1}{r-1}, \\ &= \frac{p^{y_+}(1-p)^{nr}}{B\left(y_+ + 1, nr+1\right)} B\left(y_+ + 1, nr + 1\right)\prod_{i=1}^{n}\binom{y_i+r-1}{r-1}. \end{aligned}$$

where $B(a,b)$ is the beta function

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

If we now use a flat prior for p, we see that

$$p|r, y \sim \beta\left(y_+ + 1, nr + 1\right),$$

so that, given r, we are able to simulate from p (and hence from μ). This flat prior for p means that the prior for μ is given by

$$\frac{r}{(r+\mu)^2}.$$

Denoting $\pi(r)$ as the chosen (marginal) prior distribution for r, we see that the posterior for r can be written as

$$\pi(r|y) \propto B\left(y_+ + 1, nr + 1\right)\prod_{i=1}^{n}\binom{y_i+r-1}{r-1}\pi(r).$$

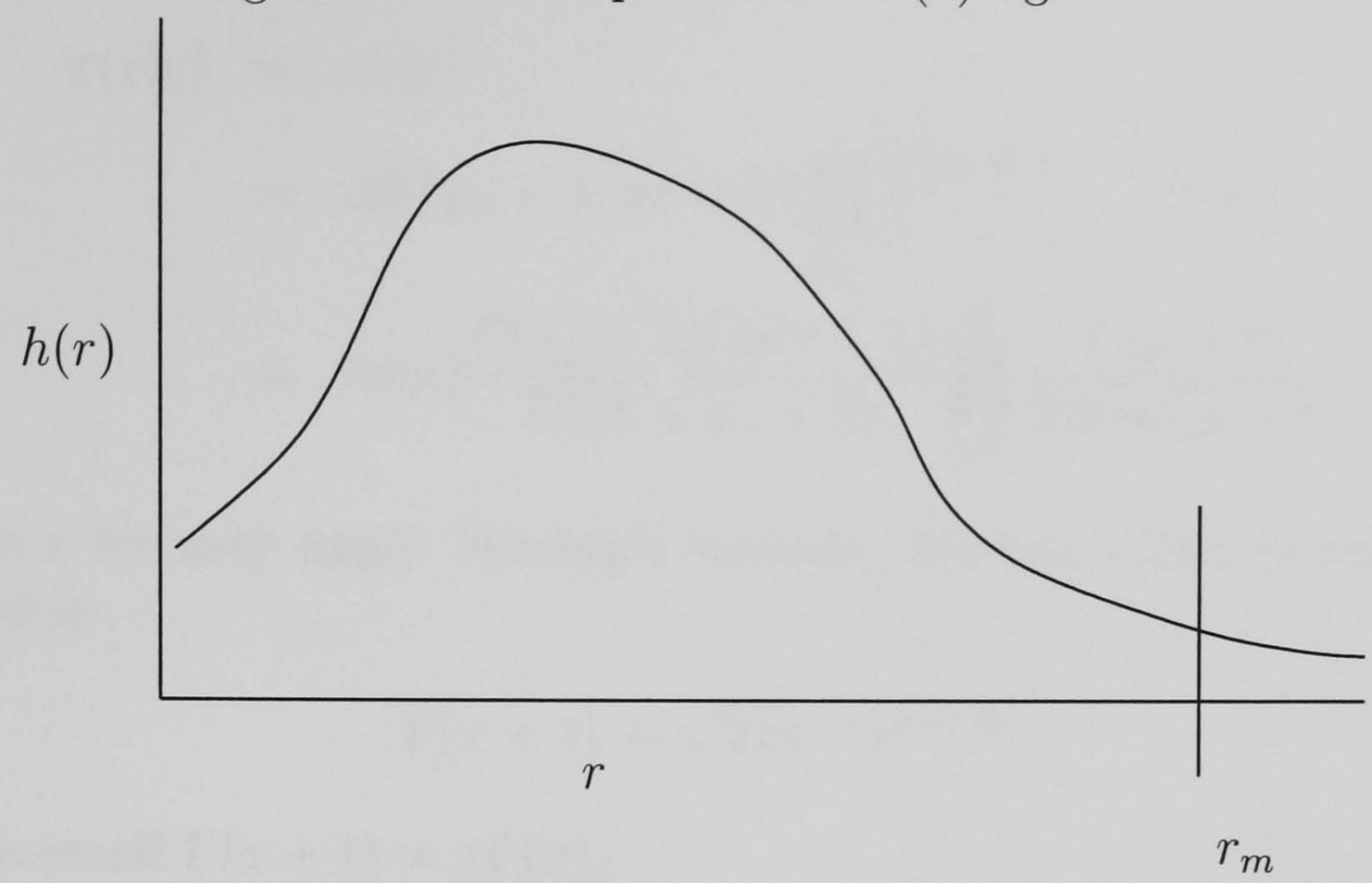
If we now define $h(r)$ so that $\pi(r|y) \propto h(r)$ we have

$$h(r) = B\left(y_+ + 1, nr + 1\right)\prod_{i=1}^{n}\binom{y_i+r-1}{r-1}\pi(r).$$

For our given data, $y_1, \ldots, y_n$, we now plot $h(r)$ against r and note that as r increases there appears to be a value after which $h(r)$ becomes negligible. We take r_m to be this value of r; this is evaluated by recording the value at which $h(r)$ drops below a selected tolerance level. This is shown in Figure 5.1.

Our next step is to evaluate $h(r^{(i)})$ over a fine grid of $r^{(i)}$ covering $(0, r_m]$. We can then obtain the empirical cumulative distribution function $\tilde{F}(r)$

Figure 5.1: Example Plot of $h(r)$ against r



which is

$$\tilde{F}(r) = \frac{\sum_{r^{(i)} \leq r} h(r^{(i)})}{\sum_{r^{(i)}} h(r^{(i)})}.$$

We now find ourselves in a position to simulate a value of μ in the following manner. Firstly we simulate u from $U(0,1)$ then using interpolation we obtain the value of r which has this value of the (empirical) cumulative distribution function. Using this value of r we simulate a value of p from the Beta distribution given above. Finally we obtain μ using

$$\mu = \frac{rp}{1-p}.$$

We can now proceed in our familar fashion and obtain an estimate for π_k (given data and k) using the proportion of times in a large sample that $g(\mu) > 1$.

We now demonstrate that the posterior distribution for r obtained above is in fact integrable, and hence can simply be transformed into a proper posterior. Recall that we defined the function $h(r)$ to be proportional to the

posterior density function, then, denoting by c a value independent of r,

$$\begin{aligned}
\pi(r|y) &= ch(r) \\
&= cB\left(y_+ + 1, nr + 1\right) \prod_{i=1}^{n} \binom{y_i + r - 1}{r - 1} \pi(r) \\
&= c\pi(r) \frac{\Gamma(y_+ + 1)\Gamma(nr + 1)}{\Gamma(nr + y_+ + 2)} \prod_{i=1}^{n} \frac{\Gamma(y_i + r)}{\Gamma(r)\Gamma(y_i + 1)}.
\end{aligned}$$

For large r we may apply Stirling's formula (Stirling 1730) to the function $\Gamma(\cdot)$, that is

$$\Gamma(x + 1) \sim \sqrt{2\pi} e^{-x} x^{x+1/2}.$$

Or, if we recall $\Gamma(x + 1) = x\Gamma(x)$,

$$\Gamma(x) \sim \sqrt{2\pi} e^{-x} x^{x-1/2}.$$

We find

$$\begin{aligned}
\pi(r|y) &\sim c\pi(r) \frac{\sqrt{2\pi} e^{-nr} \left(nr\right)^{nr+1/2}}{\sqrt{2\pi} e^{-(nr+y_++1)} \left(nr + y_+ + 1\right)^{nr+y_++3/2}} \\
&\quad \times \prod_{i=1}^{n} \left[\frac{\sqrt{2\pi} e^{-(y_i+r)} \left(y_i + r\right)^{y_i+r-1/2}}{\sqrt{2\pi} e^{-r} r^{r-1/2}} \right] \\
&\sim c\pi(r) \frac{\left(nr\right)^{nr+1/2}}{\left(nr + y_+ + 1\right)^{nr+y_++3/2}} \prod_{i=1}^{n} \left[\frac{\left(y_i + r\right)^{y_i+r-1/2}}{r^{r-1/2}} \right] \\
&\sim c\pi(r) \frac{1}{\left(nr\right)^{y_++1} \left(1 + \frac{y_++1}{nr}\right)^{nr+y_++3/2}} \prod_{i=1}^{n} \left[r^{y_i} \left(1 + \frac{y_i}{r}\right)^{y_i+r-1/2} \right].
\end{aligned}$$

