

# PREDICTIVE INFERENCE TOOLS FOR RESEARCHERS

by

Voyze G. Harris III

---

Copyright © Voyze G. Harris III 2021

A Thesis Submitted to the Faculty of the

STATISTICS AND DATA SCIENCE  
GRADUATE INTERDISCIPLINARY PROGRAM

In Partial Fulfillment of the Requirements  
For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2021

THE UNIVERSITY OF ARIZONA  
GRADUATE COLLEGE

As members of the Master's Committee, we certify that we have read the thesis prepared by Voyze Gabriel Harris III, titled *[Enter Thesis Title]* and recommend that it be accepted as fulfilling the dissertation requirement for the Master's Degree.

\_\_\_\_\_  
Dr. Dean Billheimer

Date: \_\_\_\_\_

\_\_\_\_\_  
Dr. Edward Bedrick

Date: \_\_\_\_\_

\_\_\_\_\_  
Dr. Walter Piegorsch

Date: \_\_\_\_\_

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to the Graduate College.

I hereby certify that I have read this thesis prepared under my direction and recommend that it be accepted as fulfilling the Master's requirement.

\_\_\_\_\_  
Dr. Dean Billheimer  
Master's Thesis Committee Chair  
*Biostatistics*

Date: \_\_\_\_\_



ARIZONA

# Contents

<b>1</b>	<b>Thesis Abstract</b>	<b>5</b>
<b>2</b>	<b>Introduction: Predictive Inference</b>	<b>5</b>
2.1	Why Predictive Inference? . . . . .	5
2.2	The Bayesian Parametric Prediction Format . . . . .	10
<b>3</b>	<b>Predictive Problems with Conjugate Priors</b>	<b>12</b>
3.1	Prediction of Future Successes: Beta-Binomial (Geisser p. 73) . . . . .	12
3.1.1	Derivation . . . . .	12
3.1.2	R Implementation (Beta-Binomial) . . . . .	13
3.1.3	Example . . . . .	14
3.2	Survival Time: Exponential-Gamma (Geisser p. 74) . . . . .	15
3.2.1	Derivation . . . . .	15
3.2.2	R Implementation (Exponential-Gamma) . . . . .	17
3.2.3	Example . . . . .	17
3.3	Poisson-Gamma Model (Hoff p. 43 - 50) . . . . .	19
3.3.1	Derivation . . . . .	19
3.3.2	R Implementation (Poisson-Gamma) . . . . .	21
3.3.3	Example . . . . .	22
3.4	Normal Observation with Normal-Inverse Gamma Prior . . . . .	24
3.4.1	One sample Normal-Inverse Gamma . . . . .	24
3.4.1.1	Derivation . . . . .	24
3.4.1.2	R Implementation (Normal-Inverse Gamma, 1-sample) . . . . .	25
3.4.1.3	Example . . . . .	27
3.4.2	Two-sample Normal-Inverse Gamma . . . . .	29
3.4.2.1	Derivation . . . . .	29
3.4.2.2	R Implementation (Normal-Inverse Gamma, 2-sample) . . . . .	30
3.4.2.3	Example . . . . .	30
3.4.3	$k$ -sample Normal-Inverse Gamma: Comparing multiple groups . . . . .	32
3.4.3.1	Derivation . . . . .	33
3.4.3.2	R Implementation (Normal-Inverse Gamma, $k$ -samples) . . . . .	33
3.4.3.3	Example . . . . .	35
<b>4</b>	<b>Normal Regression</b>	<b>37</b>
4.1	Least Squares Estimation Example (Hoff p. 149ff.) . . . . .	37
4.2	Bayesian Estimation for a Regression Model (Hoff p. 154ff) . . . . .	41
4.2.1	Derivation . . . . .	41
4.2.1.1	A semiconjugate prior distribution . . . . .	41
4.2.1.2	Default and weakly informative prior distributions . . . . .	42
4.2.2	R Implementation (Normal Regression) . . . . .	43
4.2.3	Example . . . . .	44
<b>5</b>	<b>Conclusion</b>	<b>45</b>
<b>6</b>	<b>Appendix</b>	<b>45</b>



# 1 Thesis Abstract

An obstacle to widespread employment of Bayesian predictive inference in scientific research is the lack of suitable computing tools. In this thesis I document several established useful models, and provide an applicable set of tools for statisticians. For each of the included models, some basic notes on mathematical derivation are presented, and predictive inference is illustrated with examples. For the details of the models and some of the examples I relied primarily on Seymour Geisser's Predictive Inference: An Introduction (1993) and Peter D. Hoff's A First Course in Bayesian Statistical Methods (2009).

An R package has been developed, the main purpose of which is to provide the researcher with a means of producing random samples from predictive distributions. For all the models, the package includes random sample generators. For those models with analytical solutions, density and distribution functions are also provided. The standard R naming convention for these function classes has been adopted: density functions are prefixed with the letter “d,” distribution functions with the letter “p,” and random generation functions with the letter “r.” Also included in all function names is the abbreviation “pred” (for predictive) and an initialism or abbreviation identifying the model itself. For example, the density function for the Beta-Binomial model is named “`dpredBB()`.” The R code for each function is included in Appendix [X-insert link to appendix here](#).

## 2 Introduction: Predictive Inference

My understanding of Bayesian Predictive Inference began with the University of Arizona course “Bayesian Statistical Theory and Applications” under Dr. Edward Bedrick. Since then it has been shaped by exposure to various sources, including articles by Bedrick and Dr. Dean Billheimer, and to a greater extent the aforementioned texts by Hoff and Geisser, and others, including Bayesian Data Analysis by Andrew Gelman et. al., Statistical Prediction Analysis by J. Aitchison and I.R. Dunsmore, and excerpts from The Signal and the Noise by Nate Silver. In the next section and what follows, the ideas expressed are an amalgam of what was learned from this body of scholarship.

### 2.1 Why Predictive Inference?

I find persuasive the assertion that the main purpose of statistics is to predict future events based on observed data, and moreover that a statistical model may be judged chiefly by the quality of its predictions. Predictions about meaningful quantities that are relevant to the object of study facilitate scientific progress in multiple ways. Nate Silver notes:

Prediction is important because it connects subjective and objective reality. Karl Popper, the philosopher of science, recognized this view. For Popper, a hypothesis was not scientific unless it was falsifiable—meaning that it could be tested in the real world by means of a prediction.

Billheimer emphasizes the improvement of scientific accuracy and reproducibility Bayesian prediction provides by “shifting our focus from ‘finding differences’ among hypothetical

parameters to predicting observable events based on our current scientific understanding.” Focus on observed data enables corroboration or refutation of current hypotheses through future experimentation, and informs decision-making by summarizing quantities of direct interest to the researcher. Bayesian predictive inference shifts the focus of statistical analysis from estimation of hypothetical parameters to statements about concrete observables.

It is not the intent of this thesis to suggest that parametric inference should be abandoned in statistical analyses. Conventional statistical inference techniques are useful for summarizing information about large quantities of data in a small number of usable values (parameter estimates), and leveraging such summaries to determine whether a particular problem merits continued attention. Indeed, the scientific discipline of mathematical statistics has developed along parametric lines, imparting greater efficiency to the solving of idealized versions of problems. However, this focus on parameters has largely ignored the impact of observable quantities which may exhibit distributions very different from our (idealized) standard models. In support of prediction in general, and Bayesian prediction in particular, I return to Nate Silver:

Making predictions based on our beliefs is the best (and perhaps even the only) way to test ourselves. If objectivity is the concern for a greater truth beyond our personal circumstances, and prediction is the best way to examine how closely aligned our personal perceptions are with that greater truth, the most objective among us are those who make the most accurate predictions. Fisher’s statistical method, which saw objectivity as residing within the confines of a laboratory experiment, is less suitable to this task than Bayesian reasoning.

Prediction is a means of discriminating between scientific hypotheses. Generally, a model may be judged by the quality of its predictions. Given competing models, the better predictor will be given more weight, and a useful model increases in utility as its predictive capability improves. With Bayesian inference, prior informed opinion about parameter distributions, updated with knowledge gained through observations, yields better predictions with each update.

To illustrate the potential difference between Bayesian prediction and the use of plug-in estimators, consider the game Pass the Pigs<sup>®</sup>, a push-your-luck dice game in which the “dice” are actually rubber pig figures. Two pig dice are thrown, and points are scored according to the combination of positions in which they come to rest. Details about the game can be found on Wikipedia here: [https://en.wikipedia.org/wiki/Pass\\_the\\_Pigs](https://en.wikipedia.org/wiki/Pass_the_Pigs)



Figure 1: Pass The Pigs<sup>®</sup>



Figure 2: Razorback

For the purpose of this example, consider the probability of a single pig landing in the “Razorback” position, which occurs when the pig is lying on its back with its legs extended upward. The irregular shape of the pig makes it difficult to assign probabilities to results other than by means of experimentation. Such an experiment was conducted at Duquesne University, and an article describing the experiment as well as Bayesian predictive inference performed on the results appeared in the *Journal of Statistics Education* Volume 14, Number 3, in 2006. The article can be accessed here: <http://jse.amstat.org/v14n3/datasets.kern.html>. Of the 11,954 recorded results for individual pigs, approxi-

mately 22.4% were Razorbacks.

Suppose in new data  $t = 4$  Razorbacks have been observed out of  $N = 10$  tosses of a single pig die, suggesting a straightforward binomial distribution with  $\theta = \text{Pr}(\text{Razorback}) = t/N = 0.4$ . Taking the Duquesne experiment into consideration, we’ll perform Bayesian prediction using various Beta prior distributions for  $\theta$ :  $\theta \sim \text{Beta}(2, 8)$ ,  $\theta \sim \text{Beta}(22, 78)$ , and  $\theta \sim \text{Beta}(224, 776)$ , and compare these results to predictions obtained from the plug-in estimator  $\theta = 0.4$ . Any number of prior distributions on  $\theta$  would satisfy the condition that  $E(\theta) \approx 0.224$ , suggested by the prior information. The specific choice of a Beta prior is made largely for computational convenience.

The researcher wants to predict how many Razorbacks will result from some number  $M$

of future pig die rolls. In this example,  $M = 100$  was used. The density curves in the plot below show the influence of the choice of prior parameters on the location and variance of the predictive distribution. Essentially, each pair of shape parameters  $(\alpha, \beta)$  in the Beta prior reflects the researcher's level of reliance on the results of the Duquesne experiment, with the "weight" given to that knowledge increasing with the shape parameters by orders of magnitude. The choice of parameters might be influenced by such things as pig tossing method (perhaps the researcher is throwing them by hand rather than dropping them by the carefully controlled method used in the Duquesne experiment), or by a need to account for pig-to-pig variation, or anything else the researcher believes introduces a deviation from the events upon which the prior information is based. Use of the scaled-up Big Pigs<sup>TM</sup> variant, for example might give reason to use a very low-weighted prior such as  $\theta \sim \text{Beta}(2, 8)$ .

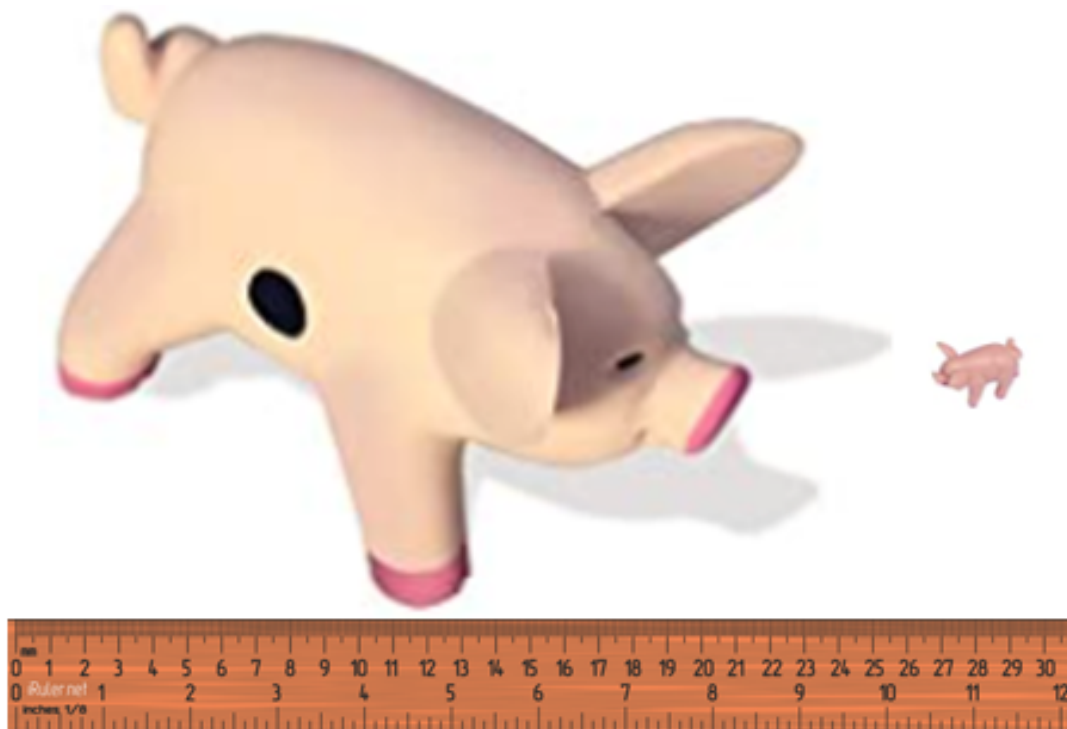
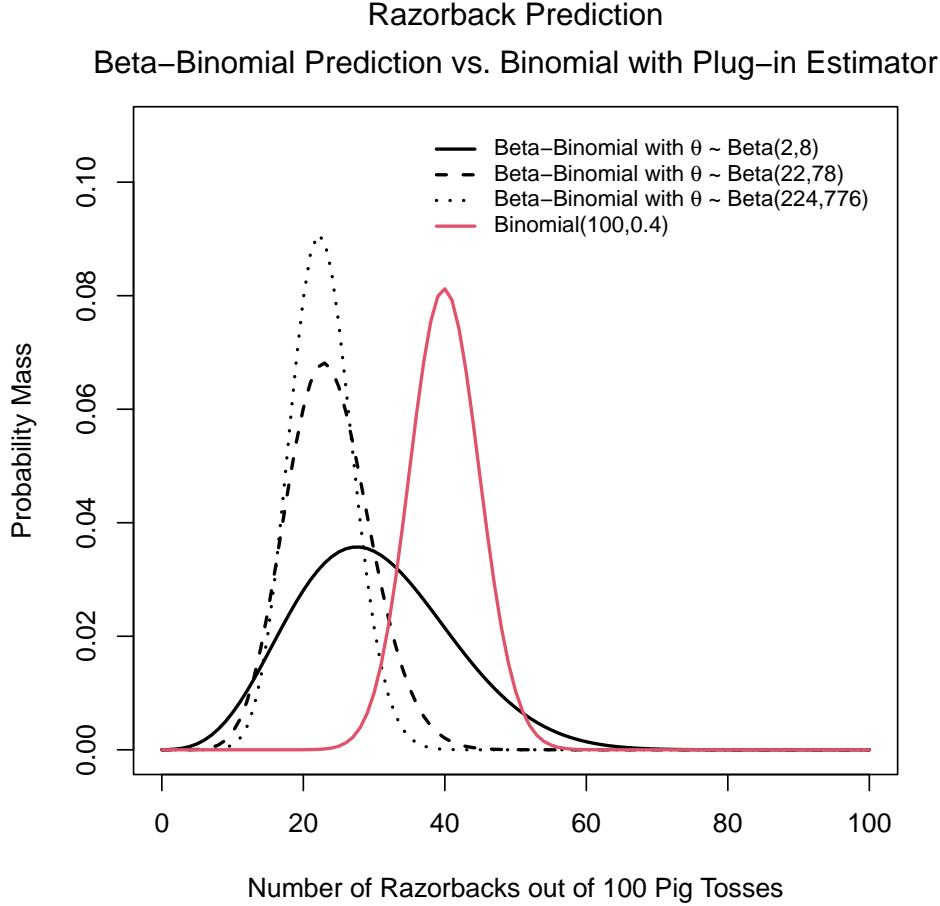


