Statistical Analysis II: Project 2 report

Kornel Howil

January 19, 2023

1 Probabilities

Knowing the bayesian network structure we can write that

$$P(C|R,S,W) = P(C|R,S) \propto P(C,R,S) \propto P(C)P(S|C)P(R|C) \tag{1}$$

and

$$P(R|C,S,W) \propto P(R,C,S,W) \propto P(C)P(S|C)P(R|C)P(W|R,S) \tag{2}$$

1.1 $\mathbf{P(C=T|R=T,S=T,W=T)}$

From equation 1 we can write that

$$P(C=T|R=T,S=T,W=T) \propto 0.5 \cdot 0.1 \cdot 0.8 = 0.04$$

and

$$P(C=F|R=T,S=T,W=T) \propto 0.5 \cdot 0.5 \cdot 0.2 = 0.05$$

Hence after normalising

$$P(C=T|R=T,S=T,W=T) = \frac{0.04}{0.04+0.05} \approx 0.44$$

1.2 $\mathbf{P(C=T|R=F,S=T,W=T)}$

From equation 1 we can write that

$$P(C=T|R=F,S=T,W=T) \propto 0.5 \cdot 0.1 \cdot 0.2 = 0.01$$

and

$$P(C=F|R=F,S=T,W=T) \propto 0.5 \cdot 0.5 \cdot 0.8 = 0.2$$

Hence after normalising

$$P(C=T|R=T,S=T,W=T) = \frac{0.01}{0.01+0.2} \approx 0.05$$

1.3 $\mathbf{P(R=T|C=T,S=T,W=T)}$

From equation 2 we can write that

$$P(R=T|C=T,S=T,W=T) \propto 0.5 \cdot 0.1 \cdot 0.8 \cdot 0.99 = 0.0396$$

and

$$P(R=F|C=T,S=T,W=T) \propto 0.5 \cdot 0.1 \cdot 0.2 \cdot 0.90 = 0.009$$

Hence after normalising

$$P(R=T|C=T,S=T,W=T) = \frac{0.0396}{0.0396+0.009} \approx 0.81$$

1.4 $\mathbf{P(R = T|C = F, S = T, W = T)}$

From equation 2 we can write that

$$P(R = T|C = F, S = T, W = T) \propto 0.5 \cdot 0.5 \cdot 0.2 \cdot 0.99 = 0.0495$$

and

$$P(R = F|C = F, S = T, W = T) \propto 0.5 \cdot 0.5 \cdot 0.8 \cdot 0.90 = 0.18$$

Hence after normalising

$$P(R = T|C = F, S = T, W = T) = \frac{0.0495}{0.0495 + 0.18} \approx 0.22$$

2 Gibbs Sampler

Using the calculated above probabilities, I implemented the Gibbs sampler in NumPy. This sampler was used to draw 100 samples 1000 times. This allowed us to make a histogram shown in the Figure 1.

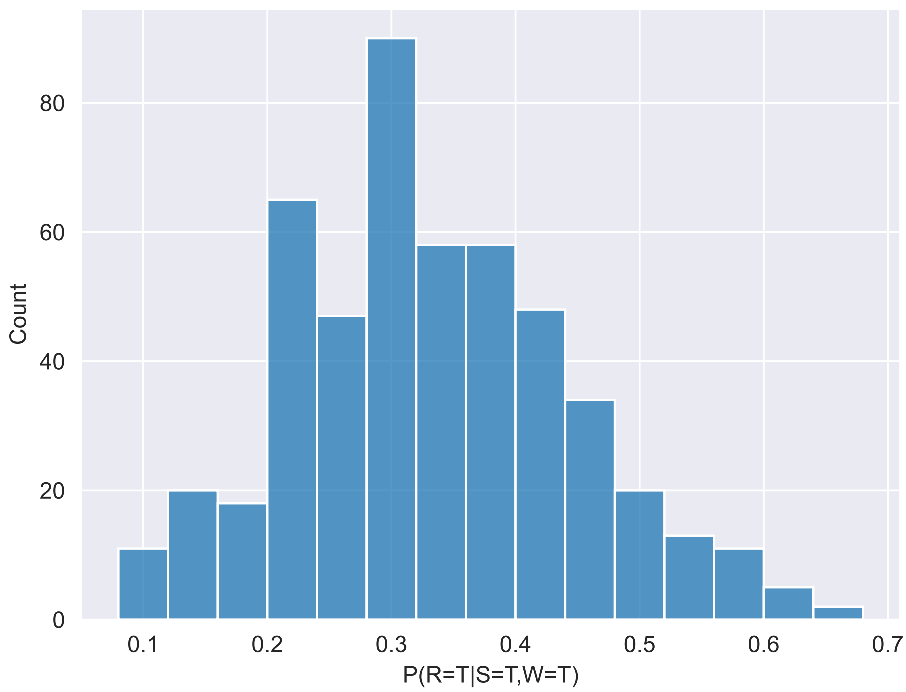

Figure 1: Histogram of values of the conditional probability $P(R = T|S = T, W = T)$ obtained by sampling 100 values from Gibbs sampler 1000 times.