As r is large we approximate $(1 + u/n)^n$ by e^u, giving

$$\pi(r|y) \sim c\pi(r) \frac{1}{r^{y_++1} e^{y_++1}} \prod_{i=1}^{n} \left[r^{y_i} e^{y_i} \right].$$

We observe that this is asymptotically

$$\begin{aligned}
\pi(r|y) &\sim c\pi(r) \frac{r^{y_+}}{r^{y_++1}} \\
&\sim c\pi(r) \frac{1}{r}.
\end{aligned}$$

So if $\pi(r)$ is constant, $\pi(r|y)$ will not be integrable because $1/r$ falls off too slowly. We note

$$\int_1^\infty \frac{1}{r}\,\mathrm{d}r \to [\log r]_1^\infty.$$

If $\pi(r) \propto 1/r$ then $\pi(r|y)$ will be integrable. However, by truncating the range of r at r_m the problem of integrability disappears as only finite ranges are involved.

5.4 Bayes factor and maximised likelihood ratio

For comparison purposes we shall also calculate the Bayes factor and the maximised likelihood ratio. This proceeds in the following manner.

5.4.1 Bayes factor calculation

We define the Bayes factor (in favour of M_1, the Poisson model, over M_2) to be

$$B = \frac{\int_0^\infty L(\mu,\infty)\pi(\mu)\,\mathrm{d}\mu}{\int_0^\infty L(\mu,1)\pi(\mu)\,\mathrm{d}\mu},$$

where the prior distribution for μ is taken to be diffuse and is given by

$$\pi(\mu) = k\mu^{-s}$$

for either $s = 0$ or $s = 1$. Note that we may ignore the constant of proportionality k as it will cancel when we calculate the ratio of integrals. We consider two prior distributions here for comparison purposes, the prior with $s = 1$ placing more weight on smaller values of μ.

We first consider the term in the numerator of the Bayes factor:

$$\begin{aligned}
\bar{L}^P &= \int_0^\infty L(\mu,\infty)\pi(\mu)\,\mathrm{d}\mu \\
&= \int_0^\infty \frac{e^{-n\mu}\mu^{y_+-s}}{\prod_1^n y_i!}\,\mathrm{d}\mu \\
&= \frac{1}{\prod_1^n y_i!}\int_0^\infty e^{-n\mu}\mu^{y_+-s}\,\mathrm{d}\mu.
\end{aligned}$$

If we now make the substitution $u = n\mu$ we obtain

$$\begin{aligned}
\bar{L}^P &= \frac{1}{\prod_1^n y_i!} \int_0^\infty e^{-u} \left(\frac{u}{n}\right)^{y_+ - s} \frac{1}{n} \, \mathrm{d}u \\
&= \frac{1}{n^{y_+ - s + 1} \prod_1^n y_i!} \int_0^\infty e^{-u} u^{y_+ - s} \, \mathrm{d}u \\
&= \frac{\Gamma(y_+ - s + 1)}{n^{y_+ - s + 1} \prod_1^n y_i!}.
\end{aligned}$$

We now perform a similar calculation for the geometric term in the denominator of the Bayes factor:

$$\begin{aligned}
\bar{L}^G &= \int_0^\infty L(\mu, 1)\pi(\mu) \, \mathrm{d}\mu \\
&= \int_0^\infty \mu^{y_+ - s}(1+\mu)^{-y_+ - n} \, \mathrm{d}\mu.
\end{aligned}$$

Using $u = 1/(\mu + 1)$ we obtain

$$\begin{aligned}
\bar{L}^G &= -\int_1^0 \left(\frac{1-u}{u}\right)^{y_+ - s} \left(\frac{1}{u}\right)^{-y_+ - n} \frac{1}{u^2} \, \mathrm{d}u \\
&= \int_0^1 (1-u)^{y_+ - s} \frac{u^{y_+ + n}}{u^{y_+ - s + 2}} \, \mathrm{d}u \\
&= \int_0^1 (1-u)^{y_+ - s} u^{n+s-2} \, \mathrm{d}u \\
&= \frac{\Gamma(n+s-1)\Gamma(y_+ - s + 1)}{\Gamma(n + y_+)}.
\end{aligned}$$

We can now see that the Bayes factor is given by

$$B = \frac{\bar{L}^P}{\bar{L}^G} = \frac{\Gamma(n + y_+)}{\Gamma(n+s-1) n^{y_+ - s + 1} \prod_1^n y_i!}.$$

5.4.2 Maximised likelihood ratio calculation

We define the maximised likelihood ratio to be

$$MLR = \frac{L(\hat{\mu}^P, \infty)}{L(\hat{\mu}^G, 1)},$$

where $\hat{\mu}^P$ is the maximum likelihood estimator (mle) of μ under the Poisson model and $\hat{\mu}^G$ is the mle of μ under the geometric model. By differentiating the likelihood functions with respect to μ and setting the resultant forms equal to zero we find that

$$\hat{\mu}^P = \hat{\mu}^G = \bar{y} = \frac{y_+}{n}.$$

We therefore find that the maximised likelihood ratio is given by

$$MLR = \frac{L(y_+/n, \infty)}{L(y_+/n, 1)} = \frac{e^{-y_+}(1 + y_+/n)^{y_+ + n}}{\prod_1^n y_i!}.$$

5.5 Examples

In order to demonstrate the use of the above method we consider applying it to randomly generated data. We take samples both of size 10 and 100 from Poisson, geometric and negative binomial (taking $r = 5$) distributions with the mean for each distribution taken to be the same. Means of $0.8, 0.9$ and 1 are considered. We use the method described above to obtain estimates of π_k considering the following prior distributions for r (note that we may ignore normalisation constants as they are absorbed into the constant κ mentioned earlier which plays no part in this method).

$$\begin{aligned} \pi_1(r) &= \frac{1}{r} \\ \pi_2(r) &= 1. \end{aligned}$$

In order to estimate π_k we set our tolerance level for $h(r)$ to be $0.01 \times \max h(r)$ and we use a simulation sample size of 100. This means that the apparent variance of our estimates is at most 0.0025, however, in this example there is additional potential error due to the use of an approximation to the posterior distribution from which we are sampling. We present the results in the following tables (Tables 5.1 to 5.6) with each table consisting of our estimates of π_k using all distributions and priors for a fixed mean and sample size. We should recall that large values of π_k provide evidence against the Poisson hypothesis. As this example is non-nested there is no classical P-value to display here. In Table 5.7 we also display the Bayes factors (as described previously) and the maximised likelihood ratio for each sample.