Figure 3: Big Pigs<sup>TM</sup> to Original (Approximate Scale)

The plot and table below illustrate the effects of the Bayesian prediction method. Perhaps most notable is the location disparity between the plug-in (Binomial) prediction and the family of Bayesian (Beta-binomial) predictive distributions. The consideration of prior knowledge is also shown to have a significant effect. In this example, the strong influence of the choice of shape parameters for the Beta prior on the mean and variance of the predictive distribution provides options for the future prognosticator. If a new set of trials closely duplicates the Duquesne experimental conditions, for example, predictions might be based on the result from the Beta(224, 776) prior. A lower-weighted prior might be used for any experimental element the researcher believes introduces a deviation from the knowledge upon which the prior information is based.





#Razorbacks Predicted out of 100 Tosses				
Prediction Method	$(\alpha, \beta)$	Prior Mean $E(\theta)$	Prediction Mean	SD
Beta-Binomial	(2,8)	0.2	29.83	11.01
Beta-Binomial	(22,78)	0.22	23.98	5.99
Beta-Binomial	(224,776)	0.224	22.43	4.51
Binomial(100, 0.4)	–	–	40	4.9

The differences in the results of the two prediction methods are easy to see in this case. Unlike the Bayesian approach, the binomial-only model uses only the observed data, does not account for any uncertainty in the binomial parameter ( $\theta$ ), and ignores historical info about Razorback frequency. If speculating whether the number of razorbacks out of the next 10 pig tosses will be greater than, say, 3, the indications of the two analysis approaches contradict each other. The statistically savvy wagerer of course realizes that the 10 die throws informing the plug-in results do not guarantee high confidence. A much less risky option for the bettor is the Bayesian method with a suitable prior parameterization.

## 2.2 The Bayesian Parametric Prediction Format

We want to predict future outcomes based on current knowledge. Specifically we're asking: for observed values  $Y_1 = y_1, \dots, Y_n = y_n$ , what is likely to be the value of the next observation,  $\tilde{Y}$ ? We want to compute  $Pr(\tilde{Y}|y_1, \dots, y_n)$ , where  $y_1, \dots, y_n$  are conditionally independent and identically distributed (i.i.d.) with respect to a population parameter (or parameters)  $\theta$ . We assign prior distribution  $\pi(\theta)$  based on some existing knowledge or beliefs. Here we are careful to satisfy ourselves that  $Y_1, \dots, Y_n$  are *exchangeable*, which enables us to rely on de Finetti's representation theorem for the conditional independence assumption. Exchangeability means essentially that the order of the observations conveys no information affecting the distribution of the data. That is,  $p(Y_1, \dots, Y_n) = p(Y_{\pi_1}, \dots, Y_{\pi_n})$  for any permutation  $\pi_1, \dots, \pi_n$  of the indices  $1, \dots, n$ .

**De Finetti's Theorem:** Let  $Y_i \in \mathcal{Y}$  for all  $i \in \{1, 2, \dots\}$ . Suppose that, for any  $n$ , our belief model for  $Y_1, \dots, Y_n$  is exchangeable. Then our model can be written as

$$p(y_1, \dots, y_n) = \int \left\{ \prod_{i=1}^n p(y_i|\theta) \right\} p(\theta) d\theta$$

for some parameter  $\theta$ , some prior distribution on  $\theta$  and some sampling model  $p(y|\theta)$ . The prior and sampling model depend on the form of the belief model  $p(y_1, \dots, y_n)$ .

[How is the following paragraph?](#)

Simply put, exchangeability of data implies (conditional) independence and identical distribution (i.i.d.) with respect to a parameter with some prior distribution. The converse is immediate from the definition of marginal probability and the commutative property of multiplication. Hoff summarizes these ideas like this (p. 29):

$$\left. \begin{array}{l} Y_1, \dots, Y_n | \theta \text{ are i.i.d} \\ \theta \sim \pi(\theta) \end{array} \right\} \Leftrightarrow Y_1, \dots, Y_n \text{ are exchangeable for all } n$$

The derivations throughout this thesis assume exchangeability, so we can formulate predictive probability as follows:

$$\begin{aligned}
p(\tilde{Y} = \tilde{y} | Y_1 = y_1, \dots, Y_n = y_n) &= \frac{p(\tilde{y}, y_1, \dots, y_n)}{p(y_1, \dots, y_n)} \\
&= \frac{\int p(\tilde{y}, y_1, \dots, y_n | \theta) \pi(\theta) d\theta}{p(y_1, \dots, y_n)} \\
&= \frac{\int p(\tilde{y} | \theta) p(y_1, \dots, y_n | \theta) \pi(\theta) d\theta}{p(y_1, \dots, y_n)} \\
&= \frac{\int p(\tilde{y} | \theta) p(\theta | y_1, \dots, y_n) p(y_1, \dots, y_n) d\theta}{p(y_1, \dots, y_n)} \\
&= \int p(\tilde{y} | \theta) p(\theta | y_1, \dots, y_n) d\theta \tag{1}
\end{aligned}$$

For prediction, we need only to characterize the observed data  $(y_1, \dots, y_n)$ , conditionally independent with respect to  $\theta$ , and supply a suitable prior distribution  $\pi(\theta)$ . Then we compute posterior  $p(\theta | y) = p(y | \theta) \pi(\theta) / \int_{\theta} p(\theta) p(y | \theta) d\theta$  and generate a prediction using (1). In some cases the integral has an analytical solution, while in others we evaluate it through monte carlo (MC) sampling.

### 3 Predictive Problems with Conjugate Priors

When data can be modeled with a distribution that suggests a conjugate prior for the parameter(s) of interest, and those parameters can reasonably be represented by that conjugate prior, the posterior calculations are greatly simplified. The four classes of models addressed in this section all exhibit this feature. The first three are single-parameter exponential families and have closed-form solutions for prediction. The fourth is a two-parameter exponential family with unknown mean and variance. While this one does not admit an analytical solution, prediction is easily accomplished by means of simple monte carlo sampling. The four classes of predictive models presented in this section are:

- Beta-Binomial ( $T = \sum Y \sim \text{Binom}(N, \theta)$  with  $\theta \sim \text{Beta}$ )
- Exponential-Gamma ( $Y \sim \text{Exp}(\theta)$  data with  $\theta \sim \text{Gamma}$ )
- Poisson-Gamma ( $Y \sim \text{Poi}(\theta)$  data  $\theta \sim \text{Gamma}$ )
- Normal-Inverse Gamma ( $Y \sim \text{Normal}(\theta, \sigma)$  data with  $\theta \sim \text{Normal}$  and  $\sigma \sim \text{Inverse Gamma}$ )

Throughout this section,  $N$  is used for the observed data sample size, and  $S$  is the size of the desired random sample of predictions. The predicted future result is indicated by surmounting the variable name with a tilde. For example, when the observed data is represented by  $y$ , the prediction is designated  $\tilde{y}$ . Parameter names are clearly indicated as they appear.

#### 3.1 Prediction of Future Successes: Beta-Binomial (Geisser p. 73)

##### 3.1.1 Derivation

Let  $Y_1, \dots, Y_N$  be exchangeable binary variables, and therefore conditionally i.i.d. with respect to some parameter  $\theta$ . That is,  $\Pr(Y_i = 1) = \theta$ , with  $Y_i = 1$  indicating success and  $Y_i = 0$  indicating failure. The number of observed successes can be represented by  $T = \sum Y_i$ , which is sufficient for  $\theta$  and has a binomial( $N, \theta$ ) distribution. That is,

$$\Pr(T = t|\theta) = \binom{N}{t} \theta^t (1 - \theta)^{N-t}.$$

Assuming  $\theta \sim \text{Beta}(\alpha, \beta)$ , we have prior distribution

$$\pi(\theta) = \frac{\Gamma(\alpha + \beta) \theta^{\alpha-1} (1 - \theta)^{\beta-1}}{\Gamma(\alpha) \Gamma(\beta)}.$$

The posterior distribution of  $\theta$  given  $Y_1, \dots, Y_N$ , then, is

$$\begin{aligned}
p(\theta|Y_1, \dots, Y_N) &= p(\theta|t) = \frac{p(t|\theta)\pi(\theta)}{\int p(t|\theta)\pi(\theta)d\theta} \\
&= \frac{\binom{N}{t}\theta^t(1-\theta)^{N-t}\frac{\Gamma(\alpha+\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}}{\int \binom{N}{t}\theta^t(1-\theta)^{N-t}\frac{\Gamma(\alpha+\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}d\theta} \\
&= \frac{\theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}}{\int \theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}d\theta} \\
&= \frac{\Gamma(N+\alpha+\beta)}{\Gamma(t+\alpha)\Gamma(N-t+\beta)}\theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1} \\
&= \text{Beta}(t+\alpha, N-t+\beta)
\end{aligned}$$

Note that the ratio of Gamma functions in the final step appears as a scaling constant that enables the  $\text{Beta}(t+\alpha, N-t+\beta)$  density function under the integrand in the denominator of the previous step to resolve to 1.

We want to predict the number  $\tilde{T}$  of successes in a future experiment involving  $M$  binary observations. That is,  $\tilde{T} = \sum_{i=1}^M Y_{N+i}$ , and we have Beta-Binomial predictive distribution

$$\begin{aligned}
\Pr[\tilde{T} = \tilde{t}|T = t] &= \int p(\tilde{T} = \tilde{t}|\theta)p(\theta|t)d\theta \\
&= \int \binom{M}{\tilde{t}}\theta^{\tilde{t}}(1-\theta)^{M-\tilde{t}}\frac{\Gamma(N+\alpha+\beta)}{\Gamma(t+\alpha)\Gamma(N-t+\beta)}\theta^{t+\alpha-1}(1-\theta)^{N-t+\beta-1}d\theta \\
&= \frac{M!}{\tilde{t}!(M-\tilde{t})!}\frac{\Gamma(N+\alpha+\beta)}{\Gamma(t+\alpha)\Gamma(N-t+\beta)}\int \theta^{\tilde{t}+t+\alpha-1}(1-\theta)^{M+N-\tilde{t}-t+\beta-1}d\theta \\
&= \frac{\Gamma(M+1)\Gamma(N+\alpha+\beta)\Gamma(\tilde{t}+t+\alpha)\Gamma(M+N-\tilde{t}-t+\beta)}{\Gamma(\tilde{t}+1)\Gamma(M-\tilde{t}+1)\Gamma(t+\alpha)\Gamma(N-t+\beta)\Gamma(M+N+\alpha+\beta)}, \quad (2)
\end{aligned}$$

an impressive combination of Gamma functions. Note that the last two factors in the numerator together with the final factor in the denominator comprise the reciprocal of the scale factor corresponding with the  $\text{Beta}(\tilde{t}+t+\alpha, M+N-\tilde{t}-t+\beta)$  kernel in the integrand, enabling the integral to resolve to 1.

