Bayesian Statistics

Convergence diagnostics

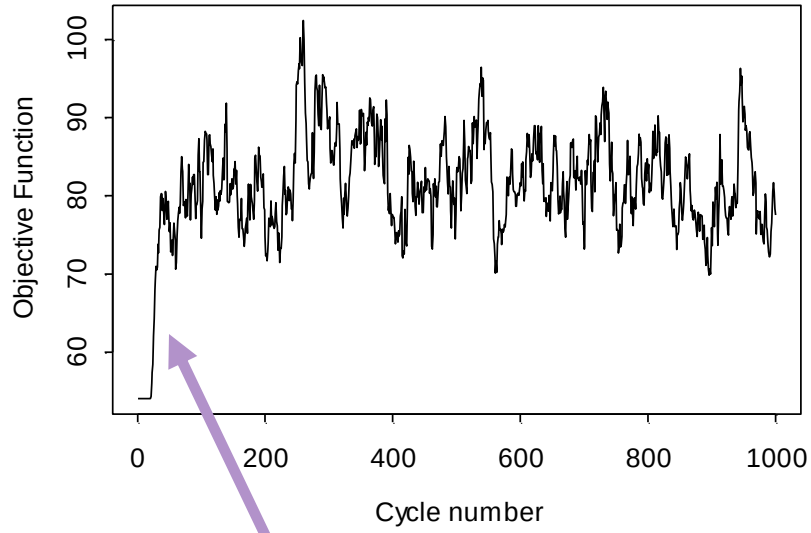
Nan Lin
Department of Mathematics
Washington University in St. Louis

Process the MCMC output

- Burn-in
- Thinning
- Monitoring convergence

- Tools
  - Traceplot
  - ACF plot
  - Acceptance (or Rejection) rate for the M-H algorithm
  - Quantitative procedures
- Software
  - R package **coda** and **BOA**

Burn-in

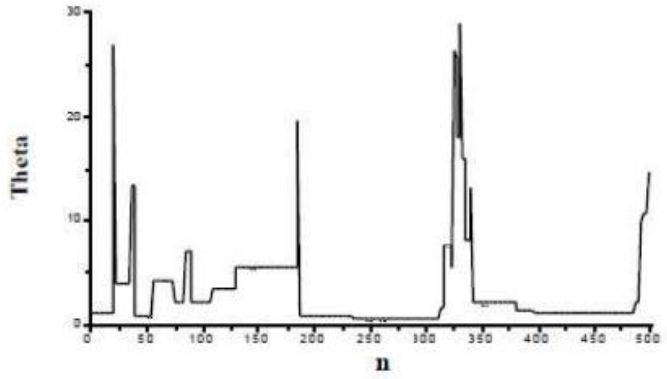


Need for a burn-in (the "pigtail")

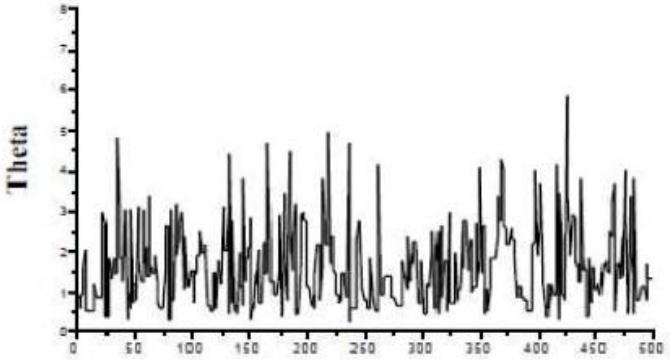
Mixing

- MCMC may have poor mixing when
  - autocorrelation is high
  - it takes a long time between jumps to different parts of the parameter space
  - Looking at the traceplot

- Poor mixing



- Well mixing



Autocorrelation

- Lag-t sample autocorrelation function

$$\text{acf}_t(\phi) = \frac{\frac{1}{S-t}\sum_{s=1}^{S-t}(\phi_s - \bar{\phi})(\phi_{s+t} - \bar{\phi})}{\frac{1}{S-1}\sum_{s=1}^{S}(\phi_s - \bar{\phi})^2},$$

  - Use R function `acf()`
  - Lower acf is better
- *Effect sample size* can be obtained using the R-command `effectivSize` in the `coda` package
- Thinning helps to reduce autocorrelation

Convergence diagnostics

- The MCMC theory guarantees that the Markov chain "eventually" will provide a good approximation to the target distribution.
- But "eventually" can be a very long time in some situations

- It is hard to prove the chain converged, but we can try to show the chain has not converged

Convergence diagnostics procedures

- Gelman and Rubin (1992)
- Geweke (1992)
- Raftery and Lewis (1992)
- And many others

- A comprehensive review is given by Cowles and Carlin (1996, JASA) "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review".