Table 5.1: π_k with mean 0.8, sample size 10

	k									
Sample	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
P, prior 1	.02	.02	.03	.03	.04	.05	.05	.05	.05	.05
P, prior 2	0	.02	.02	.02	.03	.03	.03	.04	.04	.05
NB(5), prior 1	0	0	0	0	0	0	0	0	0	.01
NB(5), prior 2	0	.02	.02	.02	.02	.02	.02	.02	.03	.03
G, prior 1	.05	.07	.08	.10	.14	.26	.50	1	1	1
G, prior 2	.02	.04	.04	.07	.14	.25	.50	1	1	1

Table 5.2: π_k with mean 0.8, sample size 100

	k									
Sample	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
P, prior 1	0	0	0	0	0	0	0	0	0	0
P, prior 2	0	0	0	0	0	0	0	0	0	0
NB(5), prior 1	0	0	0	0	0	0	0	0	0	0
NB(5), prior 2	0	0	0	0	0	0	0	0	0	0
G, prior 1	1	1	1	1	1	1	1	1	1	1
G, prior 2	1	1	1	1	1	1	1	1	1	1

Table 5.3: π_k with mean 0.9, sample size 10

	k									
Sample	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
P, prior 1	.02	.02	.03	.03	.04	.05	.05	.05	.05	.05
P, prior 2	0	.02	.02	.02	.03	.03	.03	.04	.04	.05
NB(5), prior 1	0	0	0	0	0	0	0	0	0	.01
NB(5), prior 2	0	.02	.02	.02	.02	.02	.02	.02	.03	.03
G, prior 1	.05	.07	.08	.10	.14	.26	.50	1	1	1
G, prior 2	.02	.04	.04	.07	.14	.25	.50	1	1	1

Table 5.4: π_k with mean 0.9, sample size 100

Sample	k .1	.2	.3	.4	.5	.6	.7	.8	.9	1
P, prior 1	0	0	0	0	0	0	0	0	0	0
P, prior 2	0	0	0	0	0	0	0	0	0	0
NB(5), prior 1	1	1	1	1	1	1	1	1	1	1
NB(5), prior 2	1	1	1	1	1	1	1	1	1	1
G, prior 1	1	1	1	1	1	1	1	1	1	1
G, prior 2	1	1	1	1	1	1	1	1	1	1

Table 5.5: π_k with mean 1, sample size 10

Sample	k .1	.2	.3	.4	.5	.6	.7	.8	.9	1
P, prior 1	0	0	.01	.02	.02	.03	.03	.03	.03	.04
P, prior 2	0	.01	.01	.01	.01	.01	.01	.02	.02	.03
NB(5), prior 1	.03	.08	.09	.14	.23	.44	.67	1	1	1
NB(5),prior 2	.02	.05	.05	.11	.18	.35	.62	1	1	1
G, prior 1	.03	.03	.08	.13	.21	.36	.62	1	1	1
G, prior 2	.01	.03	.04	.10	.17	.36	.60	1	1	1

Table 5.6: π_k with mean 1, sample size 100

Sample	k .1	.2	.3	.4	.5	.6	.7	.8	.9	1
P, prior 1	0	0	0	0	0	0	0	0	0	0
P, prior 2	0	0	0	0	0	0	0	0	0	0
NB(5), prior 1	0	0	0	0	0	0	0	0	0	.01
NB(5), prior 2	0	0	0	0	0	0	0	.01	.01	.01
G, prior 1	1	1	1	1	1	1	1	1	1	1
G, prior 2	1	1	1	1	1	1	1	1	1	1

Table 5.7: Bayes factors and maximised likelihood ratios

Sample			B, $s=0$	B, $s=1$	MLR
size 10	mean 0.8	P	1.544	1.716	2.035
		NB(5)	3.243	3.604	4.572
		G	1.081	1.201	1.475
	mean 0.9	P	3.771	4.190	5.951
		NB(5)	3.771	4.190	5.951
		G	0.594	0.660	0.754
	mean 1	P	2.595	2.883	3.772
		NB(5)	0.541	0.601	0.738
		G	0.541	0.601	0.762
size 100	mean 0.8	P	5.653	5.710	7.248
		NB(5)	61.793	62.417	83.071
		G	0.000	0.000	0.000
	mean 0.9	P	5345.472	5399.467	7484.709
		NB(5)	0.001	0.001	0.001
		G	0.000	0.000	0.000
	mean 1	P	42751.64	43183.47	61248.78
		NB(5)	5.332	5.386	7.025
		G	0.000	0.000	0.000

5.6 Discussion

If we consider firstly the tables concerning the samples of size 10, we see that the choice of prior distribution has very little effect on the estimates obtained for π_k and that in these examples the evidence provided by π_k is supporting, or not supporting, the Poisson hypothesis against the geometric both strongly and correctly when the actual data is taken from the Poisson or geometric respectively. In these two cases, therefore, π_k is identifying the correct model. However, the samples from the negative binomial do support the Poisson hypothesis for means of 0.8 and 0.9 but provide some evidence against it in the case where the mean is taken to be 1.

The same general pattern is observed in the tables showing the results when the sample size is 100, however in this case the sample from the negative binomial with mean of 0.9 does not support the Poisson hypothesis while those with a mean of 0.8 or 1 do. What is most striking in these tables, however is the fact that we only obtain $\pi_k = 0$ or $\pi_k = 1$ and that there is no apparent variation with k. This latter observation is explained in the

following manner. If we recall that

$$g(\mu) = \frac{\mu - \bar{y}\log(1+\mu) + \frac{1}{n}\left(\log k + \sum_{i=1}^{n} \log y_i!\right)}{\log(1+\mu)},$$

and that $\pi_k = Pr(g(\mu) > 1)$ then it is clear to see that k only enters $g(\mu)$ through the $(\log k)/n$ term. This means that in the range of k being considered there is very little change in the value of $g(\mu)\log(1+\mu)$ with k as $(\log 1)/n = 0$ and $(\log 0.1)/n = -2.3/n$. This means that in the $n = 10$ case the maximum difference between values of $g(\mu)\log(1+\mu)$ is approximately 0.23 whereas in the $n = 100$ case this difference is at most 0.023. Therefore, particularly for larger sample sizes, we observe little variation of π_k with k. We may interpret this result as a covergence of the likelihood ratio to that for the true μ, since as $n \to \infty$ $\bar{y}$ converges to μ, removing the nuisance parameter and hence giving a *fixed* likelihood ratio for Poisson to geometric.

We now consider the results we obtain from the Bayes factors (or indeed the maximised likelihood ratio which in these examples behaves in a similar fashion). Table 5.7 clearly shows that changing the prior distribution to that with $s = 1$, in this case, has very little effect. Recalling that for both the Bayes factor and MLR values greater than 1 indicate support for the Poisson hypothesis (this support increasing as the size of B or MLR increases) while values less than 1 similiarly indicate support for the geometric hypothesis, we are able to compare these with our π_k method. Examination of the tables shows that, for the smaller sample size we obtain stronger support for the (correct, in the Poisson and geometric sample case) favoured hypothesis when using the π_k method. For the larger sample size we have broad agreement between the methods, with slightly stronger conclusions arising from the π_k approach.

Once more, we are also able to plot the distribution function of the likelihood ratio regarded as a function of k, as well as an estimate of the density. In this case we must use a crude estimate of the gradient between each value of k considered. This is due to our inability to obtain a closed form for the posterior distribution above. The following Figures (5.2 to 5.4) show the plots for the case where our sample mean is 0.8 and we have a sample size of 10. A larger range of $k \in [0, 3]$ has been used in these plots in order to better observe features of the distribution. It should also be noted that different samples have been used to generate these plots than were used to give the previous tables of π_k. The use of a sample size of 10 means that the make up of these samples can be very different, explaining the different values for π_k in the range of $k \in [0, 1]$. In addition this dependence on the particular data samples explains the small difference between the plots for the Poisson

and geometric in this case. These plots appear to be very similar with the Poisson peaking around $k = 2$ and the geometric, which we may expect to have a peak below $k = 1$, peaking around $k = 1.5$.

Examination of these figures shows that, as we would expect, the density for the geometric distribution is concentrated at smaller values of k than for that of the Poisson distribution. There appears to be little or no difference between the two prior distributions considered for any of the three samples. However, in the negative binomial case we still obtain only relatively small values for π_k and thus we see no clear pattern emerging in this general case. This is to be expected as the negative binomial sample is taken from neither the Poisson nor the geomteric and therefore we may expect more inconclusive results.