### 3.1.2 R Implementation (Beta-Binomial)

This result has been used to create R functions `dpredBB()`, `ppredBB()`, and `rpredBB()` for the Beta-Binomial predictive distribtuion for density, cumulative probability, and random

sampling, respectively (see appendix for the R code). The density function `dpredBB()` relies on the R function `lgamma()` to evaluate the numerator and denominator of the predictive distribution (2) factor by factor logarithmically, and then exponentiates for the final result, evaluated at the desired integer value (or vector of values)  $\tilde{y}$  input by the user. The cdf `ppredBB()` simply calls `dpredBB()` and returns the cumulative sum of that discrete set of results. The random sampler `rpredBB()` makes use of the inverse transform method and the output from the cdf `ppredBB()`. Calls to these functions appear as follows:

```
dpredBB(tpred,N,t,M,a,b)
ppredBB(tpred,N,t,M,a,b)
rpredBB(S,N,t,M,a,b)
```

where

$\mathbf{tpred} = \tilde{t}$ , the number (or vector of numbers) of successes in a future experiment involving  $M$  binary observations

$N$  = the number of observations in the current data set

$\mathbf{t}$  = the number of successes out of the  $N$  observations in the current data set

$M$  = the number of binary observations in a future experiment

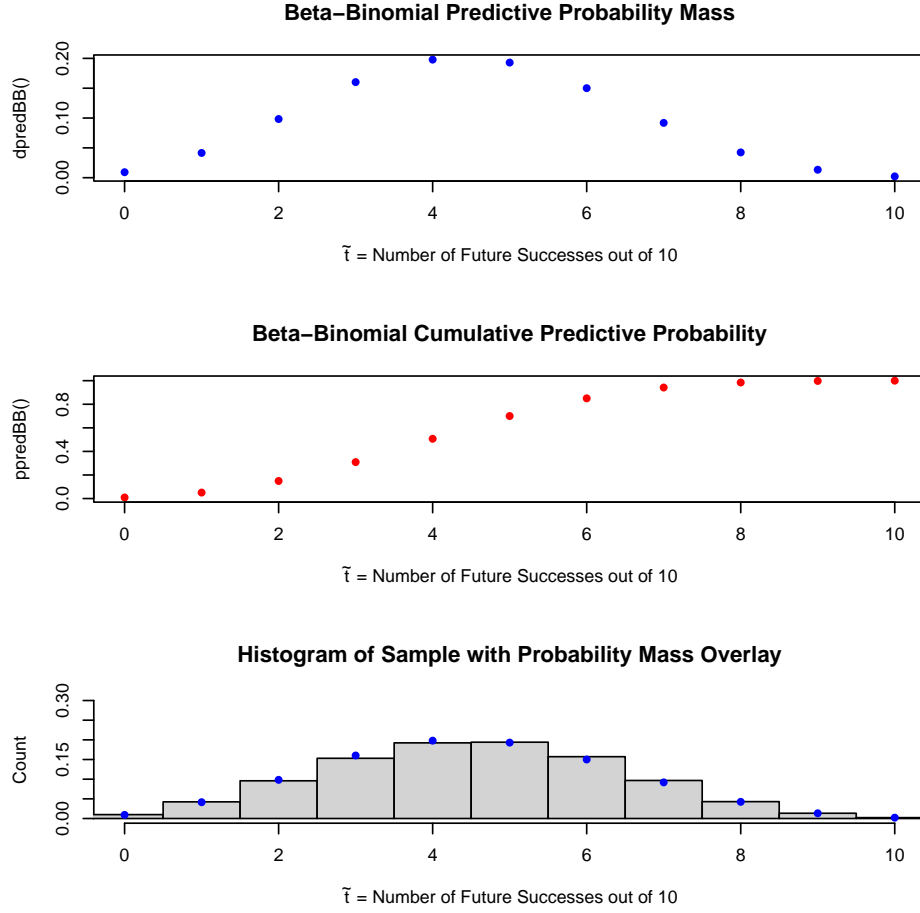
$\mathbf{a} = \alpha$ , the first shape parameter of the Beta prior

$\mathbf{b} = \beta$ , the second shape parameter of the Beta prior

$\mathbf{S}$  = the desired random sample size; that is, the number of future experiments of size  $M$  for which a predicted number of successes is desired

### 3.1.3 Example

The functions `dpredBB()` and `rpredBB()` were used in the Pass The Pigs<sup>®</sup> example in the introduction. Here, all three functions are exercised in an imagined experiment using the Big Pigs<sup>™</sup> variant. Suppose  $t = 7$  Razorbacks have been observed out of  $n = 10$  tosses of Big Pigs<sup>™</sup>, and the researcher has settled on prior distribution  $\text{Pr}(\text{Razorback}) = \theta \sim \text{Beta}(2, 8)$ . For  $M = 10$  future observations, how many successes are predicted? The figures below show the predictive distribution from `dpredBB()`, the cumulative distribution from `ppredBB()`, and a histogram of random draws from `rpredBB()`.



## 3.2 Survival Time: Exponential-Gamma (Geisser p. 74)

### 3.2.1 Derivation

Suppose  $Y_1, \dots, Y_d$  represent fully observed realizations from an exponential survival time density

$$p(y|\theta) = \theta e^{-\theta y}$$

and  $Y_{d+1}, \dots, Y_N$  represent censored copies surviving beyond the experimental time limit. We assume  $Y_1, \dots, Y_d$  are exchangeable and therefore conditionally i.i.d. with respect to  $\theta$ .

The usual exponential likelihood is used for the fully observed copies, whereas for the censored copies we need  $\Pr(Y > y|\theta) = 1 - \Pr(Y \leq y|\theta) = 1 - F(y|\theta) = 1 - (1 - e^{-\theta y}) = e^{-\theta y}$ . Here  $F$  denotes the cumulative distribution function.

Thus, with  $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$ , the overall likelihood is

$$L(\theta|y) = \prod_{i=1}^d \theta e^{-\theta y_i} \prod_{i=d+1}^N e^{-\theta y_i} = \theta^d e^{-\theta N \bar{y}}$$

Assuming a  $\text{Gamma}(\delta, \gamma)$  prior for  $\theta$ ,

$$\pi(\theta) = \frac{\gamma^\delta \theta^{\delta-1} e^{-\gamma\theta}}{\Gamma(\delta)}$$

we obtain the posterior

$$\begin{aligned} p(\theta|Y_1, \dots, Y_N) &= \frac{p(Y_1, \dots, Y_N|\theta)\pi(\theta)}{\int p(Y_1, \dots, Y_N|\theta)\pi(\theta)d\theta} \\ &= \frac{\theta^d e^{-\theta N\bar{y}} \cdot \frac{\gamma^\delta \theta^{\delta-1} e^{-\gamma\theta}}{\Gamma(\delta)}}{\int \left( \theta^d e^{-\theta N\bar{y}} \cdot \frac{\gamma^\delta \theta^{\delta-1} e^{-\gamma\theta}}{\Gamma(\delta)} \right) d\theta} \\ &= \frac{(\theta^{d+\delta-1} e^{-\theta(\gamma+N\bar{y})})}{\int (\theta^{d+\delta-1} e^{-\theta(\gamma+N\bar{y})}) d\theta} \\ &= \frac{(\gamma + N\bar{y})^{d+\delta}}{\Gamma(d+\delta)} \theta^{d+\delta-1} e^{-\theta(\gamma+N\bar{y})} \\ &= \text{Gamma}(d+\delta, \gamma + N\bar{y}), \end{aligned}$$

with the  $\text{Gamma}(d+\delta, \gamma + N\bar{y})$  density in the denominator of next to last step integrating to 1.

The survival time predictive probability density then is

$$\begin{aligned} p(\tilde{y}|Y_1, \dots, Y_N) &= \int p(\tilde{y}|\theta)p(\theta|y_1, \dots, y_N)d\theta \\ &= \int \theta e^{-\theta\tilde{y}} \cdot \frac{(\gamma + N\bar{y})^{d+\delta} \theta^{d+\delta-1} e^{-\theta(\gamma+N\bar{y})}}{\Gamma(d+\delta)} d\theta \\ &= (d+\delta)(\gamma + N\bar{y})^{d+\delta} \int \frac{\theta^{(d+\delta+1)-1} e^{-\theta(\gamma+N\bar{y}+\tilde{y})}}{(d+\delta)\Gamma(d+\delta)} d\theta \\ &= \frac{(d+\delta)(\gamma + N\bar{y})^{d+\delta}}{(\gamma + N\bar{y} + \tilde{y})^{d+\delta+1}} \int \frac{(\gamma + N\bar{y} + \tilde{y})^{d+\delta+1} \theta^{(d+\delta+1)-1} e^{-\theta(\gamma+N\bar{y}+\tilde{y})}}{\Gamma(d+\delta+1)} d\theta \\ &= \frac{(d+\delta)(\gamma + N\bar{y})^{d+\delta}}{(\gamma + N\bar{y} + \tilde{y})^{d+\delta+1}}, \end{aligned} \tag{3}$$



simplifying here by constructing a  $\text{Gamma}(d + \delta + 1, \gamma + N\bar{y} + \tilde{y})$  density in the final integrand.

### 3.2.2 R Implementation (Exponential-Gamma)

This result has been used to create R functions `dpredEG()`, `ppredEG()`, and `rpredEG()` for the Gamma-Exponential distribution for density, cumulative probability, and random sampling, respectively (see appendix for R code). These functions are exercised in the following example.

The density function `dpredEG()` evaluates the numerator and denominator of the predictive density logarithmically (using the R function `log()`) and then exponentiates to produce the result. The cdf `ppredEG()` integrates the pdf at each discrete value using the R function `integrate()`. The random sampler `rpredEG()` draws posterior  $\theta|y_1, \dots, y_i \sim \text{Gamma}(d + \delta, \gamma + \sum y_i)$  and then draws predictions from  $\text{Exp}(\theta)$ . Calls to these functions appear as follows:

```
dpredEG(ypred,y,c,dt,gm)
ppredEG(ypred,y,c,dt,gm)
rpredEG(S,y,c,dt,gm)
```

where

`ypred` =  $\tilde{y}$ , the survival time(s) in future experiments for which prediction is desired

`y` = a vector of  $N$  survival times,  $d$  observed,  $N - d$  censored

`c` = an indicator vector of length  $N$ . `c[i]=1` if the corresponding data element `y[i]` is fully observed, otherwise `c[i]=0`. Note `d = sum(c)`.

`dt` =  $\delta$ , the shape parameter of the Gamma prior

`gm` =  $\gamma$ , the rate parameter of the Gamma prior

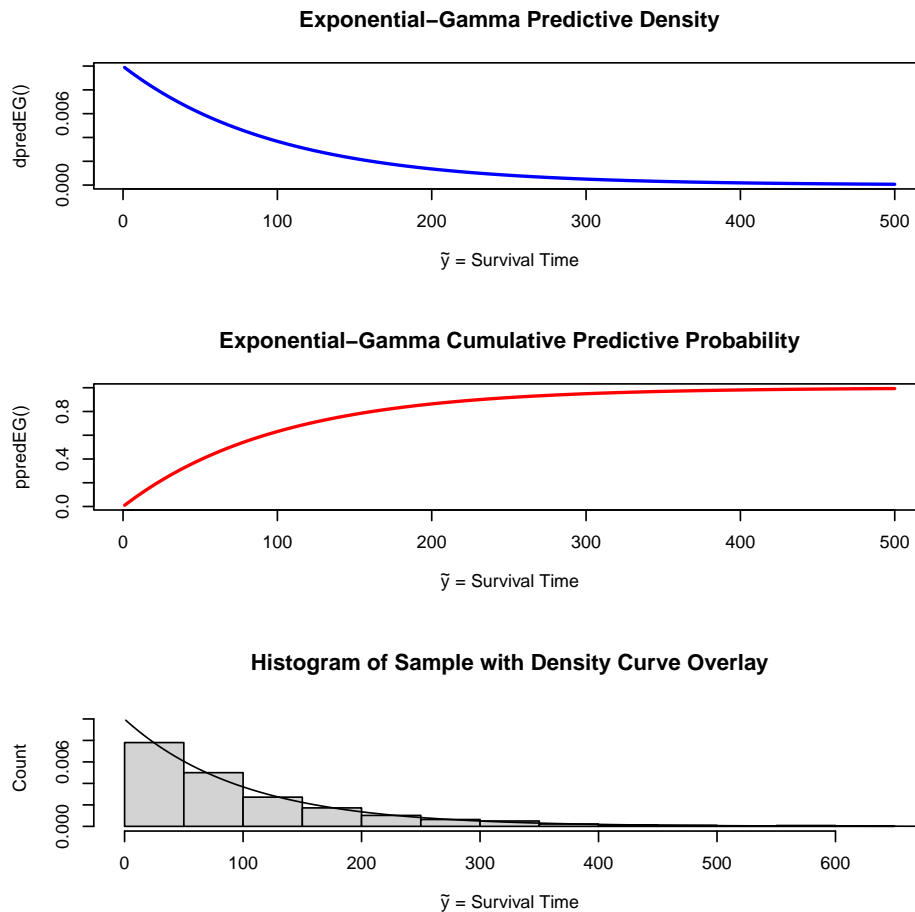
`S` = the desired random sample size

### 3.2.3 Example

**This needs updating. Also, think about the number of new observations to be predicted. Why would we (you) choose  $M=1000$ ?**

Suppose  $d = 797$  out of  $N = 1000$  copies have been observed, and the remaining 203 censored. We are interested in the number of survivors out of  $M = 1000$  future observations. For the sake of this example, survival times were generated randomly as  $y_i \sim \text{Exp}(0.01)$ , with all times exceeding an arbitrary predetermined time cutoff of 160 being set to that time. These reset times are the censored copies. We know the value of the exponential parameter ( $\theta = 0.01$ ) used to generate the data, because of the artificial way we have constructed this example. If we did not have that information, we could consider the

mean of the generated survival times (observed and censored), which was 79.9. This indicates that the expected value  $1/\theta$  of survival times must be greater than 79.9, and thus we can assume  $\theta < 1/79.9 \approx 0.0125$  with some margin. This is all to say that aiming for  $\theta = 0.01$  when choosing parameters for the prior distribution of  $\theta$  is reasonable. Since we are assuming prior  $\theta \sim \text{Gamma}(\delta, \gamma)$  we want  $\delta/\gamma \approx 0.01$ . For this example, say  $\delta = 0.5$  and  $\gamma = 50$ . The figures below illustrate the predictive probability using `dpredEG()` and `rpredEG()`, along with a histogram of a random sample taken using `rpredEG()`.