Gelman and Rubin (1992): potential scale reduction factor (PSRF)

- Based on normal theory approximation to exact Bayesian posterior inference
- Two major steps:
  - Step 1: Create an overdispersed estimate of the target distribution and use it to start **several independent sequences**.
  - Step 2: For each scalar quantity of interest, after running the Gibbs sampler chains for the desired number of iterations, say $2n$, estimate the PSRF, that is, the factor by which the scale parameter might shrink if sampling were continued indefinitely. And if the PSRF is close to 1, it indicates no violation of convergence.

Gelman and Rubin (1992)

- Step 1: Creating a starting distribution
  - Locate the high-density regions of the target distribution of x and find the K modes.
  - Approximate the high-density regions by a GMM:

$$\hat{P}(x) = \sum_{k=1}^{K} \omega_k (2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1} (x-\mu_k)\right)$$

  - Form an overdispersed distribution by first drawing from the GMM and then dividing each sample by a positive number, which results in a mixture t distributions:

$$\tilde{P}(x) \propto \sum_{k=1}^{K} \omega_k |\Sigma_k|^{-1/2} \left(\eta + (x-\mu_k)^t \Sigma_k^{-1} (x-\mu_k)\right)^{-(d+\eta)/2}$$

  - Sharpen the overdispersed approximation by downweighting regions that have relatively low density. For example, through importance resampling.

Gelman and Rubin (1992)

- Step 2: Re-estimating the target distribution
  - Independently simulate m sequences of length $2n$ from the overdispersed distribution and discard the first n iterations.
  - For each scalar parameter of interest, estimate the following quantity from the last n iterations of m sequences:
    - B: the variance between the means from m sequences;
    - W: the average of the m within-sequence variances;
    - $\bar{\bar{\theta}}$: estimate of target mean: mean of mn samples
    - $\hat{\sigma}^2$: estimate of target variance (unbiased)

$$\hat{\sigma}^2 = \frac{n-1}{n} W + \frac{1}{n} B$$

Within-chain variance

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2$$

where

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_j)^2$$

s_j^2 is just the formula for the variance of the jth chain. W is then just the mean of the variances of each chain.

W likely underestimates the true variance of the stationary distribution since our chains have probably not reached all the points of the stationary distribution.

Between-chain variance

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\bar{\theta}})^2$$

where

$$\bar{\bar{\theta}} = \frac{1}{m} \sum_{j=1}^{m} \bar{\theta}_j$$

This is the variance of the chain means multiplied by n because each chain is based on n draws.

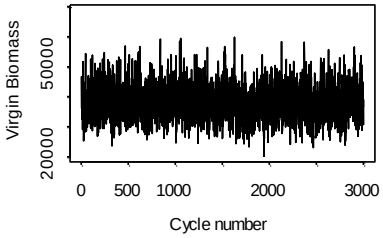
Gelman and Rubin (1992)

- Step 2: Re-estimating the target distribution (cont')
  - Estimate the posterior of target distribution as a t distribution with center $\hat{\mu}$ and scale $\sqrt{\hat{V}} = \sqrt{\hat{\sigma}^2 + B/mn}$
  - Monitor the convergence by the shrink factor $\sqrt{\hat{R}} = \sqrt{\left(\frac{\hat{V}}{W}\right)\frac{df}{df-2}}$, as it is near 1 for all scalars, collect burn-out samples.

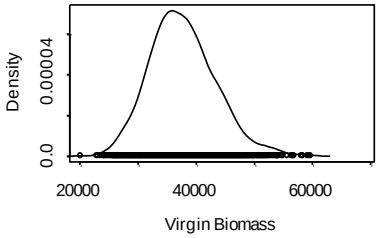
$$\sqrt{\hat{R}} = \sqrt{\left(\frac{n-1}{n} + \frac{m+1}{mn}\frac{B}{W}\right)\frac{df}{df-2}},$$

- Shrink factor approaches to 1→ Within-chain variance dominates between-chain variance, all sequences escaped the influence of starting points and traverse all target distributions.
  - If the PSRF is greater than 1.1 or 1.2, then we should perhaps run our chains out longer to improve convergence to the stationary distribution.