5.7 Another example

We now use our method of analysis to examine the dataset used by Cox (1962). The data are shown in Table 5.8 and are a sample of size 30 drawn from a Poisson with mean 0.8. We conduct our analysis in the fashion described above, again using both priors, and the results are shown in Table 5.9. The maximised likelihood ratio and Bayes factors,

$$MLR = 20.126, \quad B = \begin{cases} 14.221 & \text{if } s = 0 \\ 14.712 & \text{if } s = 1, \end{cases}$$

suggest that we consider $k \in [14, 20]$ in addition to values of $k \in (0, 1]$.

Table 5.8: Cox (1962) data

Variate value	Observed frequency
0	12
1	11
2	6
3	1
≥ 4	0

We can clearly see from the values of π_k shown that we obtain very strong support for the Poisson hypothesis and that, as we suspected, the distribution of the likelihood ratio appears to be concentrated in the range $[14, 20]$. In the analysis of the data in his 1962 paper Cox finds a ratio of roughly 60 to 1

Figure 5.2: Distribution (left) and density (right) plots for both prior 1 (top) and prior 2 (bottom) - Poisson sample

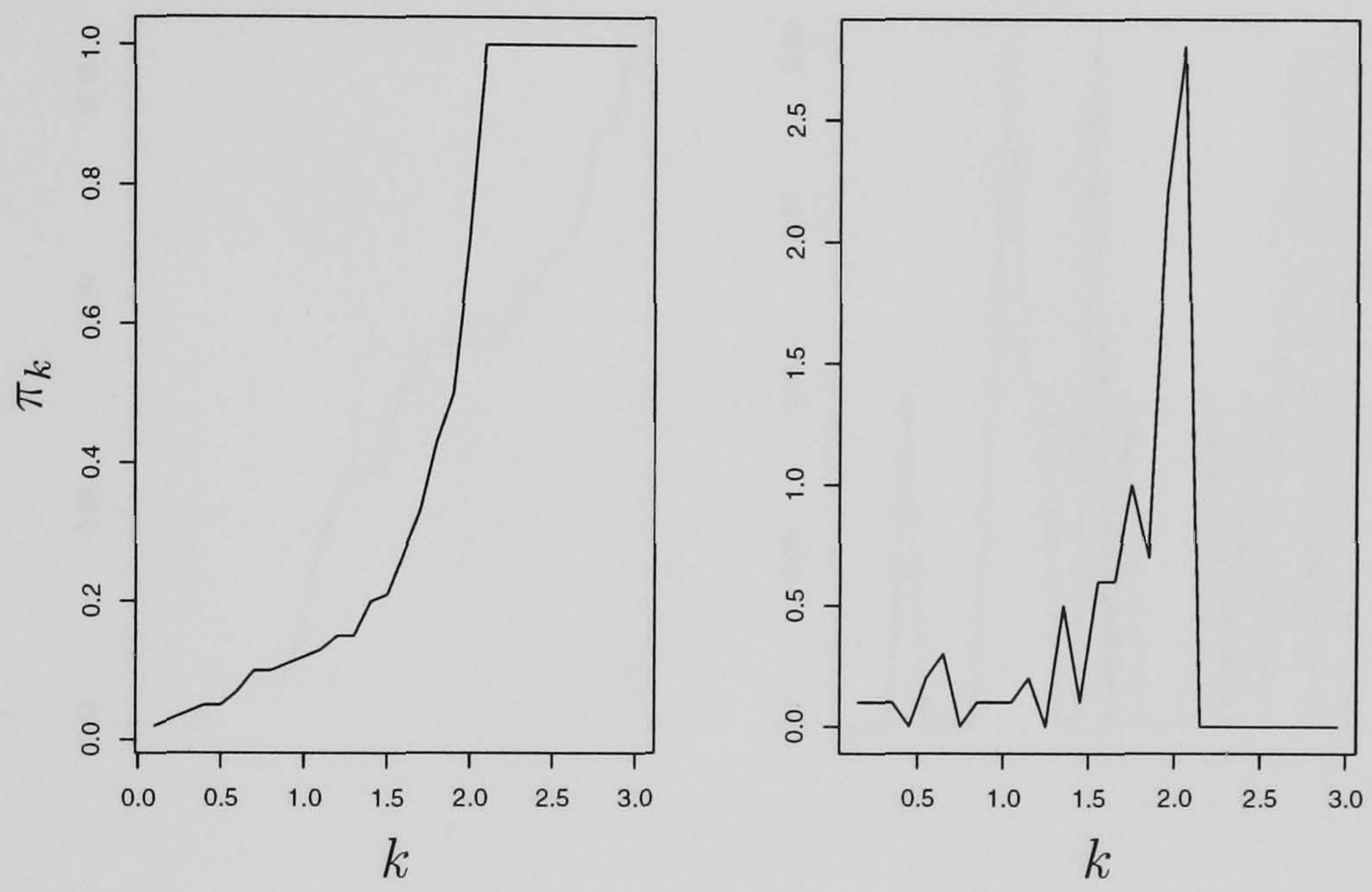
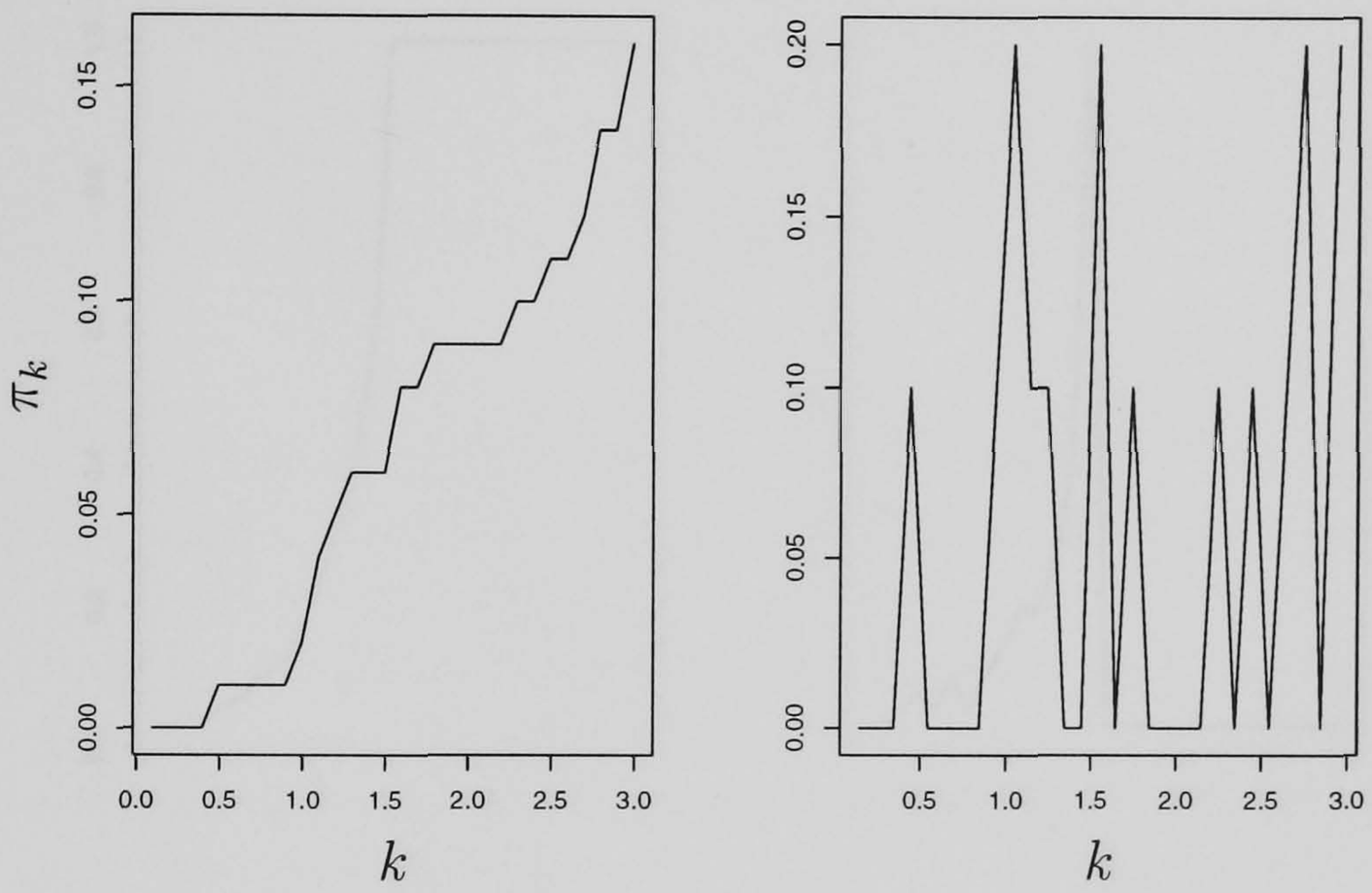