### 3.3 Poisson-Gamma Model (Hoff p. 43 - 50)

#### 3.3.1 Derivation

Suppose we have exchangeable count data observations  $Y_1, \dots, Y_N | \theta \stackrel{i.i.d.}{\sim} \text{Poisson}(\theta)$ , with  $\theta$  assumed to have prior  $\text{Gamma}(\alpha, \beta)$  distribution. That is,

$$\begin{aligned} P(Y_1 = y_1, \dots, Y_N = y_N | \theta) &= \prod_{i=1}^N p(y_i | \theta) \\ &= \prod_{i=1}^N \frac{1}{y_i!} \theta^{y_i} e^{-\theta} \\ &= \left( \prod_{i=1}^N \frac{1}{y_i!} \right) \theta^{\sum y_i} e^{-N\theta} \\ &= c(y_1, \dots, y_N) \theta^{\sum y_i} e^{-N\theta} \end{aligned}$$

and

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \text{ with } \theta, \alpha, \beta > 0.$$

Then we have posterior

$$\begin{aligned} p(\theta | y_1, \dots, y_N) &= \frac{p(y_1, \dots, y_N | \theta) \pi(\theta)}{\int_{\theta} p(y_1, \dots, y_N | \theta) p(\theta) d\theta} \\ &= \frac{p(y_1, \dots, y_N | \theta) \pi(\theta)}{p(y_1, \dots, y_N)} \\ &= \frac{1}{p(y_1, \dots, y_N)} \theta^{\sum y_i} e^{-N\theta} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \\ &= C(y_1, \dots, y_N, \alpha, \beta) \theta^{\alpha + \sum y_i - 1} e^{-(\beta + N)\theta} \\ &\propto \text{Gamma}\left(\alpha + \sum y_i, \beta + N\right). \end{aligned}$$

Here

$$\begin{aligned}
C(y_1, \dots, y_N, \alpha, \beta) &= \frac{1}{p(y_1, \dots, y_N)} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \\
&= \frac{1}{\int_{\theta} p(y_1, \dots, y_N | \theta) \pi(\theta) d\theta} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \\
&= \frac{1}{\int_{\theta} \left( \prod \frac{1}{y_i!} \right) \theta^{\sum y_i} e^{-N\theta} \left( \frac{\beta^\alpha}{\Gamma(\alpha)} \right) \theta^{\alpha-1} e^{-\beta\theta} d\theta} \cdot \left( \frac{\beta^\alpha}{\Gamma(\alpha)} \right) \\
&= \frac{1}{\left( \prod \frac{1}{y_i!} \right) \frac{\Gamma(\alpha + \sum y_i)}{(\beta + N)^{\alpha + \sum y_i}} \int_{\theta} \frac{(\beta + N)^{\alpha + \sum y_i}}{\Gamma(\alpha + \sum y_i)} \theta^{\sum y_i + \alpha - 1} e^{-(\beta + N)\theta} d\theta} \\
&= \frac{\prod_{i=1}^N y_i! (\beta + N)^{\alpha + \sum y_i}}{\Gamma(\alpha + \sum y_i)}
\end{aligned}$$

Call this constant  $C_N$  (for  $N$  observations).

Note that with an additional observation  $\tilde{y} = y_{N+1}$  the constant becomes

$$C_{N+1} = \frac{\prod_{i=1}^{N+1} y_i! (\beta + N + 1)^{\alpha + \sum_{i=1}^{N+1} y_i}}{\Gamma(\alpha + \sum_{i=1}^{N+1} y_i)}.$$

Also note that the marginal joint distribution of  $k$  observations is

$$p(y_1, \dots, y_k) = \frac{1}{C_k} \frac{\beta^\alpha}{\Gamma(\alpha)}.$$

For future observation  $\tilde{y}$ , then, we compute predictive distribution

$$\begin{aligned}
p(\tilde{y}|y_1, \dots, y_N) &= \frac{p(y_1, \dots, y_N, \tilde{y})}{p(y_1, \dots, y_N)} = \frac{p(y_1, \dots, y_{N+1})}{p(y_1, \dots, y_N)} = \frac{\frac{1}{C_{N+1}} \frac{\beta^\alpha}{\Gamma(\alpha)}}{\frac{1}{C_N} \frac{\beta^\alpha}{\Gamma(\alpha)}} = \frac{C_N}{C_{N+1}} \\
&= \frac{\frac{\prod_{i=1}^N y_i! (\beta + N)^{\alpha + \sum_{i=1}^N y_i}}{\Gamma(\alpha + \sum_{i=1}^N y_i)}}{\frac{\prod_{i=1}^{N+1} y_i! (\beta + N + 1)^{\alpha + \sum_{i=1}^{N+1} y_i}}{\Gamma(\alpha + \sum_{i=1}^{N+1} y_i)}} \\
&= \frac{\Gamma(\alpha + \sum_{i=1}^{N+1} y_i) (\beta + N)^{\alpha + \sum_{i=1}^N y_i}}{(y_{N+1}!) \Gamma(\alpha + \sum_{i=1}^N y_i) (\beta + N + 1)^{\alpha + \sum_{i=1}^{N+1} y_i}} \\
&= \frac{\Gamma(\alpha + \sum_{i=1}^N y_i + \tilde{y}) (\beta + N)^{\alpha + \sum_{i=1}^N y_i}}{(\tilde{y}!) \Gamma(\alpha + \sum_{i=1}^N y_i) (\beta + N + 1)^{\alpha + \sum_{i=1}^N y_i + \tilde{y}}} \\
&= \frac{\Gamma(\alpha + \sum y_i + \tilde{y})}{\Gamma(\tilde{y} + 1) \Gamma(\alpha + \sum y_i)} \cdot \left( \frac{\beta + N}{\beta + N + 1} \right)^{\alpha + \sum y_i} \cdot \left( \frac{1}{\beta + N + 1} \right)^{\tilde{y}} \quad (4)
\end{aligned}$$

This is a negative binomial distribution:  $\tilde{y} \sim \text{NB}(\alpha + \sum y_i, \beta + N)$

### 3.3.2 R Implementation (Poisson-Gamma)

This result has been used to create R functions `dpredPG()`, `ppredPG()`, and `rpredPG()` for the Poisson-Gamma distribution for density, cumulative probability, and random sampling, respectively (see appendix for R code). These functions are exercised in the example below.

The density function `dpredPG()` simply makes use of the R function `dnbinom()`. The cdf `ppredPG()` returns a cumulative sum of the results of `dpredPG()`. The random sampler `rpredPG()` is a bit more complicated. The difficulty is that the upper bound of the support of the predictive distribution  $p(\tilde{y}|y_1, \dots, y_n)$  is not known. Since  $p(\cdot)$  is negative binomial, we can count on it eventually decreasing toward 0 asymptotically. To establish the support, then, a method was employed to find where  $p$  comes “sufficiently close” to 0. To accomplish this a modified bisection method was devised as follows:

1. Set a desired tolerance  $\epsilon$  for the distance of the predictive distribution above zero at the upper end of its support. Currently the function leverages the local device’s numerical precision with  $\epsilon = \text{sqrt}(\text{Machine\$double.eps})$ .

2. Set lower endpoint  $L$  equal to the expected value of  $\tilde{Y}$ . That is,  $L = E(\tilde{Y}|y_1, \dots, y_N) = \frac{\alpha + \sum y_i}{\beta + N}$  (negative binomial).
3. Step to the right of  $L$  by increments in the sequence  $\{L + 2^k : k = 1, 2, \dots\}$ , setting upper endpoint  $U_{test} = L + 2^k$  using the first value of  $k$  for which  $\text{dpredPG}(L + 2^k) < 0$ .
4. Bisect  $[L, U_{test}]$  letting  $B$  be the integer nearest to the middle of the interval.
5. Test whether  $0 \leq \text{dpredPG}(B) \leq \epsilon$ . If so, accept  $U = B$  as the upper end of the support. If not:
6. Establish a new interval  $[L', U'_{test}]$  as follows:
  - (a) if  $[\text{dpredPG}(L), \text{dpredPG}(B)]$  straddles 0, then  $[L', U'_{test}] = [L, B]$ .
  - (b) if  $[\text{dpredPG}(B), \text{dpredPG}(U_{test})]$  straddles 0, then  $[L', U'_{test}] = [B, U_{test}]$ .
7. Repeat steps 3 - 5 above with the updated interval  $[L, U_{test}] = [L', U'_{test}]$  until the condition in step 5 is reached, establishing upper bound  $U$ .
8. Use  $[0, U]$  as the support upon which to draw a random sample.

Calls to these functions appear as follows:

```

dpredPG(ypred,y,alpha,beta)
ppredPG(ypred,y,alpha,beta)
rpredPG(S,y,alpha,beta)
```

where

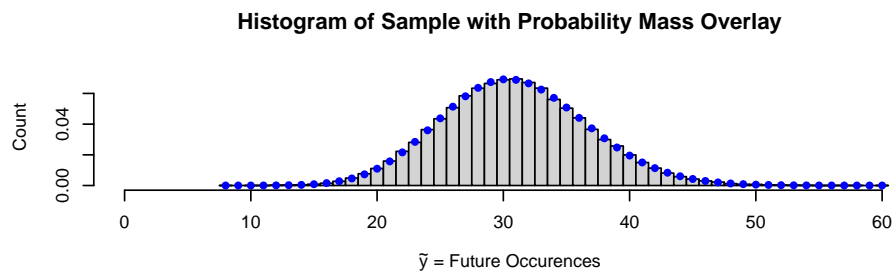
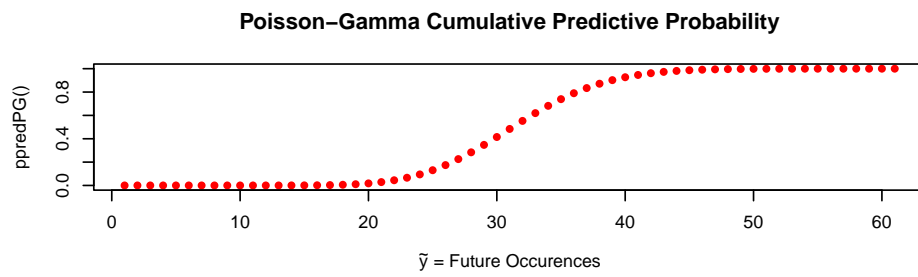
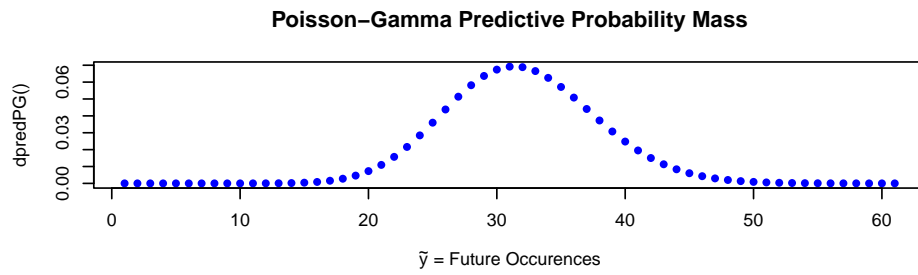
```

ypred =  $\tilde{y}$ , future count(s) for which prediction is desired
y =  $y$ , the observed counts
alpha =  $\alpha$ , the shape parameter of the Gamma prior
beta =  $\beta$ , the rate parameter of the Gamma prior
S =  $S$ , the desired random sample size
```

### 3.3.3 Example

I will come up with a different example to replace the one below.

Suppose we have 10 prior observations with counts 27, 79, 21, 100, 8, 4, 37, 15, 3, 97. Let  $\alpha = 11$  and  $\beta = 3$ . For  $\tilde{y} = 1 : 100$  possible future occurrences, the figures below show the predictive distribution from `dpredPG()`, the cumulative distribution from `ppredPG()`, and a histogram of random draws from `rpredPG()`.



### 3.4 Normal Observation with Normal-Inverse Gamma Prior

#### 3.4.1 One sample Normal-Inverse Gamma

##### 3.4.1.1 Derivation (Hoff Ch. 5)

Let  $\{Y_1, \dots, Y_N | \theta, \sigma^2\} \stackrel{i.i.d.}{\sim} \text{Normal}(\theta, \sigma^2)$  (assuming  $Y_1, \dots, Y_N$  are exchangeable). Then the joint sampling density is

$$\begin{aligned} p(y_1, \dots, y_N | \theta, \sigma^2) &= \prod_{i=1}^N p(y_i | \theta, \sigma^2) \\ &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - \theta}{\sigma}\right)^2} \\ &= (2\pi\sigma^2)^{-N/2} e^{-\frac{1}{2}\sum_{i=1}^N \left(\frac{y_i - \theta}{\sigma}\right)^2}. \end{aligned}$$

Following Hoff (p. 74-75), for joint inference on both  $\theta$  and  $\sigma$ , we assign priors

$$\frac{1}{\sigma^2} \sim \text{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$$

$$\theta | \sigma^2 \sim \text{Normal}(\mu_0, \sigma^2/\kappa_0)$$

Hoff suggests  $\sigma_0^2$  and  $\nu_0$  can be interpreted heuristically as sample variance and sample size, respectively, of prior observations, and that  $\mu_0$  and  $\kappa_0$  can be thought of as the sample mean and sample size of prior observations.

How's that  $\uparrow$ ?

From this we derive joint posterior distribution

$$\{\theta | y_1, \dots, y_N, \sigma^2\} \sim \text{Normal}(\mu_N, \sigma^2/\kappa_N)$$

$$\{\sigma^2 | y_1, \dots, y_N\} \sim \text{Inverse Gamma}(\nu_N/2, \sigma_N^2\nu_N/2).$$

where



$$\kappa_N = \kappa_0 + N$$

$$\mu_N = \frac{\kappa_0 \mu_0 + N \bar{y}}{\kappa_N}$$

$$\nu_n = \nu_0 + N$$

$$\sigma_N^2 = \frac{1}{\nu_N} \left[ \nu_0 \sigma_0^2 + (N-1)s^2 + \frac{\kappa_0 N}{\kappa_N} (\bar{y} - \mu_0)^2 \right].$$

Here  $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$  is the sample mean and  $s^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2$  is the sample variance.

From the joint posterior distribution we generate marginal samples by means of the monte carlo method (Hoff, p. 77):

$$\begin{aligned} \sigma^{2(1)} &\sim \text{Inverse Gamma}(\nu_N/2, \sigma_N^2 \nu_N/2), & \theta^{(1)} &\sim \text{Normal}(\mu_N, \sigma^{2(1)}/\kappa_N) \\ &\vdots & &\vdots \\ \sigma^{2(S)} &\sim \text{Inverse Gamma}(\nu_N/2, \sigma_N^2 \nu_N/2), & \theta^{(S)} &\sim \text{Normal}(\mu_N, \sigma^{2(S)}/\kappa_N) \end{aligned}$$

For prediction of future  $\tilde{y}|y_1, \dots, y_N, \theta, \sigma^2$ , generate  $\tilde{y}_i \sim \text{Normal}(\theta^{(i)}, \sigma^{2(i)})$ .

For prediction without the influence of any previous knowledge (Hoff p. 79), we can employ Jeffreys prior  $\tilde{p}(\theta, \sigma^2) = 1/\sigma^2$ . This leads to the same conditional distribution for  $\theta$  but a  $\text{gamma}(\frac{N-1}{2}, \frac{1}{2} \sum (y_i - \bar{y})^2)$  distribution for  $1/\sigma^2$ . This joint posterior distribution can be used to predict future  $\tilde{y}$  by first drawing  $\theta, \sigma^2$  and then simulating  $\tilde{y} \sim \text{Normal}(\theta, \sigma^2)$ . Alternatively, the joint posterior can be integrated to show that

$$\frac{\theta - \bar{y}}{s/\sqrt{N}} | y_1, \dots, y_N \sim t_{N-1}.$$

The resulting predictive distribution for  $\tilde{y}$  is a t-distribution with location  $\bar{y}$  and scale  $s\sqrt{1 + 1/N}$  and  $N - 1$  degrees of freedom (Gelman et. al. p. 66).