Mean value of the probability is 0.3347 and variance is equal to 0.0128.

3 Convergence diagnostics

3.1 Burn-in phase

Gibbs sampler was used to sample 50000 samples 10 times. Relative frequencies were calculated for each of these 10 chains. Figures 2 and 3 show relative frequencies in the function of iteration all chains for respectively C=T and R=T.

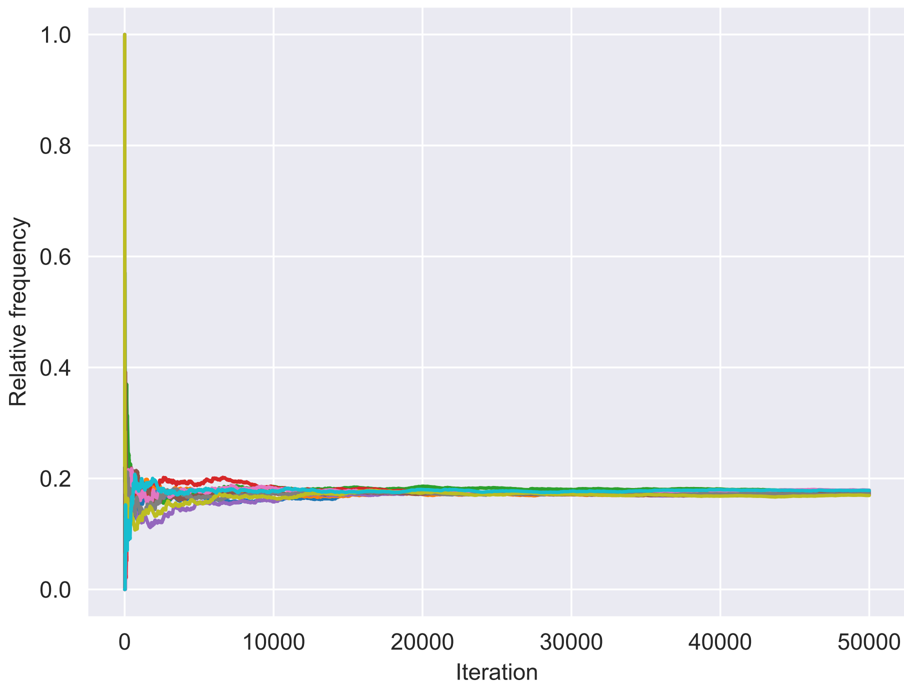

Figure 2: Relative frequencies in function of iteration of 10 chains for C=T.

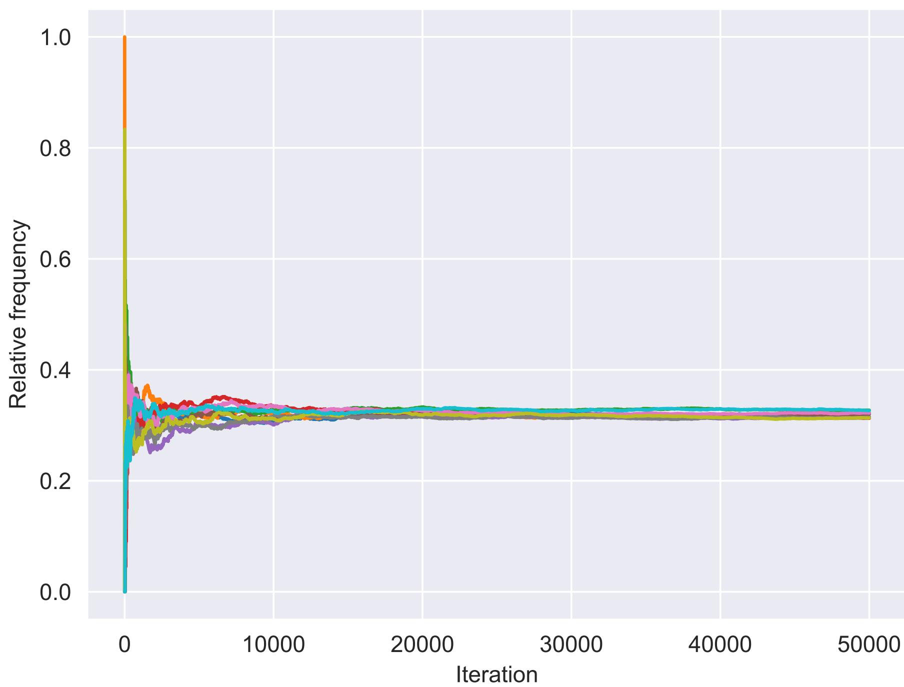

Figure 3: Relative frequencies in function of iteration of 10 chains for R=T

Using those plots we can estimate that burn-in phase lasting for 10000 iterations should be good enough.

3.2 Autocorrelation

Correlograms for both variables Cloud and Rain were produced using `plot_acf` function from `statsmodels` package.