Figure 5.3: Distribution (left) and density (right) plots for both prior 1 (top) and prior 2 (bottom) - Negative Binomial sample



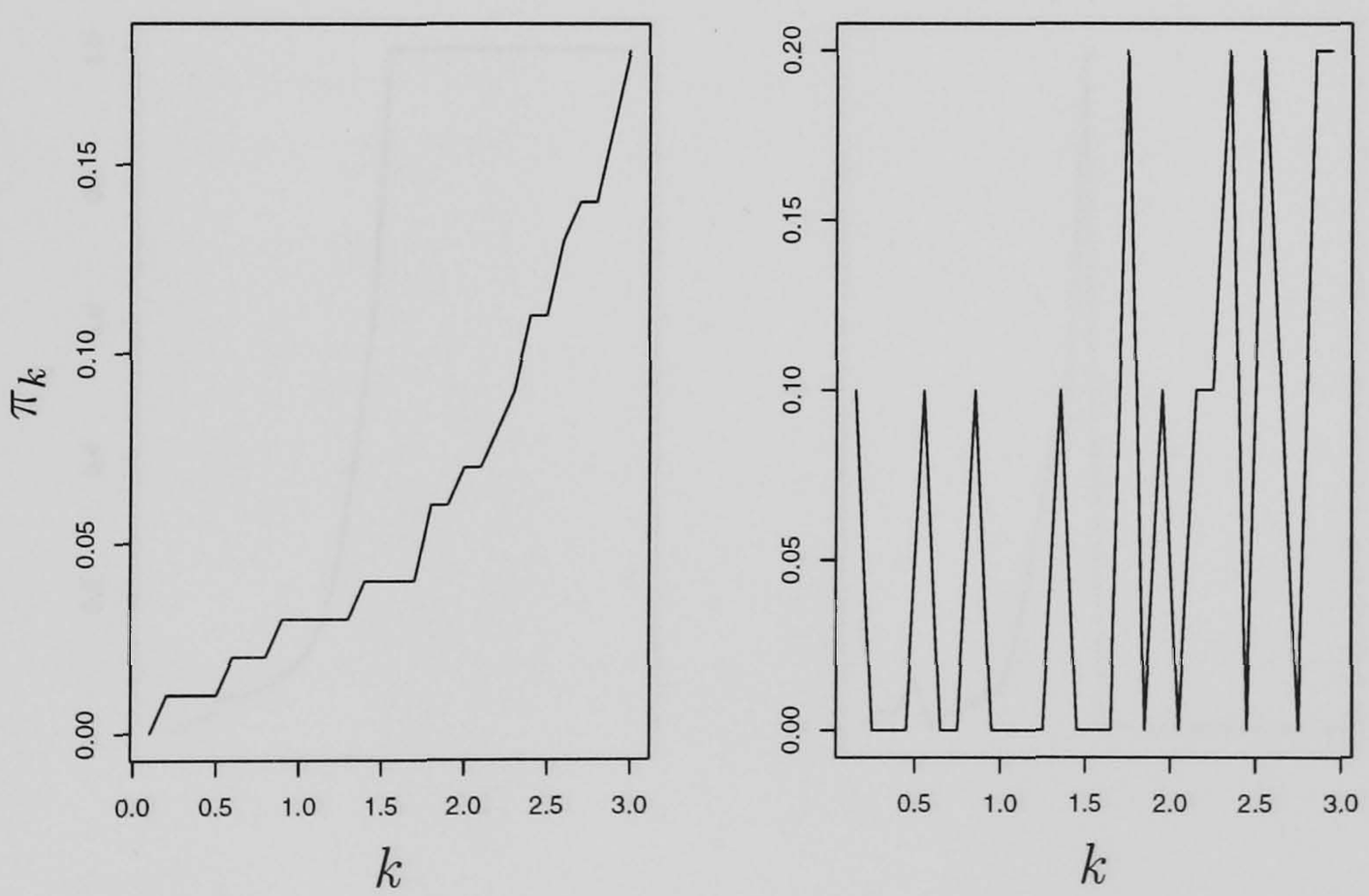
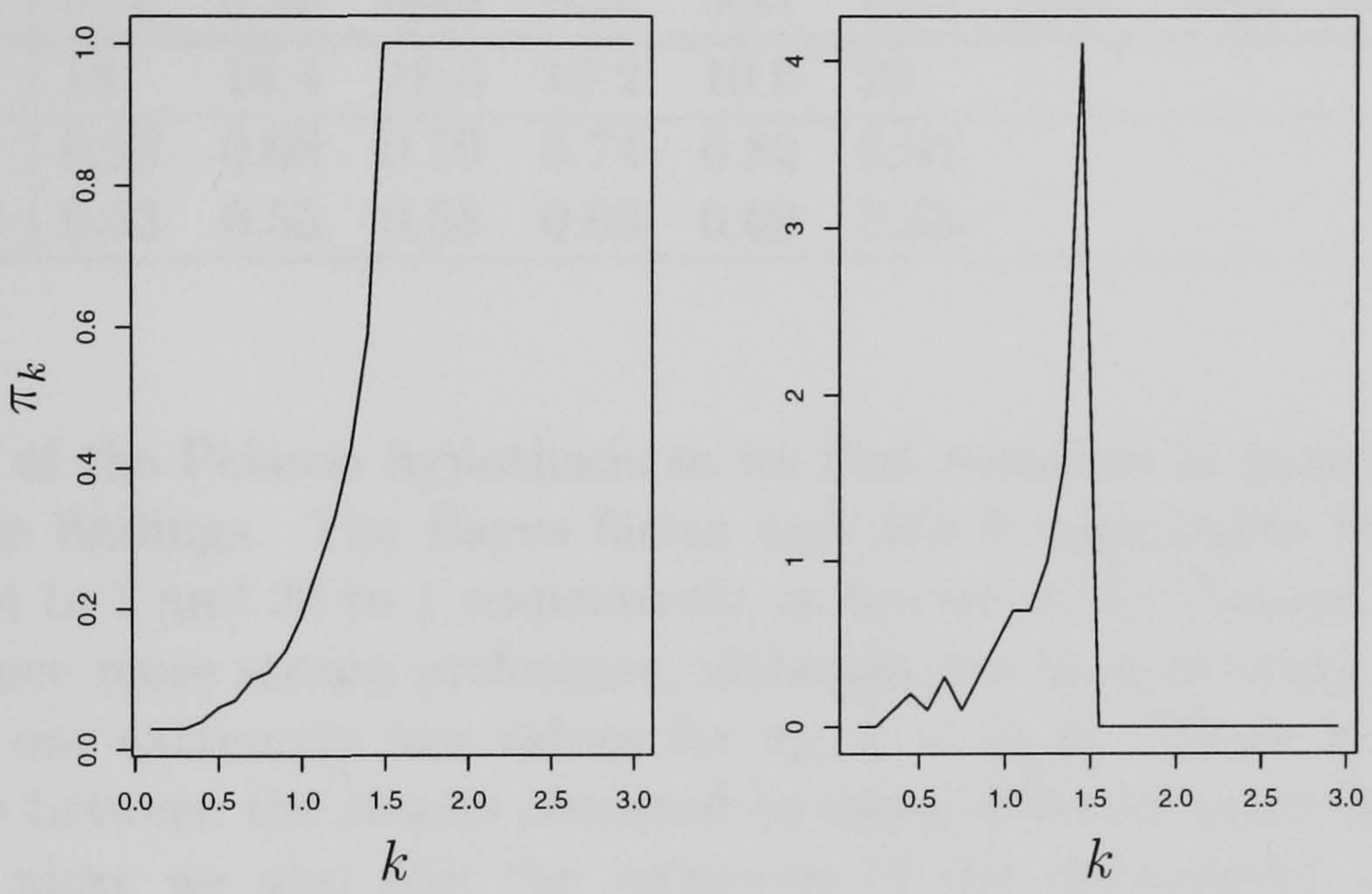