**3.4.1.2 R Implementation (Normal-Inverse Gamma, 1-sample)** R functions `dpredNormIG1()`, `ppredNormIG1()`, and `rpredNormIG1()` have been created for the Normal-Inverse Gamma distribution for density, cumulative probability, and random sampling, respectively (see appendix). These functions all include options for implementation with or without previous knowledge as desired. If Jeffreys prior is used, the functions simply implement R's Student's t-distribution functions `rt()`, `dt()`, and `pt()`, applying the location and scale parameters as described above. For predictions using previous knowledge, the functions work as follows: For the random sampler `rpredNormIG1()`, the monte

carlo method described above is directly employed. The predictive density and cumulative predictive density functions (`dpredNormIG1()` and `ppredNormID1()`, respectively) depend on the random sample. That is, the distributions generated by these two functions are approximated from a random sample returned by `rpredNormIG1()`.

`ppredNormIG1()` utilizes the empirical cumulative density function `ecdf()` from R's stats package. `dpredNormIG1()` utilizes a Kernel Density Estimation (KDE) method and R's built-in `density()` function. The KDE is computed by definition, using a Normal kernel:

$$\hat{f}_K(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{h} K\left(\frac{x - X_i}{h}\right),$$

where

$X_i$  is the random sample generated using `rpredNormIG1()`

$K$  is  $\text{Normal}(0, 1)$

$h$  is the bandwidth from R's `density()` function (that is, `h = density(Xi)$bw`)

Calls to these functions appear as follows:

```
dpredNormIG1(ypred,y,mu0,k0,sig20,nu0,S,Jeffreys)
ppredNormIG1(ypred,y,mu0,k0,sig20,nu0,S,Jeffreys)
rpredNormIG1(S,y,mu0,k0,sig20,nu0,Jeffreys)
```

where

`ypred` =  $\tilde{y}$ , value(s) for which prediction is desired

`y` =  $\mathbf{y}$ , the observed data

`mu0` =  $\mu_0$ , prior parameter for  $\theta$ , representing the mean of prior observations

`k0` =  $\kappa_0$ , prior parameter for  $\theta$ , representing the sample size of prior observations

`sig20` =  $\sigma_0^2$ , prior parameter for  $\sigma^2$ , representing the variance of prior observations

`nu0` =  $\nu_0$ , prior parameter for  $\sigma^2$ , representing the sample size of prior observations

`S` =  $S$ , the desired random sample size

`Jeffreys` = flag (defaulting to `FALSE`), controlling the decision to use Jeffrey's prior

These functions are exercised in the following example.

**3.4.1.3 Example** *Example (Hoff p. 72ff, using data from Grogan and Wirth (1981)): Midge wing length*

Grogan and Wirth (1981) provide 9 measurements of midge wing length, in millimeters:  $y = \{1.64, 1.7, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08\}$ . Previous studies suggest values  $\mu_0 = 1.9$  and  $\sigma_0^2 = 0.01$ . We choose  $\kappa_0 = \nu_0 = 1$  “...so that our prior distributions are only weakly centered around these estimates from other populations” (Hoff p. 76). We compute

$$\bar{y} = 1.804$$

$$\text{var}(y) = 0.0169$$

$$\kappa_N = 1 + 9 = 10$$

$$\mu_N = \frac{1 \cdot 1.9 + 9 \cdot 1.804}{10} = 1.814$$

$$\nu_N = 1 + 9 = 10$$

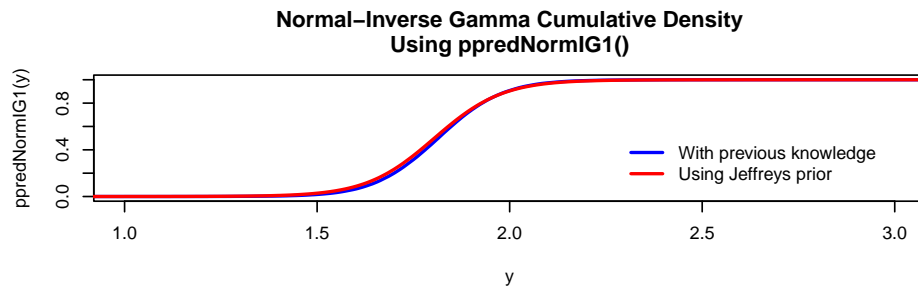
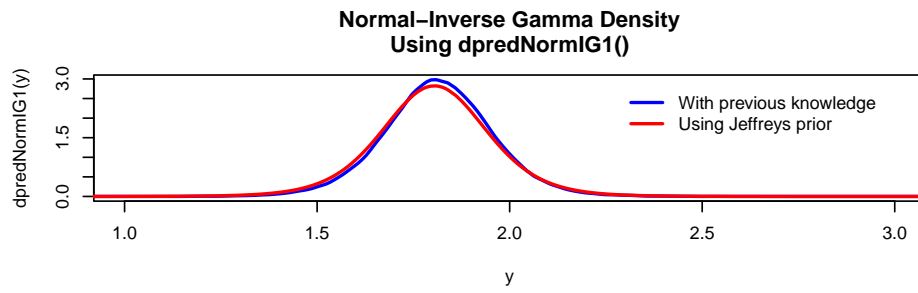
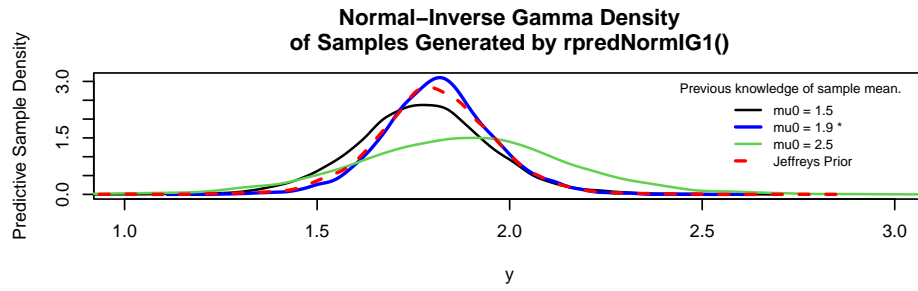
$$\sigma_N^2 = \frac{1}{10} \left[ 1 \cdot 0.01 + (9 - 1) \cdot 0.0169 + \frac{1 \cdot 9}{10} (1.804 - 1)^2 \right] = 0.0153$$

Thus  $\nu_N/2 = 5$  and  $\nu_N \sigma_N^2/2 = 0.7662$  and we have posteriors

$$\{\theta|y_1, \dots, y_N, \sigma^2\} \sim \text{Normal}(1.814, \sigma^2/10)$$

$$\{\sigma^2|y_1, \dots, y_N\} \sim \text{Inverse Gamma}(5, 0.7662)$$

The plot below illustrates the influence of previous knowledge of the population mean, and compares to the predictions resulting from Jeffreys prior.



### 3.4.2 Two-sample Normal-Inverse Gamma

**3.4.2.1 Derivation** For a Bayesian analysis comparing two exchangeable groups  $Y_{1,1}, \dots, Y_{N_1,1}$  and  $Y_{1,2}, \dots, Y_{N_2,2}$  we use the following sampling model (Hoff p. 127):

$$\begin{aligned} Y_{i,1} &= \mu + \delta + \epsilon_{i,1} \\ Y_{i,2} &= \mu - \delta + \epsilon_{i,2} \\ \{\epsilon_{i,j}\} &\sim \text{i.i.d. Normal}(0, \sigma^2). \end{aligned}$$

Letting  $\theta_1 = \mu + \delta$  and  $\theta_2 = \mu - \delta$  we see that  $\delta = (\theta_1 - \theta_2)/2$  is half the population difference in means, and  $\mu = (\theta_1 + \theta_2)/2$  is the pooled average. We'll assume conjugate prior distributions

$$\begin{aligned} \pi(\mu, \delta, \sigma^2) &= \pi(\mu) \times \pi(\delta) \times \pi(\sigma^2) \\ \mu &\sim \text{Normal}(\mu_0, \gamma_0^2) \\ \delta &\sim \text{Normal}(\delta_0, \tau_0^2) \\ \sigma^2 &\sim \text{Inverse Gamma}(\nu_0/2, \nu_0\sigma_0^2/2). \end{aligned}$$

Here  $\nu_0$  as before represents the prior sample size. The full conditional distributions follow:

$$\{\mu | \mathbf{y}_1, \mathbf{y}_2, \delta, \sigma^2\} \sim \text{Normal}(\mu_N, \gamma_N^2), \text{ where}$$

$$\mu_N = \gamma_N^2 \times \left[ \frac{\mu_0}{\gamma_0^2} + \frac{\sum_{i=1}^{N_1} (y_{i,1} - \delta) + \sum_{i=1}^{N_2} (y_{i,2} + \delta)}{\sigma^2} \right]$$

$$\gamma_N^2 = \left[ \frac{1}{\gamma_0^2} + \frac{(N_1 + N_2)}{\sigma^2} \right]^{-1}$$

$$\{\delta | \mathbf{y}_1, \mathbf{y}_2, \mu, \sigma^2\} \sim \text{Normal}(\delta_N, \tau_N^2), \text{ where}$$

$$\delta_N = \tau_N^2 \times \left[ \frac{\delta_0}{\tau_0^2} + \frac{\sum_{i=1}^{N_1} (y_{i,1} - \mu) - \sum_{i=1}^{N_2} (y_{i,2} - \mu)}{\sigma^2} \right]$$

$$\tau_N^2 = \left[ \frac{1}{\tau_0^2} + \frac{(N_1 + N_2)}{\sigma^2} \right]^{-1}$$

$$\{\sigma^2 | \mathbf{y}_1, \mathbf{y}_2, \mu, \delta\} \sim \text{Inverse Gamma}\left(\frac{\nu_N}{2}, \frac{\nu_N \sigma_N^2}{2}\right), \text{ where}$$

$$\nu_N = \nu_0 + N_1 + N_2$$

$$\nu_N \sigma_N^2 = \nu_0 \sigma_0^2 + \sum_{i=1}^{N_1} (y_{i,1} - [\mu + \delta])^2 + \sum_{i=1}^{N_2} (y_{i,2} - [\mu - \delta])^2$$

**3.4.2.2 R Implementation (Normal-Inverse Gamma, 2-sample)** The random sample generator R function `rpredNormIG2()` implements a Gibbs sampler to approximate the posterior distribution  $p(\mu, \delta, \sigma^2 | \mathbf{y}_1, \mathbf{y}_2)$ , from which to generate predictions for the two populations as follows:

1. Set initial values  $\mu = \frac{\theta_1 + \theta_2}{2}$  and  $\delta = \frac{\theta_1 - \theta_2}{2}$
2. Generate a single  $\sigma^2 | \mathbf{y}_1, \mathbf{y}_2, \mu, \delta$
3. Generate a single  $\mu | \mathbf{y}_1, \mathbf{y}_2, \delta, \sigma^2$
4. Generate a single  $\delta | \mathbf{y}_1, \mathbf{y}_2, \mu, \sigma^2$
5. Predict  $\tilde{y}_1 \sim \text{Normal}(\mu + \delta, \sigma^2)$  and  $\tilde{y}_2 \sim \text{Normal}(\mu - \delta, \sigma^2)$

The user provides the two samples  $\mathbf{y}_1$  and  $\mathbf{y}_2$  along with values for  $\mu_0, \sigma_0^2, \delta_0, \tau_0^2, \nu_0$ , and desired prediction random sample size  $S$ . The function returns  $S$  predictions for each population and the vectors of generated values for  $\mu, \delta$ , and  $\sigma^2$ . Calls to this functions appear as follows:

`rpredNormIG2(S, y1, y2, mu0, g20, d0, t20, nu0, s20)`

where

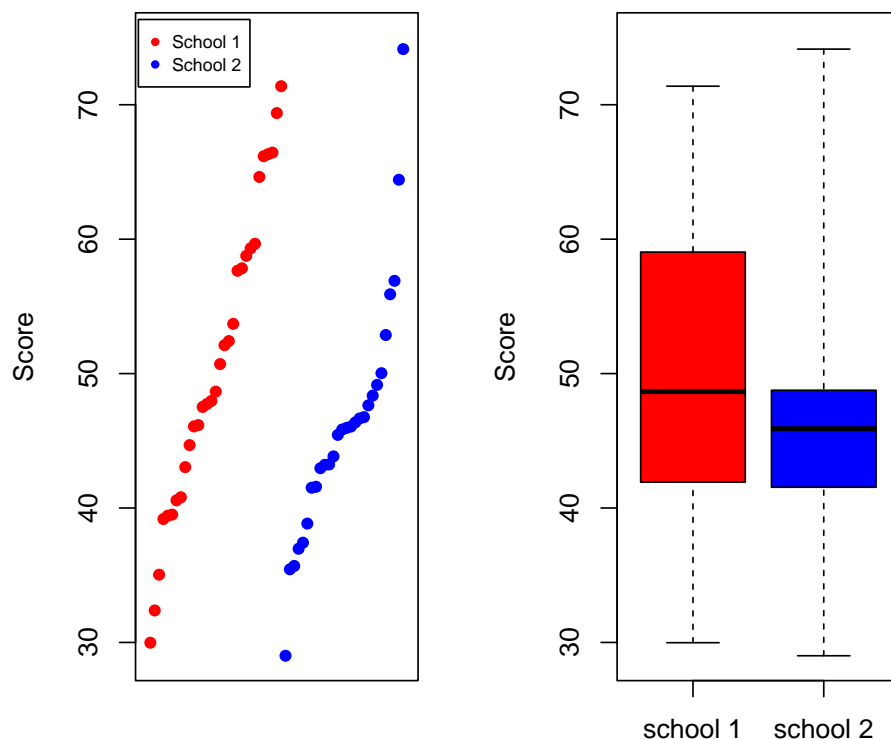
$S = S$ , the desired random sample size  
 $y1, y2 = \mathbf{y}_1, \mathbf{y}_2$ , the two sets of observed data  
 $mu0, g20 = \mu_0, \gamma_0^2$ , prior mean and variance for  $\mu$   
 $d0, t20 = \delta, \tau_0^2$ , prior mean and variance for  $\delta$   
 $nu0, s20 = \nu_0, \sigma_0^2$ , prior parameters for the variance of the Normal data error terms

**3.4.2.3 Example** (Hoff p. 128-129 *Analysis of math score data*)

Hoff provides the following example, which we reproduce here.

