MATH 559: Bayesian Theory and Methods

Selection with the Normal model

Suppose that a model is to be constructed under an assumption of exchangeability with the following components:

- Data $y_1, \ldots, y_n$ recorded;
- $f_Y(y;\theta) \equiv Normal(\mu, 1)$ – here $\theta \equiv \mu$.
- $\pi_0(\mu)$ a prior density on $\mathbb{R}$.

We consider the *true*, *data-generating* scenario where the true value of the single parameter is $\mu_0 = 2$, that is, the data are drawn *independently* from $f^*(y) \equiv Normal(2, 1)$. If we specify the prior $\pi_0(\mu) \equiv Normal(\eta, 1/\lambda)$ for some fixed $\eta \in \mathbb{R}$ and $\lambda > 0$, then from `knitr 01` we have that the *posterior* distribution is $\pi_n(\mu) \equiv Normal(\eta_n, 1/\lambda_n)$, where

$$\eta_n = \frac{n\overline{y}_n + \lambda\eta}{n+\lambda} \qquad \lambda_n = n + \lambda.$$

We may similarly consider the *random* posterior $\widetilde{\pi}_n(\theta)$, a function of θ that is random because its inputs are $Y_1, \ldots, Y_n$ instead of $y_1, \ldots, y_n$; denote the (random) mean of this distribution $\widetilde{\eta}_n$, where

$$\widetilde{\eta}_n = \frac{n\overline{Y}_n + \lambda\eta}{n+\lambda}.$$

The *posterior predictive distribution* for the 'next' data point is

$$p_n(y) \equiv f_{Y_{n+1}|Y_1,\ldots,Y_n}(y|y_1,\ldots,y_n) = \int f_Y(y;\theta)\pi_n(\theta)\, d\theta$$

We may consider also the *random* version of this expression

$$\widetilde{p}_n(y) = f_{Y_{n+1}|Y_1,\ldots,Y_n}(y|Y_1,\ldots,Y_n) = \int f_Y(y;\theta)\widetilde{\pi}_n(\theta)\, d\theta$$

then the predictive distribution itself is a *random function*, as it is a function of the random variables $Y_1, \ldots, Y_n$, not the data $y_1, \ldots, y_n$. For the *predictive* distribution in the Normal problem, $p_n(y) \equiv Normal\left(\mu_{n,1}, \lambda_{n,1}^{-1}\right)$ where

$$\mu_{n,1} = \eta_n \qquad \lambda_{n,1} = \frac{\lambda_n}{1+\lambda_n} = \frac{n+\lambda}{n+1+\lambda}$$

Thus here we have $\widetilde{\pi}_n(\mu)$ and $\widetilde{p}_n(y)$ as *random functions*, specifically

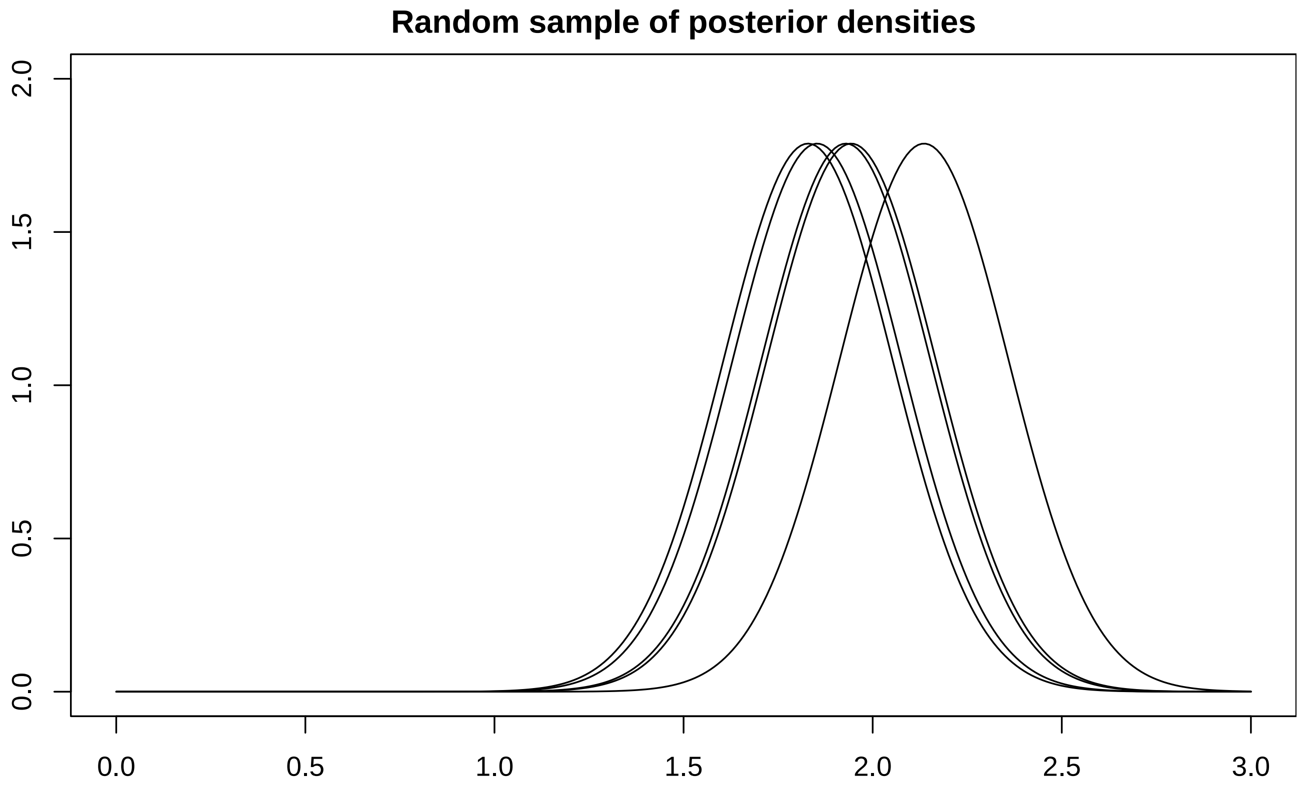
$$\widetilde{\pi}_n(\mu) \equiv Normal\left(\frac{n\overline{Y}_n + \lambda\eta}{n+\lambda}, \frac{1}{n+\lambda}\right) \qquad \widetilde{p}_n(y) \equiv Normal\left(\frac{n\overline{Y}_n + \lambda\eta}{n+\lambda}, 1 + \frac{1}{n+\lambda}\right).$$