Figure 5.4: Distribution (left) and density (right) plots for both prior 1 (top) and prior 2 (bottom) - Geometric sample



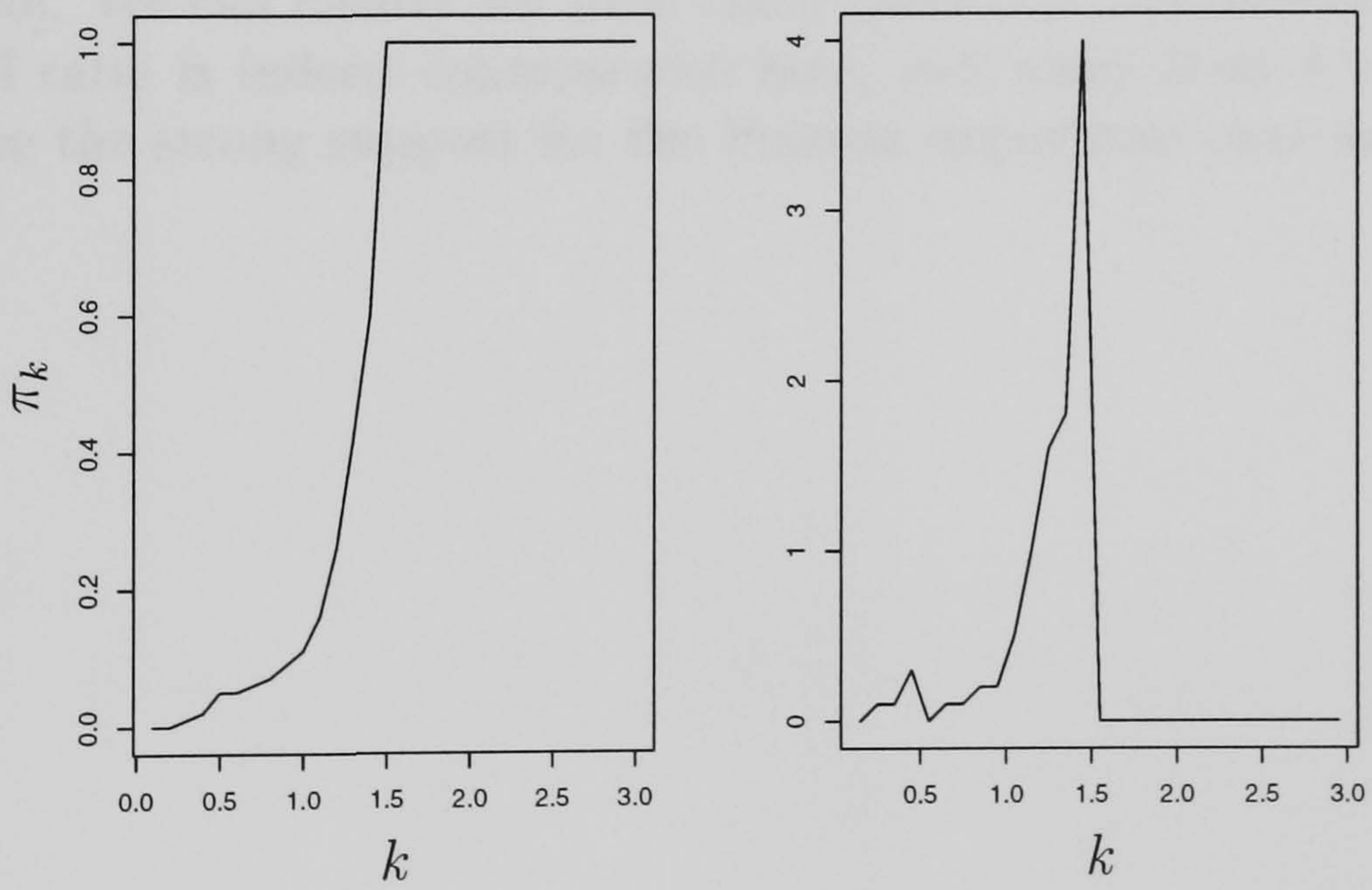


Table 5.9: π_k for the Cox (1962) data

k	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
prior 1	0	0	0	0	0	0	0	0	0	0
prior 2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

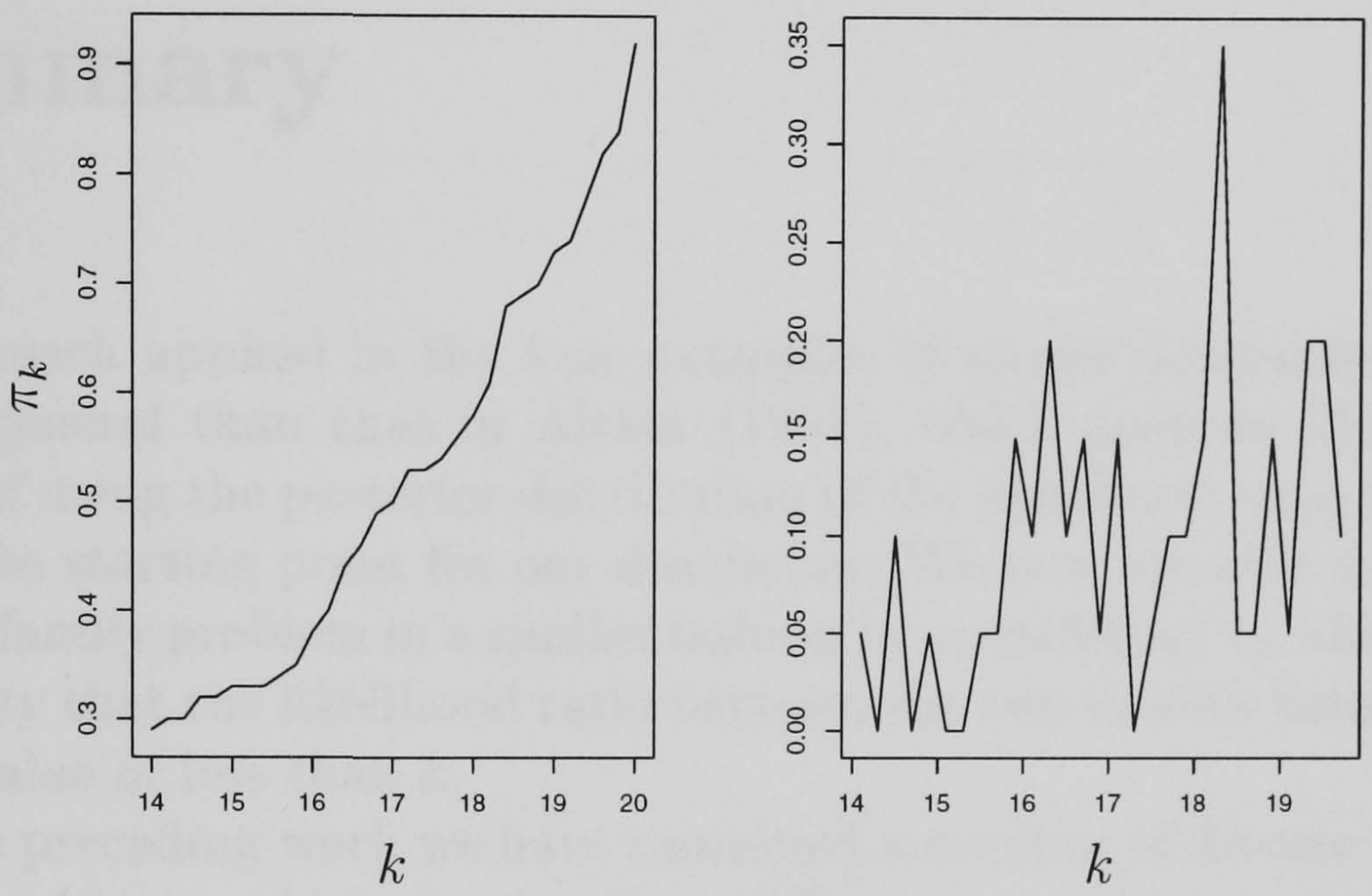
k	14	14.4	14.8	15.2	15.6	16	16.4	16.8	17.2	17.6
prior 1	0.29	0.30	0.32	0.33	0.34	0.38	0.44	0.49	0.53	0.54
prior 2	0.28	0.30	0.33	0.35	0.37	0.37	0.40	0.44	0.47	0.49