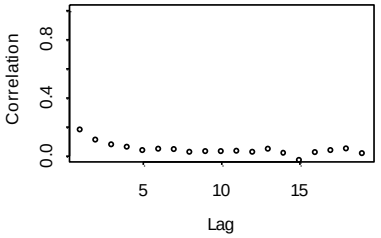
Trace - Chain #1



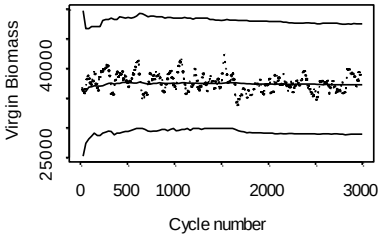
Kernel Density - Chain #1



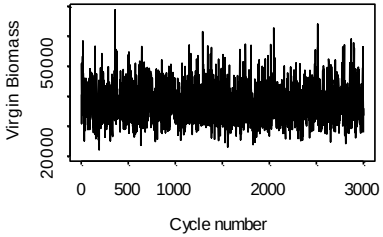
ACF vs. Lag - Chain #1



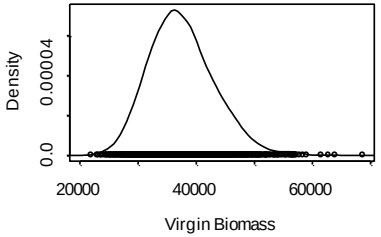
Cumulative patterns - Chain #1



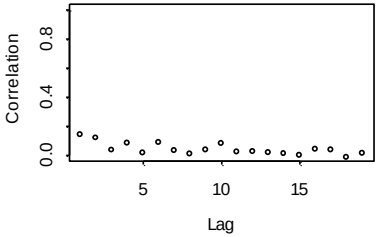
Trace - Chain #2



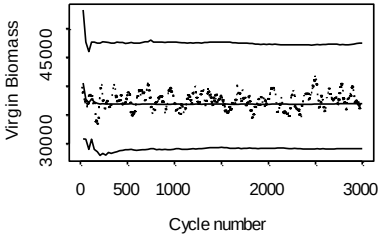
Kernel Density - Chain #2



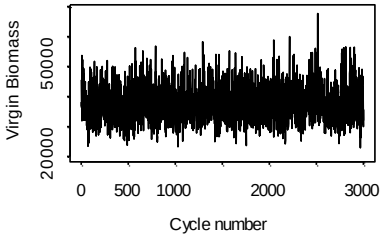
ACF vs. Lag - Chain #2



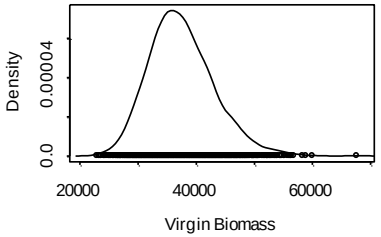
Cumulative patterns - Chain #2



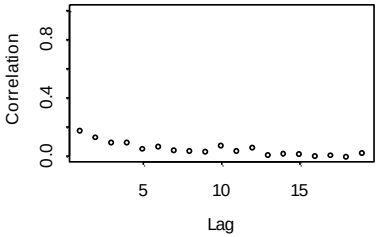
Trace - Chain #3



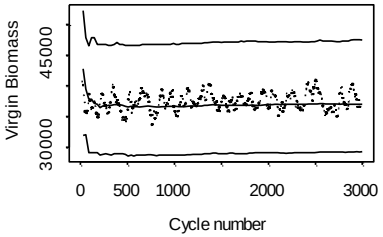
Kernel Density - Chain #3



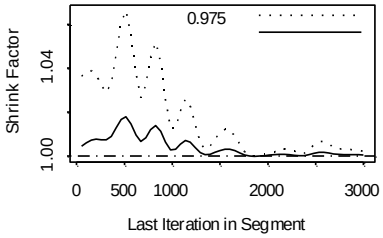
ACF vs. Lag - Chain #3



Cumulative patterns - Chain #3



Gelman and Rubin Shrink Factors



Gelman and Rubin (1992)