To illustrate the random nature of these functions, we consider five replicate data sets generated from the true model $f^*(y) \equiv Normal(2, 1)$, and plot the derived posterior in each case; under this data generating model

$$\overline{Y}_n \sim Normal(2, 1/n).$$

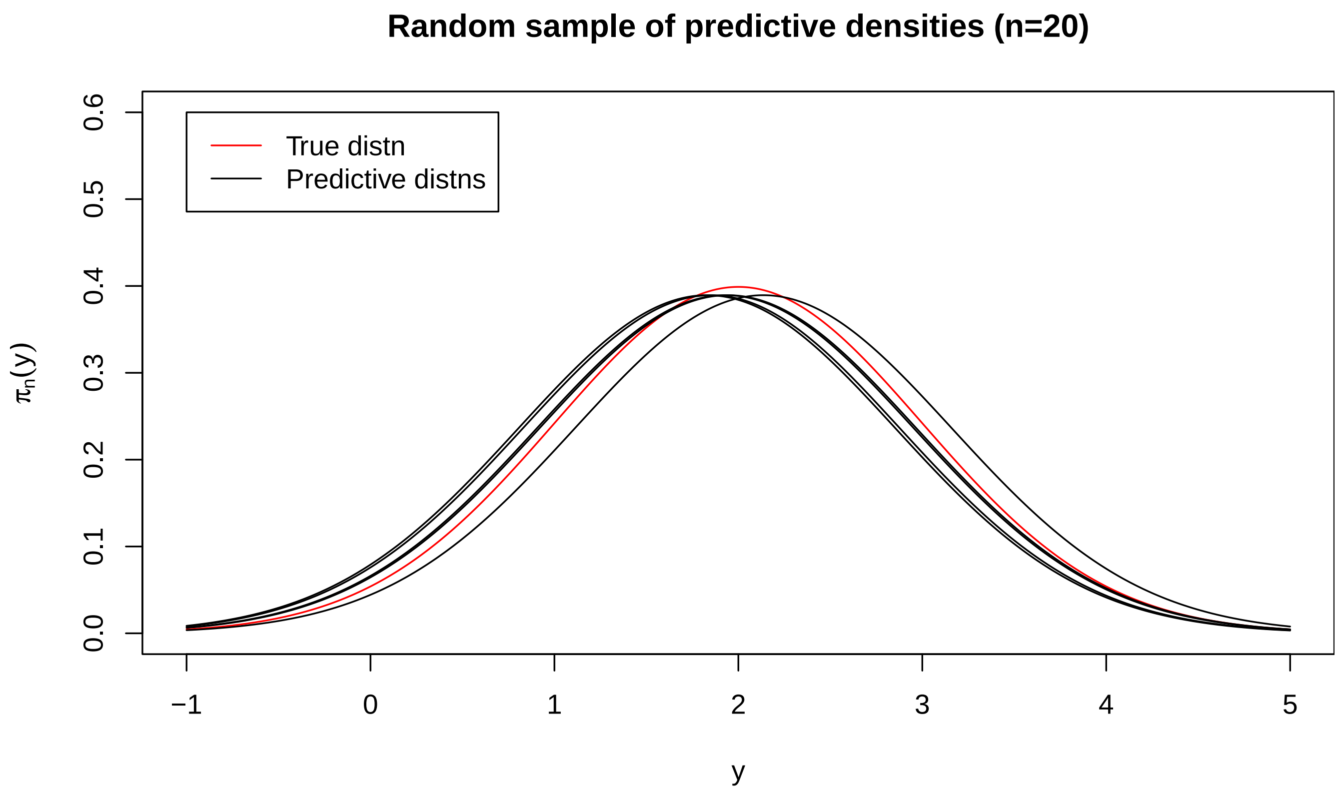
We take the prior hyperparameters to be $\eta = 0$ and $\lambda = 0.1$.

```
set.seed(2134)
n<-20;nreps<-5
mu0<-2;sigma0<-1
eta<-0; lambda<-0.1
lambda.n<-n+lambda; lambda.n1<-lambda.n/(1+lambda.n)
par(mar=c(3,3,2,0))
xv<-seq(0,3,by=0.01)
yv<-dnorm(xv,0,1)
plot(xv,yv,type='n',main='Random sample of posterior densities',ylim=range(0,2))
for(irep in 1:nreps){
    ybar<-rnorm(1,mu0,sqrt(1/n))
    eta.n<-(n*ybar+lambda*eta)/(n+lambda)
    yv<-dnorm(xv,eta.n,sqrt(1/lambda.n))
    lines(xv,yv)
}
```