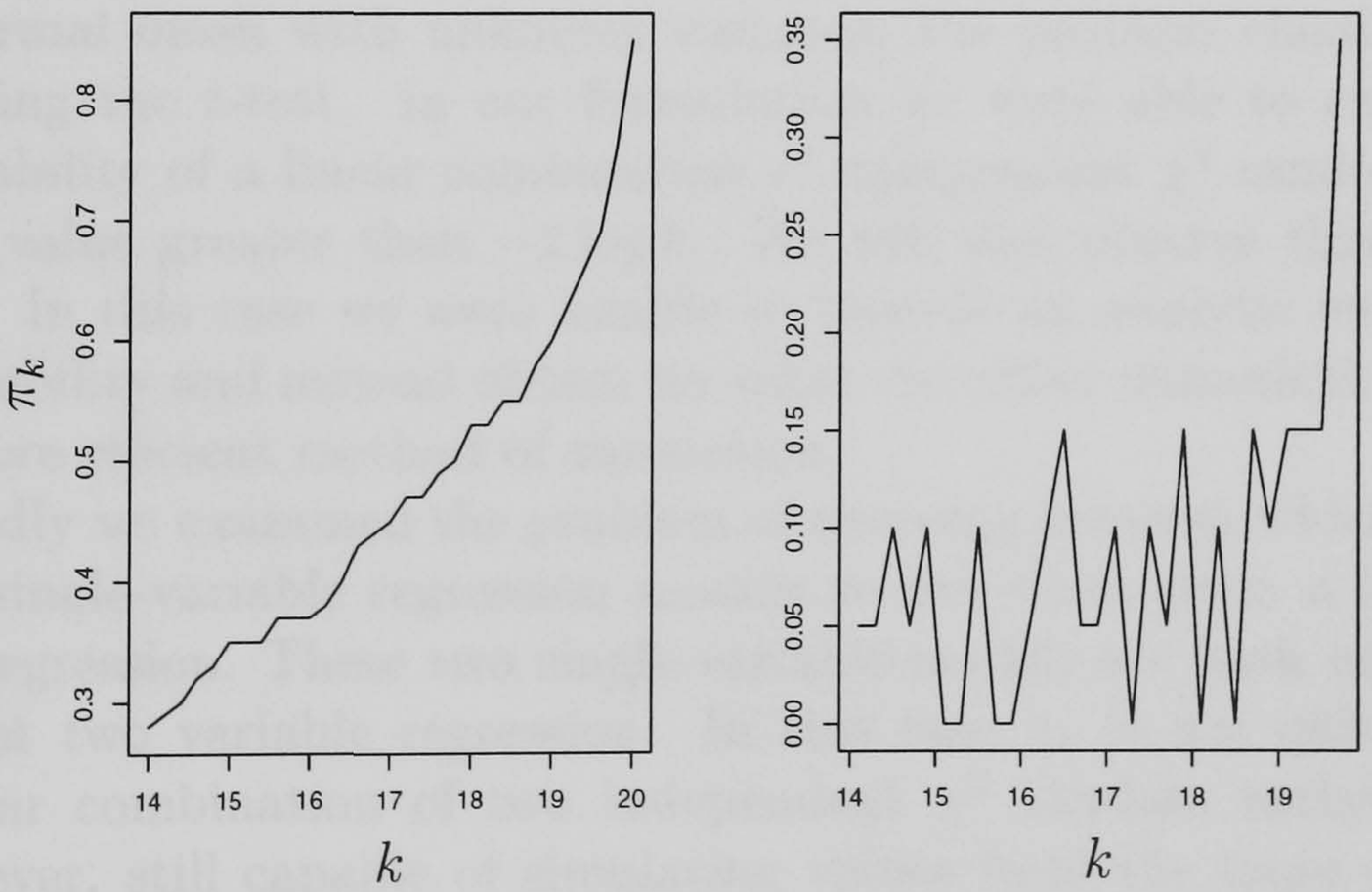
k	18	18.4	18.8	19.2	19.6	20
prior 1	0.58	0.68	0.70	0.74	0.82	0.92
prior 2	0.53	0.55	0.58	0.63	0.69	0.85

in favour of the Poisson hypothesis so we find ourselves in broad agreement with these findings. The Bayes factor and MLR approaches find ratios of around 14 to 1 and 20 to 1 respectively, in favour of the Poisson hypothesis. This is once more strong preference, although not as convincing as the Cox result or our extremely low values for π_k, $k \in (0, 1]$. There is again little difference between the results obtained by using different prior distributions.

Once more we also plot the estimates of the distribution and density as functions of k. Figure 5.5 shows these for both priors using the range $k \in [14, 20]$. We can clearly see from these plots that the distribution of the likelihood ratio is indeed concentrated here, well away from $k = 0$, so that we can see the strong support for the Poisson hypothesis once more.

Figure 5.5: Distribution (left) and density (right) plots for both prior 1 (top) and prior 2 (bottom)





Chapter 6

Summary

The approach applied in the four examples of model comparison problems is more general than that in Aitkin (1997), which used an identical basic concept of using the posterior distribution of the likelihood ratio statistic and formed the starting point for our discussion. We now are able to handle the common family problem in a similar fashion by considering π_k, the (posterior) probability that the likelihood ratio between the two models being compared takes a value of less than k.

In the preceding work we have examined a number of diverse model comparison problems, which despite their differences can now be handled using the same general approach. We firstly considered the familiar case of testing a Normal mean with unknown variance, the problem classically solved by applying the t-test. In our formulation we were able to express π_k as the probability of a linear combination of independent χ^2 random variables taking a value greater than $-2\log k$. We will also observe this in another example. In this case we were unable to provide an analytic expression for this probability and instead obtain its value via either numerical integration, or the more efficient method of simulation.

Secondly we examined the problem of choosing between which of the two possible single-variable regression models to use when given a two variable Normal regression. These two single-variable models are both nested within the parent two variable regression. In this case π_k is not quite as simple as a linear combination of two independent χ^2 random variables but we are, however, still capable of simulating values from the more complicated distributions involved and hence we can evaluate π_k in a straightforward fashion.