Math score data for two public high schools in the United States (we'll call them School 1 and School 2) were based on results of a national exam, standardized to produce a nationwide mean of 50 and a standard deviation of 10. The data consisted of 31 individual

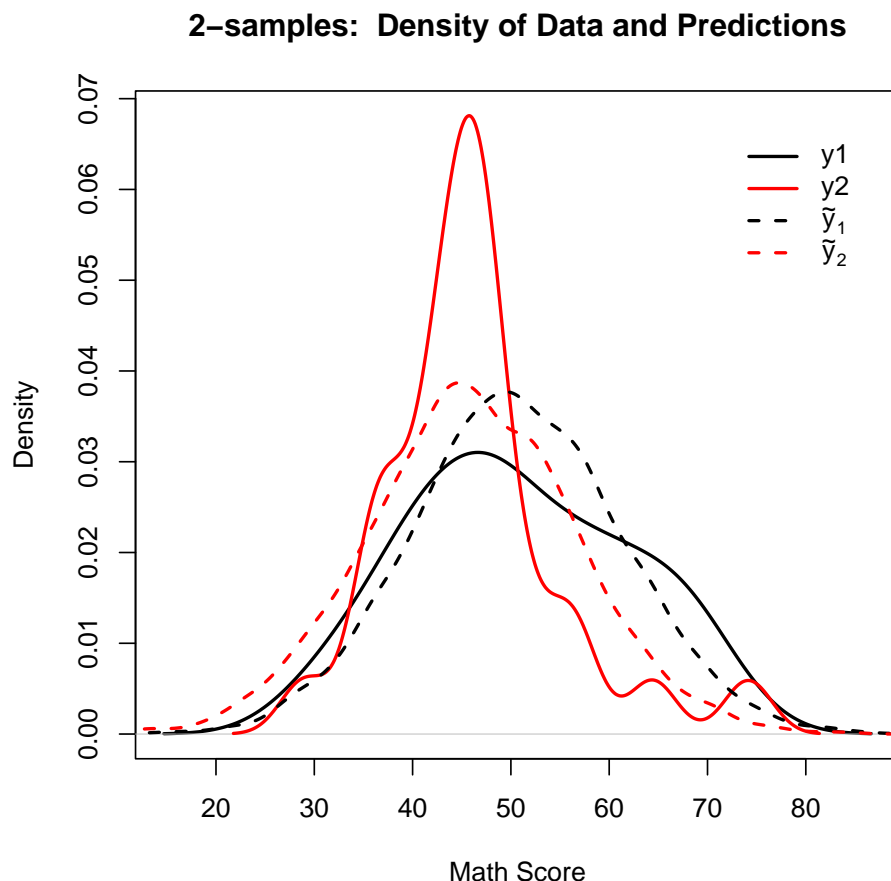
scores for School 1 and 28 for School 2. The plots below show give a visual summary of the two data sets.



The mean scores of the two samples are  $\bar{y}_1 = 50.81$  and  $\bar{y}_2 = 46.15$  for School 1 and School 2, and the standard deviations were 11.3 and 9.1, respectively. A standard t-test yields a  $p$ -value of  $p = 0.087$ , in response to which we would fail to reject the hypothesis that the population means  $\theta_1$  and  $\theta_2$  are different. Hoff points out that it is easy to imagine that re-sampling might result in a few more low scores from School 1 and a few more high scores from School 2—enough perhaps to result in a  $p$ -value below 0.05 and the opposite conclusion. Why not employ a Bayesian analysis that allows information to be shared across the groups and relies on prior knowledge of the similarities of the two groups?

We're interested in predicting a future math score by a student from each school. Unless the two schools were known in advance to be extremely exceptional, reasonable prior parameters can be based on this information. For the prior distributions of  $\mu$  and  $\sigma^2$ , we'll take  $\mu_0 = 50$  and  $\sigma_0^2 = 10^2 = 100$ , although this latter value is likely to be an overestimate of the within-school sampling variability. We'll make these prior distributions somewhat diffuse, with  $\gamma_0^2 = 25^2 = 625$  and  $\nu_0 = 1$ . For the prior distribution on  $\delta$ , choosing  $\delta_0 = 0$  represents the prior opinion that  $\theta_1 > \theta_2$  and  $\theta_2 > \theta_1$  are equally probable. Finally, since the scores are bounded between 0 and 100, half the difference between  $\theta_1$  and  $\theta_2$  must be less than 50 in absolute value, so a value of  $\tau_0^2 = 25^2 = 625$  seems reasonably diffuse.

Anything else needed here? – yes something below the plot describing the results. Mention shared variance so data from different groups influencing each other. The results of a call to `rpredNormIG2()` with the above input values for  $\mathbf{y}_1, \mathbf{y}_2, \mu_0, \sigma_0^2, \delta_0, \tau_0^2$ , and  $N$  are summarized in the following plot.



### 3.4.3 $k$ -sample Normal-Inverse Gamma: Comparing multiple groups

I expanded this paragraph to explain exchangeability in the hierarchical model:

For two-level data consisting of groups and units within groups, denote data  $\mathbf{Y}_1, \dots, \mathbf{Y}_k$  where  $\mathbf{Y}_j = \{Y_{1,j}, \dots, Y_{N_j,j}\}, j = 1, \dots, k$ . Here we stipulate exchangeability of  $Y_{1,j}, \dots, Y_{N_j,j}$ , and thus we can model the data within group  $j$  as conditionally i.i.d. given some parameter  $\phi_j$ . We also assume that  $\phi_1, \dots, \phi_k$  are exchangeable and thus conditionally i.i.d. given a parameter  $\psi$ , which has prior distribution  $\pi(\psi)$ .

We have the hierarchical Normal model (Hoff p. 132):

$$\phi_j = \{\theta_j, \sigma^2\}, p(y|\phi_j) = \text{Normal}(\theta_j, \sigma^2) \text{ (within-group model)}$$

$$\psi = \{\mu, \tau^2\}, p(\theta_j|\psi) = \text{Normal}(\mu, \tau^2) \text{ (between-groups model)}$$



We use standard semiconjugate Normal and Inverse Gamma prior distributions for the fixed but unknown parameters in the model:

$$\sigma^2 \sim \text{Inverse Gamma} \left( \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2} \right)$$

$$\tau^2 \sim \text{Inverse Gamma} \left( \frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2} \right)$$

$$\mu \sim \text{Normal} (\mu_0, \gamma_0^2)$$

**3.4.3.1 Derivation** As with the two-sample problem, joint posterior inferences for the unknown parameters can be made by constructing a Gibbs sampler to approximate the posterior distribution  $p(\theta_1, \dots, \theta_k, \mu, \tau^2, \sigma^2 | \mathbf{y}_1, \dots, \mathbf{y}_k)$ . For this we need the full conditional distribution of each parameter (Hoff pp. 134-135):

$$\{\mu | \theta_1, \dots, \theta_k, \tau^2\} \sim \text{Normal} \left( \frac{\frac{k\bar{\theta}}{\tau^2} + \frac{\mu_0}{\gamma_0^2}}{\frac{k}{\tau^2} + \frac{1}{\gamma_0^2}}, \frac{1}{\frac{k}{\tau^2} + \frac{1}{\gamma_0^2}} \right)$$

$$\{\tau^2 | \theta_1, \dots, \theta_k, \mu\} \sim \text{Inverse Gamma} \left( \frac{\eta_0 + k}{2}, \frac{\eta_0 \tau_0^2 + \sum (\theta_j - \mu)^2}{2} \right)$$

$$\{\theta_j | y_{1,j}, \dots, y_{n,j}, \sigma^2\} \sim \text{Normal} \left( \frac{\frac{n_j \bar{y}_j}{\sigma^2} + \frac{1}{\tau^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}} \right)$$

$$\{\sigma^2 | \theta_1, \dots, \theta_k, \mathbf{y}_1, \dots, \mathbf{y}_k\} \sim \text{Inverse Gamma} \left( \frac{1}{2} \left[ \nu_0 + \sum_{j=1}^k n_j \right], \frac{1}{2} \left[ \nu_0 \sigma_0^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{i,j} - \theta_j)^2 \right] \right).$$

Here  $\bar{\theta} = \frac{1}{k} \sum_{j=1}^k \theta_j$ . Note that  $\sum \sum (y_{i,j} - \theta_j)^2$  is the sum of squared residuals across all groups, conditional on the within-group means, and so the conditional distribution concentrates probability around a pooled-sample estimate of the variance.

**3.4.3.2 R Implementation (Normal-Inverse Gamma, k-samples)** The R function `rpred-NormIGk()` implements a Gibbs sampler for posterior approximation of each unknown quantity by sampling from its full conditional distribution. From these posteriors, predictions are generated, as follows:

1. Set prior parameter values:

$$\begin{aligned} &\nu_0, \sigma_0^2 \text{ for } p(\sigma^2) \\ &\eta_0, \tau_0^2 \text{ for } p(\tau^2) \\ &\mu_0, \gamma_0^2 \text{ for } p(\mu). \end{aligned}$$

2. Set initial states for the unknown parameters:

$$\begin{aligned}\theta_1^{(1)} &= \bar{\mathbf{y}}_1, \dots, \theta_k^{(1)} = \bar{\mathbf{y}}_k \\ \mu^{(1)} &= \text{mean} \left( \theta_1^{(1)}, \dots, \theta_k^{(1)} \right) \\ \tau^{2(1)} &= \text{var} \left( \theta_1^{(1)}, \dots, \theta_k^{(1)} \right) \\ \sigma^{2(1)} &= \text{mean} \left( \text{var}(\mathbf{y}_1), \dots, \text{var}(\mathbf{y}_k) \right)\end{aligned}$$

3. For  $s \in \{1, \dots, S\}$ , sample

$$\begin{aligned}\text{(a)} \quad \mu^{(s+1)} &\sim p \left( \mu | \theta_1^{(s)}, \dots, \theta_k^{(s)}, \tau^{2(s)} \right) \\ \text{(b)} \quad \tau^{2(s+1)} &\sim p \left( \tau^2 | \theta_1^{(s)}, \dots, \theta_k^{(s)}, \mu^{(s+1)} \right) \\ \text{(c)} \quad \sigma^{2(s+1)} &\sim p \left( \sigma^2 | \theta_1^{(s)}, \dots, \theta_k^{(s)}, \mathbf{y}_1, \dots, \mathbf{y}_k \right) \\ \text{(d)} \quad \theta_j^{(s+1)} &\sim p \left( \theta_j | \mu^{(s+1)}, \tau^{2(s+1)}, \sigma^{2(s+1)}, \mathbf{y}_j \right) \text{ for } j \in \{1, \dots, k\}\end{aligned}$$

4. For  $s \in \{1, \dots, S\}$ , generate prediction  $\tilde{y}_j^{(s)} \sim \text{Normal} \left( \theta_j^{(s)}, \sigma^{2(s)} \right)$

Calls to this functions appear as follows:

$$\text{rpredNormIGk}(\mathbf{S}, \mathbf{Y}, \text{nu0}, \text{s20}, \text{eta0}, \text{t20}, \text{mu0}, \text{g20})$$

where

$\mathbf{S} = S$ , the number of predictions desired, where a single prediction has  $k$  elements, one corresponding to each data group  $\mathbf{Y}_1, \dots, \mathbf{Y}_k$ . That is,  $\tilde{\mathbf{Y}}_j = \tilde{Y}_{1,j}, \dots, \tilde{Y}_{s,j}$ , for  $j \in \{1, \dots, k\}$ ,  $s \in \{1, \dots, S\}$ .

$\mathbf{Y}$  = the  $k$  sets of observed data (one column of data and one column of group indices)

$\text{nu0} = \nu_0$ , prior parameter for  $\sigma^2$  representing prior sample size

$\text{s20} =$ , prior parameter for  $\sigma^2$  representing within-sample variance

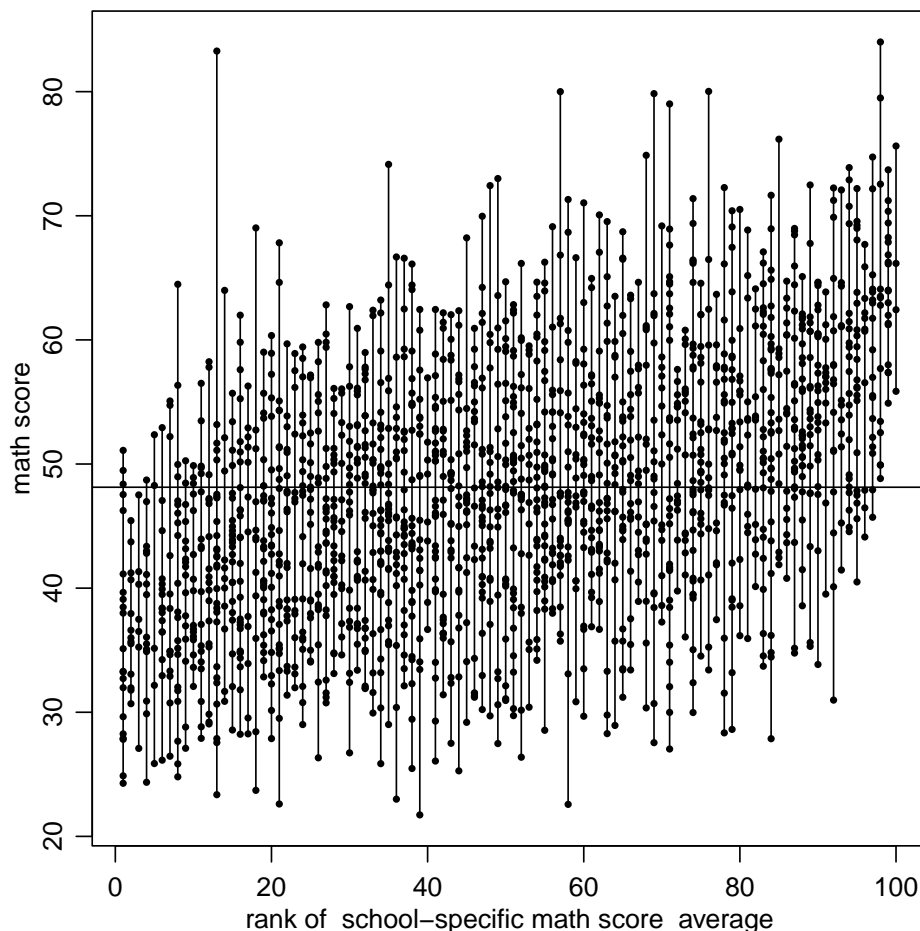
$\text{eta0} = \eta_0$ , prior parameter for  $\tau^2$  representing prior sample size

$\text{t20} = \tau_0^2$ , prior parameter for  $\tau^2$  representing between-sample variance

$\text{mu0} = \mu_0$ , prior parameter for  $\mu$  representing the mean of the pooled average from prior knowledge

$\text{g20} = \gamma_0^2$ , prior parameter for  $\mu$  representing the variance of the pooled average from prior knowledge

**3.4.3.3 Example** Returning to the math scores example, data for 10th-grade students from 100 large urban schools (each having 10th-grade enrollment of at least 400) is summarized in the following plot. Each school's results are represented by a vertical segment, with points plotted at the individual students' math scores. The horizontal line is drawn at the grand mean of the individual school means.