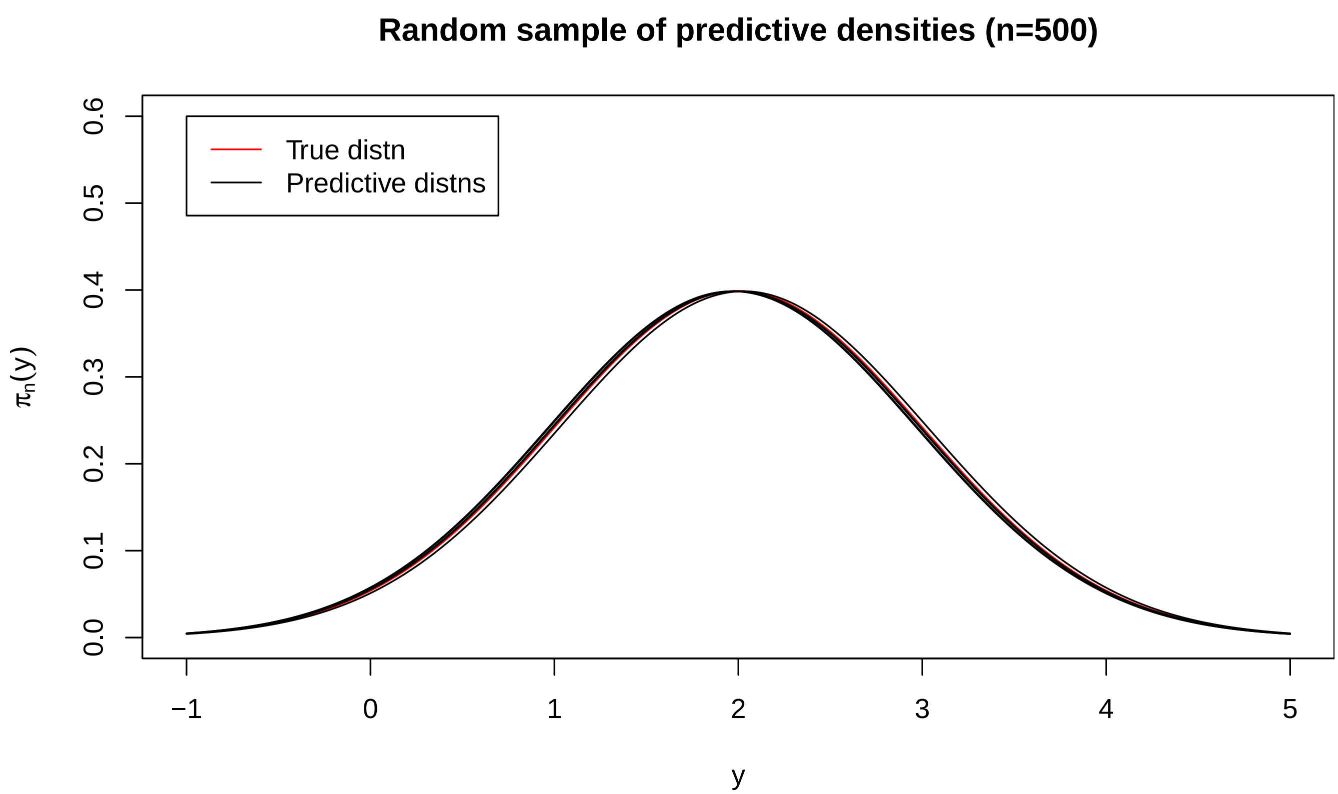


The posterior predictive distribution $p_n(y)$ can be regarded as a Bayesian estimate of the true data generating distribution $f^*(y)$. In this Normal model, and by standard arguments, as $Y_1, Y_2, \ldots$ are drawn independently from $f^*(y) \equiv Normal(2,1)$, we have that $\overline{Y}_n \xrightarrow{a.s.} 2$, and so as n increases we can see that $\widetilde{p}_n(y)$ converges (pointwise almost surely, and weakly) to $f^*(y)$.

```
xv<-seq(-1,5,by=0.01)
yv<-dnorm(xv,2,1)
par(mar=c(4,4,4,0))
plot(xv,yv,type='l',main='Random sample of predictive densities (n=20)',
    ylim=range(0,0.6),col='red',xlab='y',ylab=expression(pi[n](y)))
set.seed(2134)
for(irep in 1:nreps){
    ybar<-rnorm(1,mu0,sqrt(1/n))
    eta.n<-(n*ybar+lambda*eta)/(n+lambda)
    yv<-dnorm(xv,eta.n,sqrt(1/lambda.n1))
    lines(xv,yv)
}
legend(-1,0.6,c('True distn','Predictive distns'),col=c('red','black'),lty=1)
```



For $n = 500$, we practically recover $f^*(y)$ in each replicate.



The KL divergence between $f^*(y)$ and $p_n(y)$ is

$$KL(f^*, p_n) = \int \log\left(\frac{f^*(y)}{p_n(y)}\right) f^*(y)\, dy = \int \log(f^*(y)) f^*(y)\, dy - \int \log(p_n(y)) f^*(y)\, dy. \qquad (\diamondsuit)$$

The first term in ($\diamondsuit$) is a constant which does not depend on the inference model. The random variable version $KL(f^*, \widetilde{p}_n)$ can also be considered.

The following statistics can be used for model selection:

- **Training loss:** The *training loss*, T_n, is a measure that approximates the KL divergence based on the sample

$$T_n \equiv T(Y_1, \ldots, Y_n) = -\frac{1}{n}\sum_{i=1}^{n} \log \widetilde{p}_n(Y_i)$$

  which can be regarded as a sample-based estimator of the second term in ($\diamondsuit$), with the data drawn independently from f^*. In this form, T_n is random variable as it depends on $\widetilde{p}_n$.

  We have in the Normal case that

$$\log \widetilde{p}_n(y) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{n+\lambda+1}{n+\lambda}\right) - \frac{1}{2}\frac{n+\lambda}{n+\lambda+1}(y - \widetilde{\eta}_n)^2$$

  so therefore

$$T_n = \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\left(\frac{n+\lambda+1}{n+\lambda}\right) + \frac{1}{2}\frac{n+\lambda}{n(n+\lambda+1)}\sum_{i=1}^{n}(Y_i - \widetilde{\eta}_n)^2$$

- **Generalization loss:** The *generalization loss*, G_n, is the second term in ($\diamondsuit$):

$$G_n \equiv G(Y_1, \ldots, Y_n) = -\int \log \widetilde{p}_n(y) f^*(y)\, dy.$$

  This can only be computed precisely if $f^*(y)$ is known. In our Normal example, using the calculation above and denoting by $\phi(y)$ the standard Normal density, we have that

$$G_n = \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\left(\frac{n+\lambda+1}{n+\lambda}\right) + \frac{1}{2}\frac{n+\lambda}{n+\lambda+1}\int_{-\infty}^{\infty}(y - \widetilde{\eta}_n)^2\phi(y-2)dy.$$