We also studied the problem of deciding whether time series data came from a general AR(1) model or from a non-stationary random walk, a special case of the auto-regressive model. Once more we were able to express π_k as

the probability of a linear combination of independent χ^2 random variables taking a value greater than $-2\log k$. In this case, however, because of particular properties of the χ^2_2 distribution (this is an exponential distribution with mean 2) we are able to complete the integration analytically and obtain an exact solution for π_k (in terms of the distribution function of the χ^2 distribution) without the need for either numerical integration or simulation.

The final example analysed does not share this property. Indeed, the problem of deciding whether a Poisson or a geometric distribution is better supported by a sample of discrete data raises another problem. Once we have nested both these distributions within the negative binomial distribution we find that there is no conjugate prior that we can place on the index of the negative binomial distribution and hence we are unable to obtain a straightforward posterior to use for the simulation we require in order to estimate π_k. Details of the method we employ for our simulation can be found in the appropriate chapter but we may say here that it allows us to simulate (at least approximately) from *any* posterior distribution irrespective of the existence of conjugate priors. Using this we complete the simulation and are therefore able to evaluate π_k.

The examples studied in the previous chapters have demonstrated the use of an alternative method of model comparison which, as shown in the introduction for a simple example, avoids the Lindley paradox. Further demonstration of the lack of agreement between our conclusions, and indeed those from standard frequentist methods, with those from Bayes factor methods when considering the specific cases discussed in the body of the thesis would be an option for continued study. As mentioned in the introduction, and emphasised later in this section, this is not possible for all the examples considered, when considering the uninformative priors that we have chosen to use.

The ability of our method to analyse the common family problem is not a property exhibited by the standard Bayes factor methods, as we have seen by our inability to demonstrate their use with the uninformative priors selected for our comparison. Indeed it is straightforward to adapt our method to allow *any* prior distribution to be placed on the parameters as the particular approach applied in the Poisson versus geometric example can be used in any situation where exact analytic or more standard simulation techniques are impossible to find or apply. Common family problems are more difficult by Bayes factor methods because the informative reference priors, if they exist, for the two models may well be different; the common-family approach that we apply allows a single common prior for the nuisance parameter under both models.

We have seen that it is relatively simple to apply Monte Carlo simulation

methods to these examples, even for cases where the nuisance parameter prior does not have an analytic form. We have seen that for *any* such problem we are now able to use the simulation method (at least approximately) to obtain an estimate for π_k.

Although nested and common-family problems cover a great deal of model comparison problems they are by no means the only form that these can take. Conjectural ways of dealing with the comparison of more general models, apart from Dempster's suggestion of using the posterior distribution of the deviance under each model, include using mixtures or other (artificial) functions of the two models to force them into a composite single family, to which the common-family approach can then be applied.

Appendix A

Bibliography

Aitkin, M., Anderson, D., Francis, B. & Hinde, J. (1989) *Statistical modelling in GLIM*. Oxford, Oxford University Press.

Aitkin, M. (1991) Posterior Bayes factors (with Discussion). *Journal of the Royal Statistical Society*, **B 53**, 111-142.

Aitkin, M. (1992) Evidence and the posterior Bayes factor. *Mathematical Scientist*, **17**, 15-25.

Aitkin, M. (1997) The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. *Statistics and Computing*, **7**, 253-261.

Aitkin, M. (1998) Simpson's Paradox and the Bayes factor. *Journal of the Royal Statistical Society*, **B 60**, 269-270.

Akaike, H. (1973) Information theory and the extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds. B. N. Petior & F. Csaki), pp. 267-281. Budapest: Akademiai Kiado.

Atkinson, A. C. (1970) A method for discriminating between models (with Discussion). *Journal of the Royal Statistical Society*, **B 32**, 323-353.

Bartlett, M. S. (1957) A comment on D. V. Lindley's statistical paradox. *Biometrika*, **44**, 533-534.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis (Second edition)*. New York, Springer-Verlag.

Berger, J. O. & Sellke, T. (1987) Testing a point null hypothesis: the irreconcilability of P values and evidence. *Journal of the American Statistical Association*, **82**, 112-122.

Berry, S. M. (1998) Understanding and testing for heterogeneity across 2×2 tables: application to meta-analysis. *Statistics in Medicine*, **17**, 2353-2369.

Carlin, B. P. & Chib, S. (1995) Bayesian model choice via Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society*, **B 57**, 473-484.

Cox, D. R. (1958) Some problems connected with statistical inference. *Annals of Mathematical Statistics*, **29**, 357-372.

Cox, D. R. (1962) Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society*, **B 24**, 406-424.

Dempster, A. P. (1974,97) The direct use of likelihood for significance testing. In *Proc. Conf. Foundational Questions in Statistical Inference* (eds. O. Barndorff-Nielsen, P. Blaesild & G. Sihon), pp. 335-354. University of Aarhus. Reprinted in *Statistics and Computing*, **7**, 247-252.

Edwards, A. W. F. (1972) *Likelihood.* Cambridge, Cambridge University Press.

Feller, W. (1968) *An Introduction to Probability Theory and its Applications (Volume I - Third edition).* New York, John Wiley.

Freud, J. E. (1962) *Mathematical Statistics.* Englewood Cliffs, Prentice-Hall.

Graybill, F. A. & Mood, A. M. (1963) *Introduction to the Theory of Statistics (Second edition).* New York, McGraw-Hill.

Hoel, P. G. (1962) *Introduction to Mathematical Statistics (Third edition).* New York, John Wiley.

Kent, J. T. (1982) Robust properties of likelihood ratio tests. *Biometrika*, **69**, 19-27.

Kent, J. T. (1986) The underlying structure of nonnested hypothesis tests. *Biometrika*, **73**, 333-343.

Krzanowski, W. J. (2000) *Principles of Multivariate Analysis: A User's Perspective (Revised edition).* Oxford, Oxford University Press.

Lindley, D. V. (1957) A statistical paradox. *Biometrika*, **44**, 187-192.

Lindley, D. V. (1993) On the presentation of evidence. *Mathematical Scientist*, **18**, 60-63.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979) *Multivariate Analysis.* London, Academic Press.

Mariott, J. M. & Newbold, P. (1998) Bayesian comparison of ARIMA and stationary ARMA models. *International Statistical Review*, **66**, 323-336.

Nelder, J. A. & Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society*, **A 135**, 370-384.

O'Hagan, A. (1994) *Kendall's Advanced Theory of Statistics, Volume 2B. Bayesian Inference.* Edward Arnold.

O'Hagan, A. (1995) Fractional Bayes factors for model comparison (with Discussion). *Journal of the Royal Statistical Society*, **B 57**, 99-138.

Pitman, E. J. G. (1936) Sufficient statistics and intrinsic accuracy. *Proceedings of the Cambridge Philosophy Society*, **32**, 567-579.

Pitman, E. J. G. (1937) The "closest" estimates of statistical parameters. *Proceedings of the Cambridge Philosophy Society*, **33**, 212-222.

Royall, R. (1997) *Statistical Evidence: A Likelihood paradigm.* Boca Raton, Chapman and Hall - CRC.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Severini, T. A. (2000) *Likelihood Methods in Statistics.* Oxford, Oxford University Press.

Shafer, G. (1982) Lindley's paradox (with Discussion). *Journal of the American Statistical Association*, **77**, 325-351.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., Naylor, J. C. & Dransfield, M. (1985) The implementation of the Bayesian paradigm. *Communications in Statistics*, **A 14**, 1079-1102.

Smith, A. F. M. & Spiegelhalter, D. J. (1980) Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society*, **B 42**, 213-220.

Spiegelhalter, D. J. & Smith, A. F. M. (1982) Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society*, **B 44**, 377-387.

Stirling, J. (1730) Methodus differentialis, sive tractatus de summation et interpolation serierum infinitarium. *London.* English translation by Holliday, J. (1749) The differential method: A treatise of the summation and interpolation of infinite series.

Stone, M. (1997) Discussion of papers by Dempster and Aitkin. *Statistics and Computing*, **7**, 263-264.

Williams, E.J. (1959) *Regression Analysis.* New York, John Wiley.