For prediction, we'll use the following prior values (Hoff p. 137):

- $\sigma_0^2$  : 100 (within-school variance)
- $\nu_0$  : 1 (prior sample size)
- $\tau_0^2$  : 100 (between-school variance)
- $\eta_0$  : 1 (prior sample size)
- $\mu_0$  : 50 (prior mean of school means)
- $\gamma_0^2$  : 25 (prior variance of school means)

In the example below the observed test score data from three individual schools are compared with their predictions. The schools chosen were numbers 5, 67, and 92 from the study, which had the minimum average math score, maximum average math score, and closest to the overall average math score, respectively.

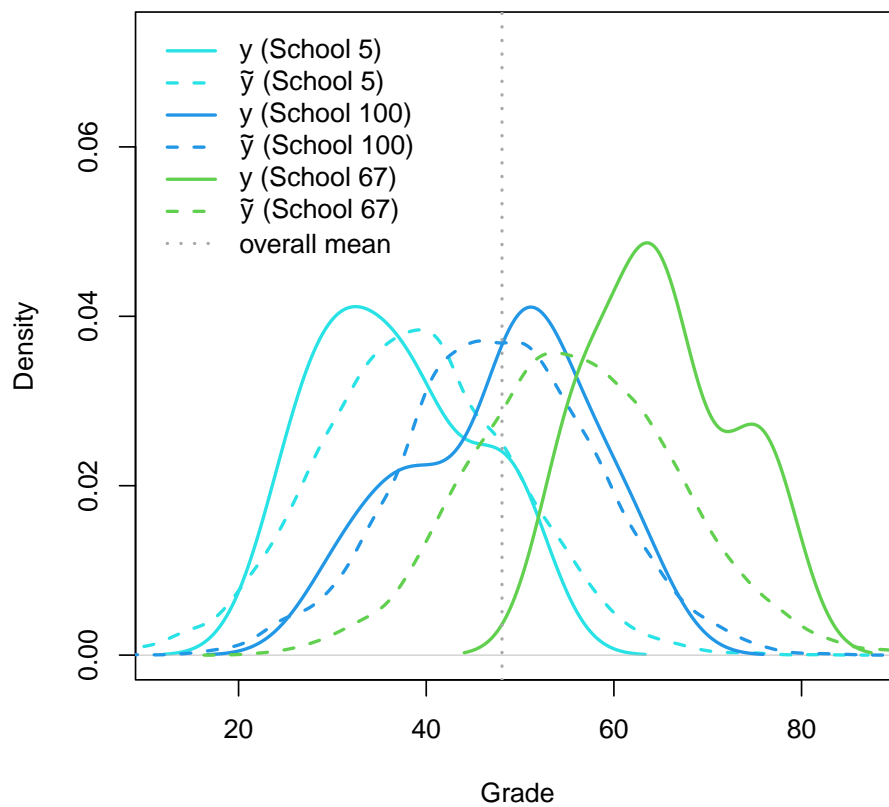
	school	average
min average	5	36.58
max average	67	65.02
grand mean	–	48.13
closest to overall	92	48.18

The plot below shows the density curves of the observed math scores for these three schools and the predictions of future math scores. The observed data is shown using solid lines, and the predictive densities are displayed with dashed lines. The predictions are “pulled” toward the overall mean, which is indicated on the plot with the dotted gray vertical line.

As before, what is being predicted?

See expansion of first sentence in above paragraph. Does that suffice?

**Densities of School Data and Predictions**



## 4 Normal Regression

Starting with observations  $Y_1, \dots, Y_N$  and explanatory variables  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , with  $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,p}\}$  where  $Y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i$  and  $\epsilon_1, \dots, \epsilon_n \stackrel{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma^2)$ , we have joint probability density

$$\begin{aligned} p(y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^n p(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 \right\}. \end{aligned} \quad (5)$$

If we let  $\mathbf{y} = (y_1, \dots, y_n)^T$  and let  $\mathbf{X}$  be the  $n \times p$  matrix whose  $i$ th row is  $\mathbf{x}_i$ , then we can express this joint probability in terms of the Multivariate Normal distribution. The Normal regression model is

$$\{\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2\} \sim \text{Multivariate Normal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}),$$

where  $\mathbf{I}$  is the  $p \times p$  identity matrix and

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_1 x_{1,1} + \dots + \beta_p x_{1,p} \\ \vdots \\ \beta_1 x_{n,1} + \dots + \beta_p x_{n,p} \end{pmatrix} = \begin{pmatrix} E[Y_1 | \boldsymbol{\beta}, \mathbf{x}_1] \\ \vdots \\ E[Y_n | \boldsymbol{\beta}, \mathbf{x}_n] \end{pmatrix}$$

The density (5) depends on  $\boldsymbol{\beta}$  through the residuals  $(y_i - \boldsymbol{\beta}^T \mathbf{x}_i)$ . We compute the ordinary least squares estimates

$$\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

and

$$\hat{\sigma}_{ols}^2 = \frac{SSR(\hat{\boldsymbol{\beta}}_{ols})}{(n-p)} = \frac{\sum (y_i - \hat{\boldsymbol{\beta}}_{ols}^T \mathbf{x}_i)^2}{(n-p)}.$$

### 4.1 Least Squares Estimation Example (Hoff p. 149ff.)

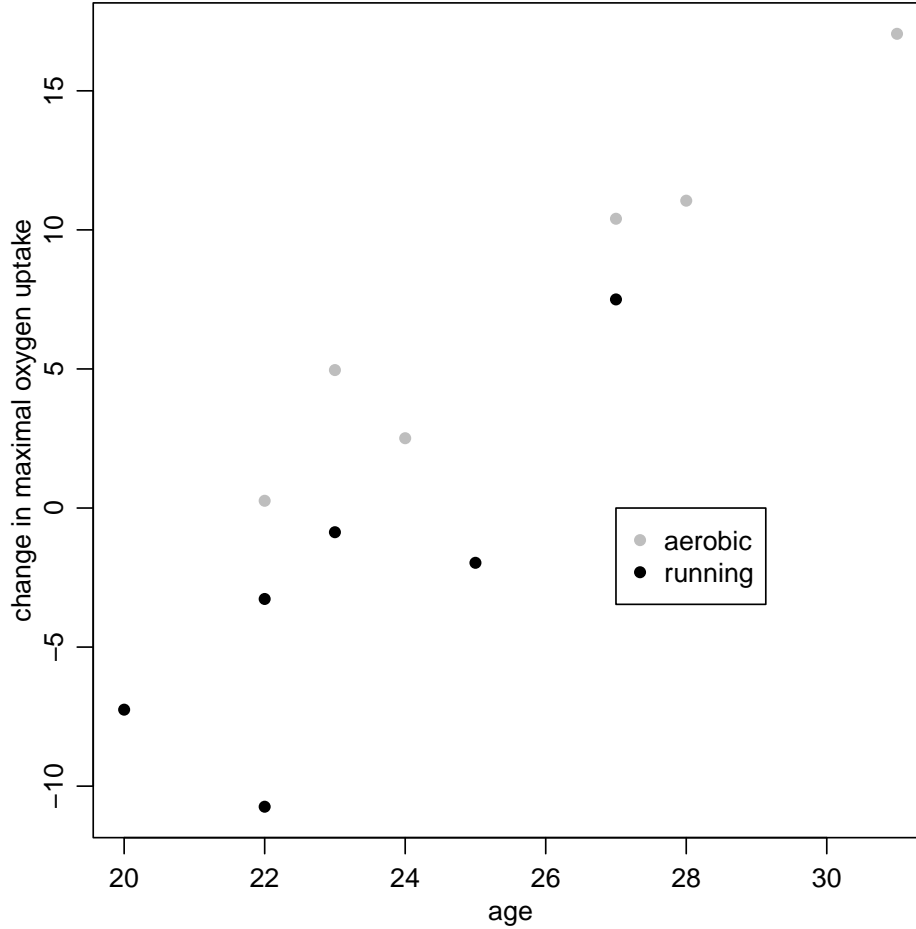
Your example 4.1 is much better described than the other examples. You should try to bring the others up to this standard.

Here we reproduce the example provided by Hoff, which will be used to illustrate Bayesian prediction for a regression model.

*Example: Oxygen uptake (from Kuehl (2000), Hoff p. 149ff)*

Twelve healthy men who did not exercise regularly were recruited to take part in a study of the effects of two different exercise regimens on oxygen uptake. Six of the twelve men were randomly assigned to a 12-week flat-terrain running program, and the remaining six were assigned to a 12-week step aerobics program. The maximum oxygen uptake of each

subject was measured (in liters per minute) while running on an inclined treadmill, both before and after the 12-week program. Of interest is how a subject's change in maximal oxygen uptake may depend on which program they were assigned to. However, other factors, such as age, are expected to affect the change in maximal uptake as well. The results are shown here:



Hoff's regression model:

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i, \text{ where} \quad (6)$$

$x_{i,1} = 1$  for each subject  $i$   
 $x_{i,2} = 0$  if subject  $i$  is on the running program, 1 if on aerobic  
 $x_{i,3} = \text{age of subject } i$   
 $x_{i,4} = x_{i,2} \times x_{i,3}$

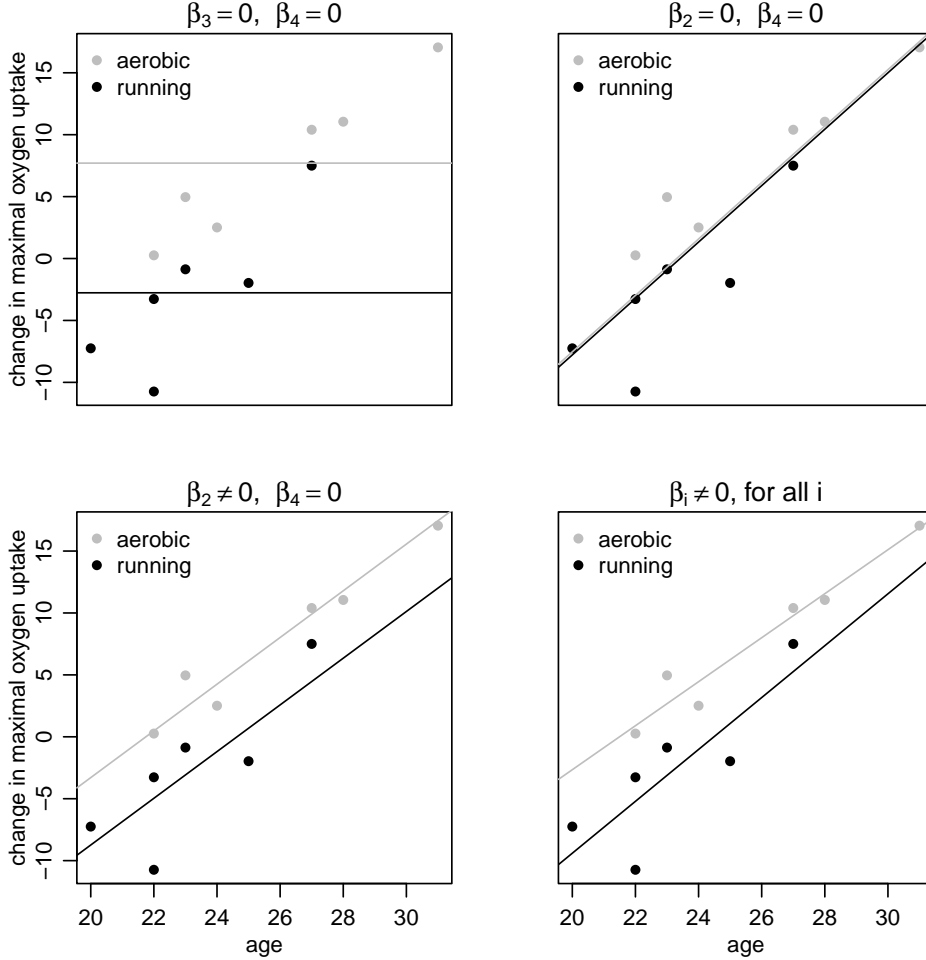
Under this model the conditional expectations of  $Y$  for the two different levels of  $x_{i,2}$  are

$$E[Y|\mathbf{x}] = \beta_1 + \beta_3 \times (\text{age}) \text{ if } x_{i,2} = 0 \text{ (running program), and}$$

$$E[Y|\mathbf{x}] = (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times (\text{age}) \text{ if } x_{i,2} = 1 \text{ (aerobic program)}$$

In other words, the model assumes that the relationship is linear in age for both exercise groups, with the difference in intercepts given by  $\beta_2$  and the difference in slopes given by

$\beta_4$ . If we assumed that  $\beta_2 = \beta_4 = 0$ , then we would have identical lines for both groups. If we assumed  $\beta_2 \neq 0$  and  $\beta_4 = 0$  then we would have a different line for each group but they would be parallel. Allowing all coefficients to be non-zero gives us two unrelated lines. Some different possibilities are depicted graphically below:



Let's find the least squares regression estimates for the model (6), and use the results to evaluate the differences between the two exercise groups. The ages of the 12 subjects, along with their observed changes in maximal oxygen uptake, are

$$\mathbf{x}_3 = (23, 22, 22, 25, 27, 20, 31, 23, 27, 28, 22, 24)$$

$$\mathbf{y} = (-0.87, -10.74, -3.27, -1.97, 7.50, -7.25, 17.05, 4.96, 10.40, 11.05, 0.26, 2.51),$$

with the first six elements of each vector corresponding to the subjects in the running group and the latter six corresponding to subjects in the aerobics group. After constructing the  $12 \times 4$  matrix  $\mathbf{X} = (\mathbf{x}_1 \mathbf{x}_2 \mathbf{x}_3 \mathbf{x}_4)$ , the matrices  $\mathbf{X}^T \mathbf{X}$  and  $\mathbf{X}^T \mathbf{y}$  can be computed, from which we get  $\boldsymbol{\beta}_{ols} = (-51.29, 13.11, 2.09, -0.32)^T$ :

This means that the estimated linear relationship between uptake and age has an intercept and slope of -51.29 and 2.09 for the running group, and  $-51.29 + 13.11 = -38.18$  and  $2.09 - 0.32 = 1.77$  for the aerobics group. These two lines are plotted in the fourth panel of

Figure XX. We obtain unbiased estimate  $\sigma^2 = SSR(\hat{\beta}_{ols})/(n - p) = 8.54$ , and use this to compute the standard error of the components of  $\hat{\beta}_{ols}$ , which are 12.25, 15.76, 0.53, and 0.65, respectively. Comparing the values of  $\hat{\beta}_{ols}$  to their standard errors suggests that the evidence for differences between the two exercise regimens is not very strong.