  Writing

$$\begin{aligned}\int_{-\infty}^{\infty}(y - \widetilde{\eta}_n)^2\phi(y-2)dy &= \int_{-\infty}^{\infty}(y - 2 + 2 - \widetilde{\eta}_n)^2\phi(y-2)dy \\ &= \int_{-\infty}^{\infty}(y-2)^2\phi(y-2)dy + \int_{-\infty}^{\infty}(2 - \widetilde{\eta}_n)^2\phi(y-2)dy = 1 + (2 - \widetilde{\eta}_n)^2\end{aligned}$$

  we have that

$$G_n = \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\left(\frac{n+\lambda+1}{n+\lambda}\right) + \frac{1}{2}\frac{n+\lambda}{n+\lambda+1}\left(1 + (2 - \widetilde{\eta}_n)^2\right)$$

- **Entropy:** The first term in ($\diamondsuit$) is often denoted $-S$, where

$$S = -\int \log(f^*(y)) f^*(y)\, dy$$

  and is termed the *entropy* of f^*. With $f^*(y) \equiv Normal(2, 1)$, we have that

$$S = \frac{1}{2}\log(2\pi) + \frac{1}{2} \simeq 1.418939.$$

  and

$$G_n - S = \frac{1}{2}\log\left(\frac{n+\lambda+1}{n+\lambda}\right) + \frac{1}{2}\frac{n+\lambda}{n+\lambda+1}\left(1 + (2 - \widetilde{\eta}_n)^2\right) - \frac{1}{2}$$

  The quantity $G_n - S$ is termed the *generalization error*: note that $G_n \geq S$ (with probability 1) as the KL divergence is non-negative. Note that as $n \longrightarrow \infty$, $G_n \stackrel{a.s.}{\longrightarrow} S$.

- **Cross-validation loss:** The *cross-validation* loss, C_n, is defined by

$$C_n = -\frac{1}{n}\sum_{i=1}^{n}\log \widetilde{p}_n^{(-i)}(Y_i)$$

where $\widetilde{p}_n^{(-i)}(y)$ is the posterior predictive distribution derived from the random variables with Y_i omitted. From above, we have

$$C_n = \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\left(\frac{n+\lambda}{n-1+\lambda}\right) + \frac{1}{2}\frac{(n-1+\lambda)}{n(n+\lambda)}\sum_{i=1}^{n}(Y_i - \widetilde{\eta}_n^{(-i)})^2$$

where for $i = 1, \ldots, n$

$$\widetilde{\eta}_n^{(-i)} = \frac{\sum\limits_{j\neq i} Y_j + \eta\lambda}{n-1+\lambda}.$$

We have for arbitrary y that

$$\begin{aligned}
\mathbb{E}_{\widetilde{\pi}_n}\left[\frac{1}{f_Y(y;\theta)}\right] &\equiv \int_{-\infty}^{\infty}(2\pi)^{1/2}\exp\{(y-\mu)^2/2\}\widetilde{\pi}_n(\mu)\,d\mu \\
&= \int_{-\infty}^{\infty}(2\pi)^{1/2}\exp\left\{\frac{1}{2}(y-\mu)^2\right\}\left(\frac{\lambda_n}{2\pi}\right)^{1/2}\exp\left\{-\frac{\lambda_n}{2}(\mu-\widetilde{\eta}_n)^2\right\}d\mu \\
&= \lambda_n^{1/2}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}\left[\lambda_n(\mu-\widetilde{\eta}_n)^2 - (y-\mu)^2\right]\right\}.
\end{aligned}$$

Completing the square

$$\lambda_n(\mu-\widetilde{\eta}_n)^2 - (y-\mu)^2 = (\lambda_n - 1)\left(\mu - \frac{\lambda_n\widetilde{\eta}_n - y}{\lambda_n - 1}\right)^2 - \frac{\lambda_n}{\lambda_n - 1}(y-\widetilde{\eta}_n)^2$$

and so therefore computing the integral (the integrand is the kernel of a Normal density) we get

$$\mathbb{E}_{\widetilde{\pi}_n}\left[\frac{1}{f_Y(y;\theta)}\right] = (2\pi)^{1/2}\left(\frac{\lambda_n}{\lambda_n-1}\right)^{1/2}\exp\left\{\frac{1}{2}\frac{\lambda_n}{\lambda_n-1}(y-\widetilde{\eta}_n)^2\right\}$$

so therefore as $\lambda_n = n+\lambda$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\log\mathbb{E}_{\widetilde{\pi}_n}\left[\frac{1}{f_Y(Y_i;\theta)}\right] = \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\left(\frac{n+\lambda}{n-1+\lambda}\right) + \frac{1}{2}\frac{n+\lambda}{n(n-1+\lambda)}\sum_{i=1}^{n}(Y_i-\widetilde{\eta}_n)^2.$$