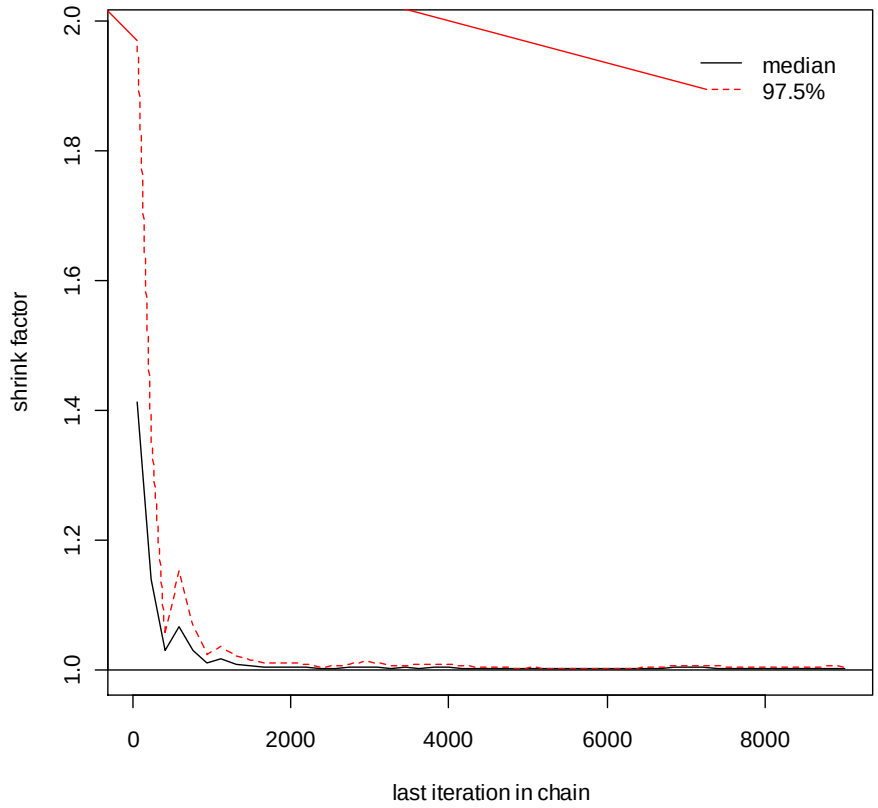
Steps (for each parameter):

1. Run $m \geq 2$ chains of length $2n$ from overdispersed starting values.
2. Discard the first n draws in each chain.
3. Calculate the within-chain and between-chain variance.
4. Calculate the estimated variance of the parameter as a weighted sum of the within-chain and between-chain variance.
5. Calculate the potential scale reduction factor.

```
> mh.draws1 <- mh.gamma(10000, start = 1, burnin = 1000, cand.sd = 1, shape = 1.7, rate = 4.4)
Acceptance Rate: 0.2720302
> mh.draws2 <- mh.gamma(10000, start = 1, burnin = 1000, cand.sd = 1, shape = 1.7, rate = 4.4)
Acceptance Rate: 0.2810312
> mh.draws3 <- mh.gamma(10000, start = 1, burnin = 1000, cand.sd = 1, shape = 1.7, rate = 4.4)
Acceptance Rate: 0.2779198
> mh.draws4 <- mh.gamma(10000, start = 1, burnin = 1000, cand.sd = 1, shape = 1.7, rate = 4.4)
Acceptance Rate: 0.2664741
> mh.draws5 <- mh.gamma(10000, start = 1, burnin = 1000, cand.sd = 1, shape = 1.7, rate = 4.4)
Acceptance Rate: 0.2584732
>
> mh.list <- mcmc.list(list(mcmc(mh.draws1), mcmc(mh.draws2), mcmc(mh.draws3), mcmc(mh.draws4), mcmc(mh.draws5)))
>
> gelman.diag(mh.list)
Potential scale reduction factors:

     Point est. 97.5% quantile
[1,]       1.00            1.00

> gelman.plot(mh.list)
```



Gelman and Rubin (1992)

- Shrink factor approaches to 1→ Within-chain variance dominates between-chain variance, all sequences escaped the influence of starting points and traverse all target distributions.

- A multivariate version of the statistic exists (Brooks and Gelman, 1997).

- Some concerns
  - Rely on the user's ability to find a start distribution.
  - Rely on normal approximation for diagnosing convergence to the true posterior.
  - Can be inefficient due to the need for multiple sequences and discarding a large number of early iterations.