## 4.2 Bayesian Estimation for a Regression Model (Hoff p. 154ff)

### 4.2.1 Derivation

**4.2.1.1 A semiconjugate prior distribution** Hoff proposes a semiconjugate prior distribution for  $\beta$  and  $\sigma^2$  to be used when there is information available about the parameters. The sampling density of the data is

define  $SSR(\beta)$  I think you should write out the full conditional distributions for the different parts of the updating scheme, or at least the expressions for  $V$  and  $m$ . Similar to what you did in Chapter 3. See below def of  $SSR$  and Gibbs sampler steps with  $V$  and  $m$  written out. Better?

$$p(\mathbf{y}|\mathbf{X}, \beta, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}SSR(\beta)\right\} = \exp\left\{-\frac{1}{2\sigma^2}[\mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta]\right\},$$

and for priors we choose  $\beta \sim \text{Multivariate Normal}(\beta_0, \Sigma_0)$  and  $1/\sigma^2 = \gamma \sim \text{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$ .

Here

$$\begin{aligned} SSR(\beta) &= \sum_{i=1}^n (y_i - \beta^T x_i)^2 \\ &= \sum_{i=1}^n (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta \end{aligned}$$

is the sum of squared residuals, which is minimized by the ordinary least squares estimate  $\beta_{ols} = (\mathbf{X}^T\mathbf{X}^{-1}\mathbf{X}^T\mathbf{y})$ .

Thus we have

$$\begin{aligned} p(\beta|\mathbf{y}, \mathbf{X}, \sigma^2) &\propto p(\mathbf{y}|\mathbf{X}, \beta, \sigma^2) \times p(\beta) \\ &\propto \exp\left\{-\frac{1}{2}(-2\beta^T\mathbf{X}^T\mathbf{y}/\sigma^2 + \beta^T\mathbf{X}^T\mathbf{X}\beta/\sigma^2) - \frac{1}{2}(-2\beta^T\Sigma_0^{-1}\beta_0 + \beta^T\Sigma_0^{-1}\beta)\right\} \\ &= \exp\left\{\beta^T(\Sigma_0^{-1}\beta_0 + \mathbf{X}^T\mathbf{y}/\sigma^2) - \frac{1}{2}\beta^T(\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2)\beta\right\} \end{aligned}$$

and

$$\begin{aligned} p(\gamma|\mathbf{y}, \mathbf{X}, \beta) &\propto p(\gamma)p(\mathbf{y}|\mathbf{X}, \beta, \gamma) \\ &\propto [\gamma^{\nu_0/2-1}\exp(-\gamma \times \nu_0\sigma_0^2/2)] \times [\gamma^{n/2}\exp(-\gamma \times SSR(\beta)/2)] \\ &= \gamma^{(\nu_0+n)/2-1}\exp(-\gamma[\nu_0\sigma_0^2 + SSR(\beta)]/2), \end{aligned}$$

which we recognize as a gamma density, so that

$$\{\sigma^2|\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}\} \sim \text{Inverse Gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta})]/2).$$

Constructing a Gibbs sampler to approximate the joint posterior distribution  $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})$  is then straightforward: given current values  $\{\boldsymbol{\beta}^{(s)}, \sigma^{2(s)}\}$ , new values can be generated by

1. updating  $\boldsymbol{\beta}$ :

- (a) compute  $\mathbf{V} = \text{Var}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^{2(s)}] = (\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^{2(s)})^{-1}$  and  
 $\mathbf{m} = \text{E}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^{2(s)}] = (\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^{2(s)})^{-1} (\Sigma_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}^T\mathbf{y}/\sigma^{2(s)})$
- (b) sample  $\boldsymbol{\beta}^{(s+1)} \sim \text{Multivariate Normal}(\mathbf{m}, \mathbf{V})$

2. updating  $\sigma^2$ :

- (a) compute  $\text{SSR}(\boldsymbol{\beta}^{(s+1)})$
- (b) sample  $\sigma^{2(s+1)} \sim \text{Inverse Gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta}^{(s+1)})]/2).$

To create a sample from the predictive distribution of responses: for each  $s \in \{1, \dots, S\}$ , draw  $\epsilon^{(s)} \sim \text{Normal}(0, \sigma^{2(s)})$ . Then compute

$$y^{(s)} = \boldsymbol{\beta}^{(s)T}\mathbf{X} + \epsilon.$$

#### 4.2.1.2 Default and weakly informative prior distributions First three sentences of 4.2.1.2 need some editing.

In situations where prior information is unavailable or difficult to quantify, an alternative “default” class of prior distributions is given. Specification of the prior parameters  $(\boldsymbol{\beta}_0, \Sigma_0)$  and  $(\nu_0, \sigma_0^2)$  that represent actual prior information for a Bayesian analysis can be difficult. For a prior distribution that is not going to represent real prior information about the parameters, we choose one that is as minimally informative as possible. The resulting posterior distribution, then, will represent the posterior information of someone who began with little knowledge of the population being studied. Here we will employ Zellner’s “ $g$ -prior” (Zellner, 1986). We choose  $\boldsymbol{\beta}_0 = \mathbf{0}$  and  $\Sigma_0 = k(\mathbf{X}^T\mathbf{X})^{-1}$ ,  $k = g\sigma^2$ ,  $g > 0$ , which satisfies a desired condition that the regression parameter estimation be invariant to changes in the scale of the regressors. With this, equations ?? and ?? reduce to

$$\text{Var}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] = [\mathbf{X}^T\mathbf{X}/(g\sigma^2) + \mathbf{X}^T\mathbf{X}/\sigma^2]^{-1} = \frac{g}{g+1}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \quad (7)$$

$$\text{E}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] = [\mathbf{X}^T\mathbf{X}/(g\sigma^2) + \mathbf{X}^T\mathbf{X}/\sigma^2]^{-1}\mathbf{X}^T\mathbf{y}/\sigma^2 = \frac{g}{g+1}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \quad (8)$$

Letting

$$\mathbf{V} = \frac{g}{g+1}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \text{ and } \mathbf{m} = \frac{g}{g+1}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

we arrive at posteriors

$$\{\sigma^2|\mathbf{y}, \mathbf{X}\} \sim \text{Inverse Gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}_g]/2) \quad (9)$$

$$\{\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2\} \sim \text{Multivariate Normal}\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{ols}, \frac{g}{g+1}\sigma^2[\mathbf{X}^T\mathbf{X}]^{-1}\right). \quad (10)$$

Here  $\text{SSR}_g = \mathbf{y}^T\mathbf{y} - \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} = \mathbf{y}^T(\mathbf{I} - \frac{g}{g+1}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$ .

Simple monte carlo approximation can be used to sample from the joint posterior density  $p(\sigma^2, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$  as follows. Here  $g$  is typically set to the size of the data sample ~~number~~ of prior observations. You say "g is typically set to the number of prior observations". I don't know what you mean by "prior observations". Better? Then:

1. sample  $\sigma^2 \sim \text{Inverse Gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}_g]/2)$
2. sample  $\boldsymbol{\beta} \sim \text{Multivariate Normal}\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{ols}, \frac{g}{g+1}\sigma^2[\mathbf{X}^T\mathbf{X}]^{-1}\right)$ .

To create a sample from the predictive distribution of responses, draw  $\epsilon \sim \text{Normal}(0, \sigma^2)$ . Then for each triplet  $(\boldsymbol{\beta}, \sigma^2, \epsilon)$  we have

$$y = \boldsymbol{\beta}^T\mathbf{X} + \epsilon.$$

#### 4.2.2 R Implementation (Normal Regression)

The R function `rpredNormReg(S,Xpred,X,y,beta0,Sigma0,nu0,s20,gprior)` approximates the joint posterior density  $p(\sigma^2, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$  using one of the two methods described above, generates  $S$  triplets  $(\boldsymbol{\beta}^{(s)}, \sigma^{2(s)}, \epsilon^{(s)}) \sim \text{Normal}(0, \sigma^{2(s)})$ , and returns  $S$  predictions  $y = X_{pred}\boldsymbol{\beta}^{(s)} + \epsilon^{(s)}$ . Calls to this functions appear as follows:

`rpredNormReg(S,Xpred,X,y,beta0,Sigma0,nu0,s20,gprior)`

where

$S = S$ , the number of MCMC realizations (the desired random sample size)

$X_{pred}$  = the vector or matrix of vectors of explanatory variables upon which to base prediction

$\mathbf{X} = \mathbf{X}$ , the design matrix associated with the observed data

$\mathbf{y} = \mathbf{y}$ , the observed response vector

$\text{beta0} = \boldsymbol{\beta}_0$ , location vector for the Multivariate Normal prior distribution of  $\boldsymbol{\beta}$

$\text{Sigma0} = \boldsymbol{\Sigma}_0$ , covariance matrix for the Multivariate Normal prior distribution of  $\boldsymbol{\beta}$

$\text{nu0}, \text{s20} = \nu_0, \sigma_0^2$ , prior parameters for the Inverse-Gamma prior distribution of  $\sigma^2$

$\text{gprior}$  = a flag controlling the decision to use Zellner's location-invariant g-prior

Note: the function defaults to `gprior=TRUE`, in which case input values for `beta0`, `Sigma0`, `nu0`, and `s20` are ignored. If the user wants to employ Hoff's semi-conjugate prior as defined in section 4.2.1.1 above, all input variables must be specified, with `gprior = FALSE`.

### 4.2.3 Example

In the example below (Hoff data and code found [here](#)) to employ Hoff's semi-conjugate prior we use

$\beta_0 = \hat{\beta}_{ols} = (-51.29, -51.29, -51.29, -51.29)$  (ordinary least squares estimator of  $\beta$ )

$$\Sigma_0 = (X^T X)^{-1} \sigma^2 n = \begin{pmatrix} 1801.4 & -1801.4 & -77.02 & 77.02 \\ -1801.4 & 2981.28 & 77.02 & -122.03 \\ -77.02 & 77.02 & 3.32 & -3.32 \\ 77.02 & -122.03 & -3.32 & 5.07 \end{pmatrix} \quad (\text{sampling variance of } \hat{\beta}_{ols})$$

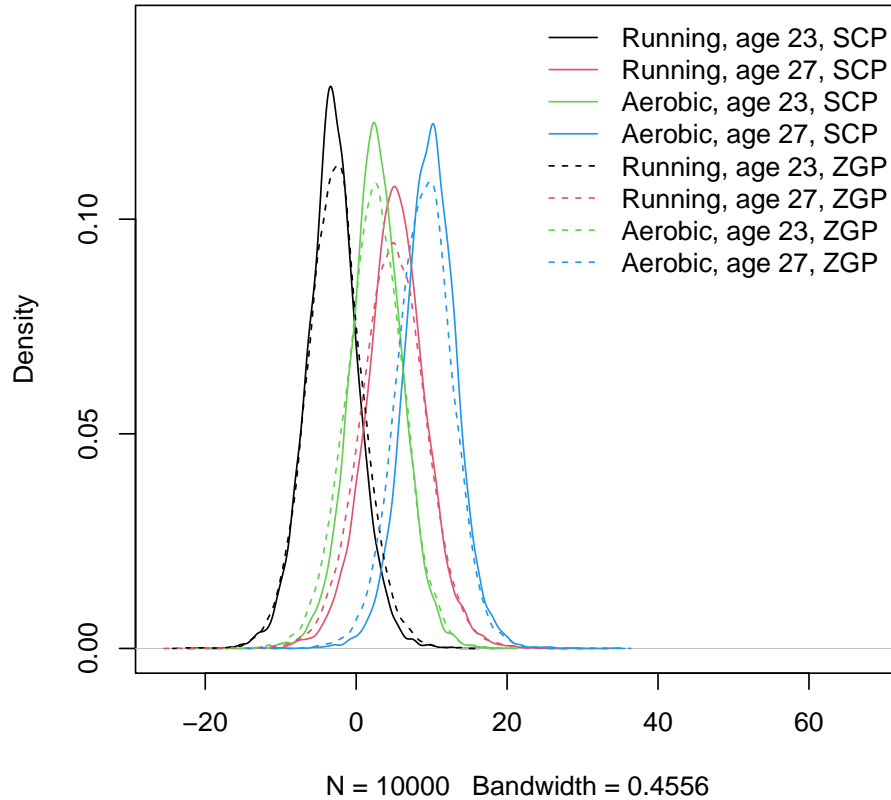
$\nu_0 = 1$  (prior sample size)

$\sigma_0^2 = \frac{\sum e_i}{n-1} = 6.21$  (variance of the residuals)

$S = 5000$  (sample size for predictive distribution random draw)

Inspection of the plot below shows the predicted distributions using Zellner's g-prior shrink toward 0, and have greater variance than those predicted using Hoff's semi-conjugate prior.

**Semi-conjugate prior vs. Zellner's g-prior**



**5 Conclusion**

**6 Appendix**

**7 References**