Now

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i-\widetilde{\eta}_n)^2 = \sum_{i=1}^{n}\left(Y_i - \frac{n\overline{Y}_n+\eta\lambda}{n+\lambda}\right)^2 &= \frac{1}{(n+\lambda)^2}\sum_{i=1}^{n}\left((n+\lambda)Y_i - \sum_{j=1}^{n}Y_j - \eta\lambda\right)^2 \\
&= \frac{1}{(n+\lambda)^2}\sum_{i=1}^{n}\left((n-1+\lambda)Y_i - \sum_{j\neq i}Y_j - \eta\lambda\right)^2 \\
&= \frac{(n-1+\lambda)^2}{(n+\lambda)^2}\sum_{i=1}^{n}\left(Y_i - \frac{\sum\limits_{j\neq i}Y_j+\eta\lambda}{n-1+\lambda}\right)^2 = \frac{(n-1+\lambda)^2}{(n+\lambda)^2}\sum_{i=1}^{n}\left(Y_i - \widetilde{\eta}_n^{(-i)}\right)^2
\end{aligned}$$

and so we have verified that

$$C_n = \frac{1}{n}\sum_{i=1}^{n}\log\mathbb{E}_{\widetilde{\pi}_n}\left[\frac{1}{f_Y(Y_i;\theta)}\right].$$

- **WAIC:** The *widely applicable information criterion* (or WAIC), W_n, is defined by

$$W_n = T_n + \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}_{\widetilde{\pi}_n}[\log f_Y(Y_i;\theta)]$$

where T_n is the training loss. It can be shown that $W_n = C_n + \mathrm{O}_p(n^{-2})$ and so W_n provides the basis of a tractable approximation strategy.

Studying the properties of W_n as a random variable is not easy, but we can compute the numerical version of this statistic. However, it is not always straightforward to compute $\mathrm{Var}_{\pi_n}[\log f_Y(y_i;\mu)]$ analytically, so instead it is often approximated by sampling the posterior distribution $\pi_n(\mu)$, and using the samples to compute the variance numerically. That is, if we sample N times from $\pi_n(\mu)$ to obtain sampled values $\mu^{(1)},\ldots,\mu^{(N)}$, we can approximate

$$\mathrm{Var}_{\pi_n}[\log f_Y(y;\mu)] \simeq \frac{1}{N}\sum_{j=1}^{N}(s(y;\mu^{(j)}) - \overline{s}(y))^2$$

where

$$s(y;\mu) = \log f_Y(y;\mu) \qquad \overline{s}(y) = \frac{1}{N}\sum_{j=1}^{N} s(y;\mu^{(j)}).$$

- **Marginal likelihood (or prior predictive):** The normalizing constant that appears in the denominator of the (random) posterior $\widetilde{\pi}_n(\theta)$ is

$$Z_n \equiv Z(Y_1,\ldots,Y_n) = \int \prod_{i=1}^{n} f_Y(Y_i;\theta)\pi_0(\theta)\,d\theta.$$

which is the value of the (random) joint pdf $f_{Y_{1:n}}(Y_{1:n}) \equiv f_{Y_1,\ldots,Y_n}(Y_1,\ldots,Y_n)$. The quantity Z_n is termed the *marginal likelihood*, or *prior predictive* distribution. Here, by the usual complete-the-square calculations

$$\begin{aligned}
f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) &= \int_{-\infty}^{\infty}\left(\frac{1}{2\pi}\right)^{n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left\{-\frac{\lambda}{2}(\mu-\eta)^2\right\}d\mu \\
&= \left(\frac{1}{2\pi}\right)^{n/2}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\overline{y}_n)^2\right\}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}\left[n(\mu-\overline{y}_n)^2+\lambda(\mu-\eta)^2\right]\right\}d\mu \\
&= \left(\frac{1}{2\pi}\right)^{n/2}\left(\frac{\lambda}{n+\lambda}\right)^{1/2}\exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}(y_i-\overline{y}_n)^2+\frac{n\lambda}{n+\lambda}(\overline{y}_n-\eta)^2\right]\right\}.
\end{aligned}$$

Therefore, recalling that $\lambda_n = n+\lambda$

$$\log Z_n = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log\lambda - \frac{1}{2}\log\lambda_n - \frac{1}{2}\left[\sum_{i=1}^{n}(y_i-\overline{y}_n)^2 + \frac{n\lambda}{\lambda_n}(\overline{y}_n-\eta)^2\right]$$

We have by definition that $p_n(y_{n+1}) = z_{n+1}/z_n$ and hence

$$\log\widetilde{p}_n(y_{n+1}) = \log z_{n+1} - \log z_n$$

Finally $F_n = -\log Z_n$ is the *free energy*. We can also report

$$\overline{F}_n = -\frac{1}{n}\log Z_n.$$

In large samples, the quantities T_n, G_n, C_n and W_n are numerically very similar, and have the same limiting value.