Geweke (1992)

- Assumes that the intent of the Bayesian analysis is to estimate the mean $\mathrm{E}[g(\theta)]$ of some function $g(\theta)$ of interest
- Use methods from spectral analysis to assess convergence
  - Collect $g(\theta^{(j)})$ after each iteration
  - Treat $\{g(\theta^{(j)})\}_{j=1,p}$ as time series and compute spectral density $S_G(\omega)$.
  - Use numerical standard error (NSE) and relative numerical efficiency (RNE) to monitor convergence.

Geweke (1992)

- After running the chain for n iterations, estimate $\mathrm{E}[g(\theta)]$

$$\overline{g}_n = \frac{\sum_{i=1}^{n} g(\boldsymbol{\theta}^{(i)})}{n},$$

  - Asymptotic variance is $\mathrm{NSE} = S_g(0)/n$.
- Then test whether the mean from the first n_A iterations is equal to that from the last n_B iterations based on the asymptotic result

$$\frac{\bar{g}_n^A - \bar{g}_n^B}{\sqrt{\frac{S_g^A(0)}{n_A} + \frac{S_g^B(0)}{n_B}}} \to N(0,1)$$

  - Rejection means the chain has not converged
- Geweke suggests using $n_A = .1n$ and $n_B = .5n$.

Geweke (1992)

- Determine sufficient iterations:
  - Relative numerical efficiency (RNE)

$$\text{RNE} = \text{var}[g(\theta)]/S_G(0) = (2\pi)^{-1} \int_{-\pi}^{\pi} S_G(\omega)d\omega / S_G(0) \longrightarrow 1$$

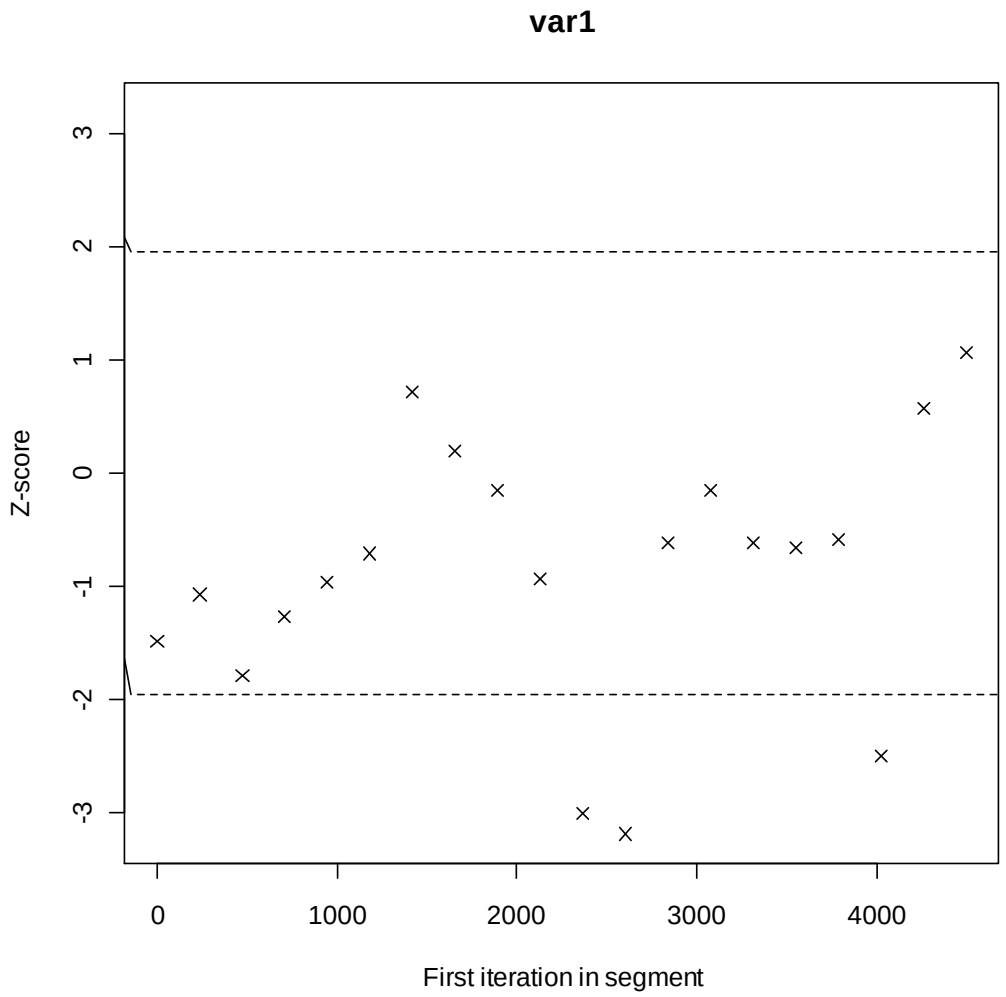
    - Indicating the number of draws would be required to produce the same numerical accuracy if the draws had been made from an iid sample drawn directly from the posterior distribution.
- Summary
  - Requires just a single chain
  - Sensitive to the spectral window

```
> geweke.diag(mh.draws) # the output is a Z-score

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  var1
-1.489

> geweke.plot(mh.draws)
```