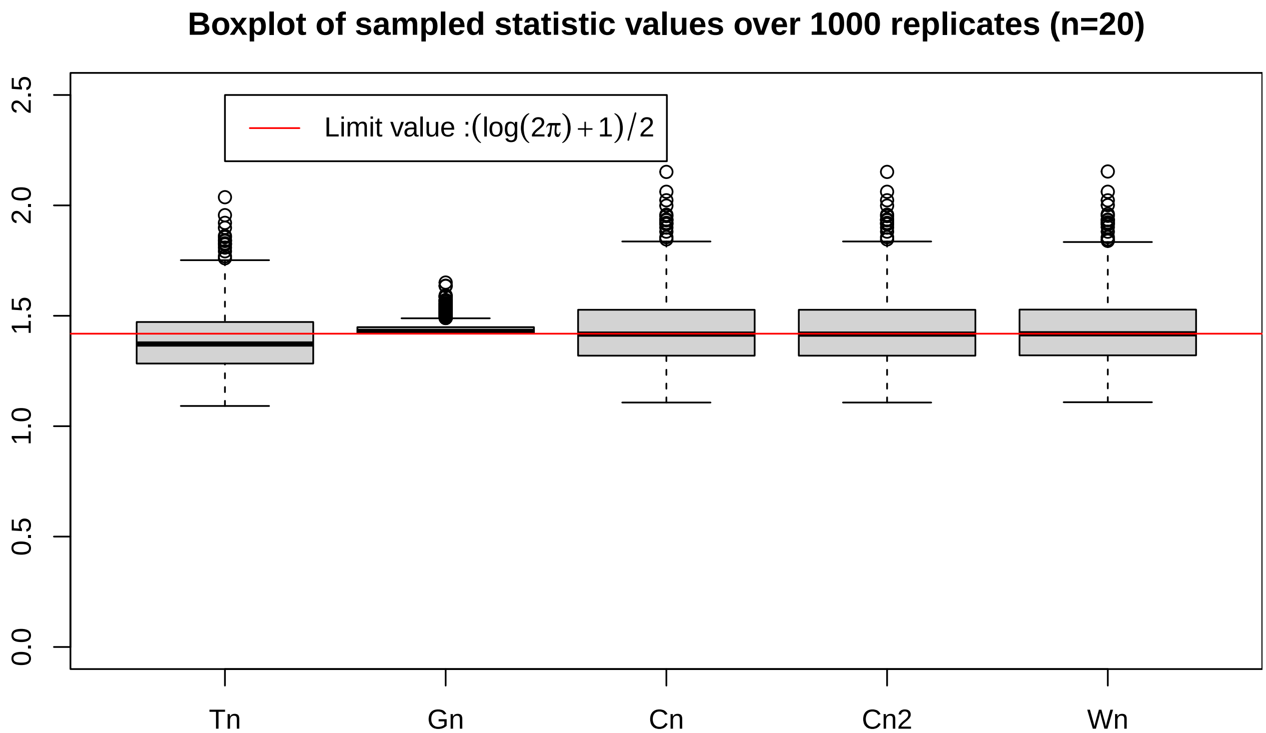
Simulation Study:

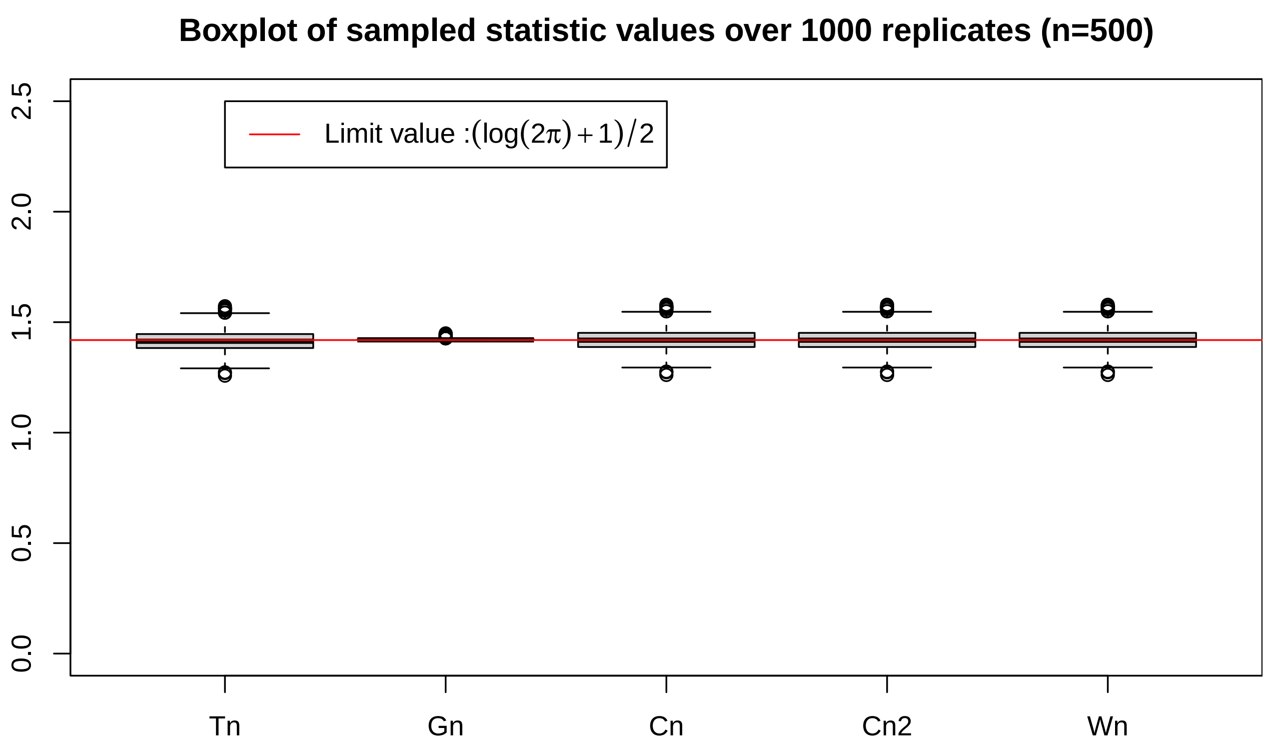
```
set.seed(2134)
n<-20;nreps<-1000
mu0<-2;sigma0<-1
eta<-0; lambda<-0.1
lambda.n<-n+lambda; lambda.n1<-lambda.n/(1+lambda.n)
lambda.ni<-n-1+lambda; lambda.ni1<-lambda.ni/(1+lambda.ni)
Y<-matrix(rnorm(n*nreps,2,1),ncol=n)
const<-0.5*log(2*pi)-0.5*log(lambda.n1)
eta.n<-(n*apply(Y,1,mean)+eta*lambda)/lambda.n
Tn<-const+0.5*lambda.n1*apply((Y-eta.n)^2,1,sum)/n
Gn<-const+0.5*lambda.n1*(1+(eta.n-mu0)^2)
dsq<-function(xv,ev,lv){
    dv<-xv*0
    for(j in 1:length(xv)){
        dv[j]<-xv[j]-(sum(xv[-j])+ev*lv)/(length(xv)-1+lv)
    }
    return(sum(dv^2))
}
Cn<-const+0.5*lambda.ni1*apply(Y,1,dsq,ev=eta,lv=lambda)/n
Cn2<-const+0.5*apply((Y-eta.n)^2,1,sum)/(n*lambda.ni1)
ssq<-function(xv){
    return(sum((xv-mean(xv)^2)))
}
variance.term<-function(xv,ev,lv,N=10000){

    #Monte Carlo calculation
    en<-(sum(xv)+ev*lv)/(length(xv)+lv)
    ln<-length(xv)+lv

    mu<-rnorm(N,en,sqrt(1/ln))
    d<-outer(xv,mu,'-')

    return(mean(apply(dnorm(d,log=T),1,var)))
}
Wn<-Tn+apply(Y,1,variance.term,ev=eta,lv=lambda)
logZn<--0.5*n*log(2*pi)+0.5*log(lambda)-0.5*log(lambda.n)-0.5*apply(Y,1,ssq)-
        0.5*n*lambda*(apply(Y,1,mean)-eta)^2/lambda.n
Fn<--logZn
Fnbar<-Fn/n
lbl<-c(expression(T[n]),expression(G[n]),expression(C[n]),expression(C[n2]),expression(W[n]))
par(mar=c(4,4,3,0))
boxplot(cbind(Tn,Gn,Cn,Cn2,Wn),labels=lbl,ylim=range(0,2.5))
title('Boxplot of sampled statistic values over 1000 replicates (n=20)')
abline(h=0.5*(log(2*pi)+1),col='red')
legend(1,2.5,c(expression(paste('Limit value :',(log(2*pi)+1)/2))),col='red',lty=1)
```





Means across the 1000 replicate data sets for $n = 500$: each is approximately $(\log(2\pi) + 1)/2 \simeq 1.418939$.

```
+       Tn       Gn       Cn      Cn2       Wn
+ 1.416024 1.421383 1.420992 1.420992 1.421005
```