Raftery and Lewis (1992)

- This method is intended both to detect convergence to the stationary distribution and to provide a way of bounding the variance of estimates of quantiles of functions of parameters
- Suppose that we want to measure some posterior quantile of interest q.
- If we define some acceptable tolerance r for q and a probability s of being within that tolerance, the Raftery and Lewis diagnostic will calculate the number of iterations N and the number of burn-ins M necessary to satisfy the specified conditions
- The diagnostic was designed to test the number of iterations and burn-in needed by first running and testing shorter pilot chain.

Raftery and Lewis (1992): Inputs

1. Select a posterior quantile of interest q (for example, the 0.025 quantile).
2. Select an acceptable tolerance r for this quantile (for example, if $r = 0.005$, then that means we want to measure the 0.025 quantile with an accuracy of ± 0.005).
3. Select a probability s, which is the desired probability of being within (q-r, q+r).
4. Run a "pilot" sampler to generate a Markov chain of minimum length given by rounding up

$$n_{\min} = \left[\Phi^{-1}\left(\frac{s+1}{2}\right) \frac{\sqrt{q(1-q)}}{r} \right]^2$$

where $\Phi^{-1}(\cdot)$ is the inverse of the normal CDF.

Raftery and Lewis (1992): Output

```
> raftery.diag(mh.draws, q = 0.025, r = 0.005, s = 0.95)

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

 Burn-in  Total Lower bound  Dependence
 (M)      (N)   (Nmin)       factor (I)
 16       17303 3746         4.62
```

- M: number of burn-ins necessary
- N: number of iterations necessary in the Markov chain
- $N_{\min}$: minimum number of iterations for the "pilot" sampler
- I: dependence factor, interpreted as the proportional increase in the number of iterations attributable to serial dependence.

  High dependence factors (> 5) are worrisome and may be due to influential starting values, high correlations between coefficients, or poor mixing.

Raftery and Lewis (1992)

- The Raftery-Lewis diagnostic will differ depending on which quantile q you choose.
- Estimates tend to be conservative in that it will suggest more iterations than necessary.
- It only tests marginal convergence on each parameter.
- It often works well with simple models.

Heidelberg and Welch (1983)

- The Heidelberg and Welch diagnostic calculates a test statistic (based on the Cramer-von Mises test statistic) to accept or reject the null hypothesis that the Markov chain is from a stationary distribution.

- The diagnostic consists of two parts.

Heidelberg and Welch (1983): Part I

1. Generate a chain of N iterations and define an α level.
2. Calculate the test statistic on the whole chain. Accept or reject null hypothesis that the chain is from a stationary distribution.
3. If null hypothesis is rejected, discard the first 10% of the chain. Calculate the test statistic and accept or reject null.
4. If null hypothesis is rejected, discard the next 10% and calculate the test statistic.
5. Repeat until null hypothesis is accepted or 50% of the chain is discarded. If test still rejects null hypothesis, then the chain fails the test and needs to be run longer.

Heidelberg and Welch (1983): Part II

- If the chain passes the first part of the diagnostic, then it takes the part of the chain not discarded from the first part to test the second part.
- The **halfwidth** test calculates half the width of the $100(1-\alpha)\%$ credible interval around the mean.
- If the ratio of the halfwidth and the mean is lower than some ϵ, then the chain passes the test. Otherwise, the chain must be run out longer.

```
> heidel.diag(mh.draws)

     Stationarity start     p-value
     test         iteration
var1 passed       1         0.109

     Halfwidth Mean  Halfwidth
     test
var1 passed    0.386 0.0208
```

About convergence diagnostics

- Can NOT prove the chain converges
- Be cautious when using these diagnostics;
- Use a variety of diagnostic tools rather than any single one;
- Learn as much as possible about the target density before applying the MCMC algorithm