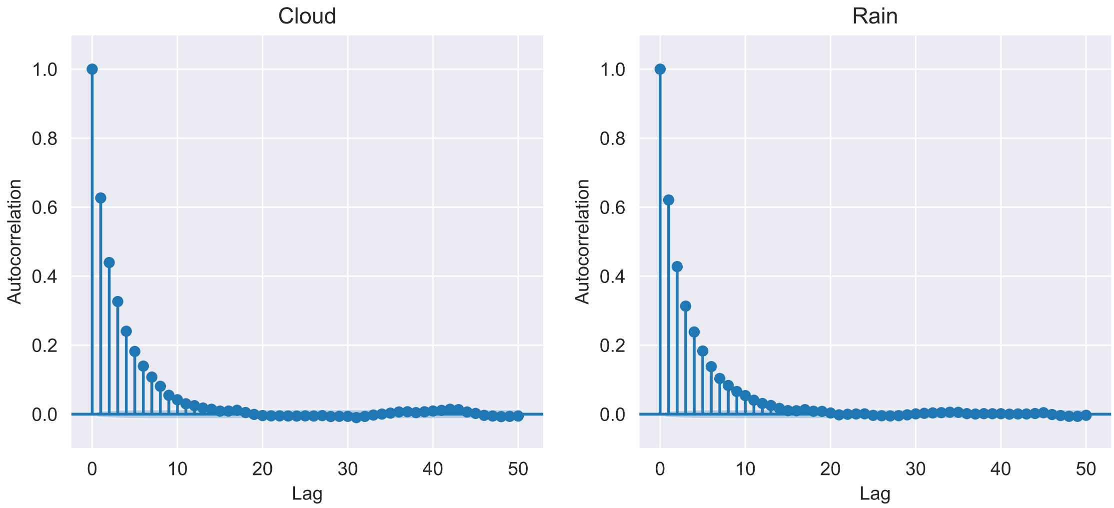

Figure 4: Correlograms for both variables Cloud (left) and Rain (right).

It can be seen from both figures that autocorrelation from lag greater than 15 is negligible. Hence, we can assume that keeping every 15-th sample would be sufficient way to thin-out these chains.

4 100 samples after burn_in and thinning_out

Using chosen above values of burn_in and thin_out once again 100 samples (after burn_in and thinning_out) were drawn 500 times. Histogram of obtained conditional probabilities $P(R = T|S = T, W = T)$ is shown on the Figure 5.

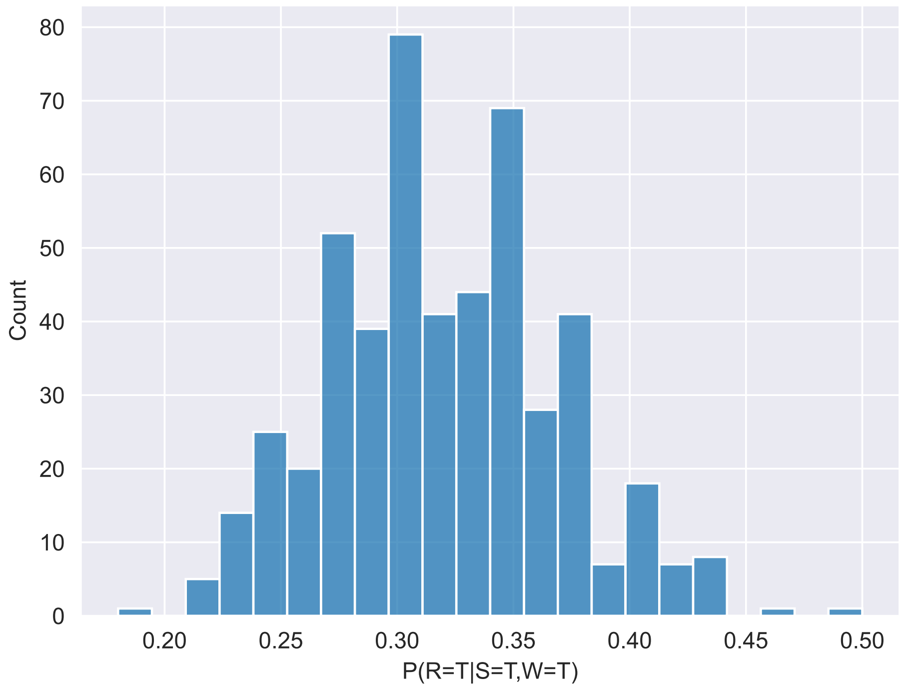

Figure 5: Histogram of values of the conditional probability $P(R = T|S = T, W = T)$ obtained by sampling 100 values (after burn_in and thinning_out) from Gibbs sampler 1000 times.

The mean value of the conditional probability is 0.3196 and the variance is equal to 0.0024. We can see that variance is approximately 5 times smaller than in the run without burning_in and thinning_out. Hence, we can be more certain about the value of this probability.

5 Final results

5.1 Final run of gibbs sampler

For the final estimation of the $P(R = T|S = T, W = T)$ gibbs sampler was run 500 times for $n = 50000$ iterations with chosen earlier values of burn_in and thin_out. Histogram of obtained conditional probabilities is shown on the Figure 6.

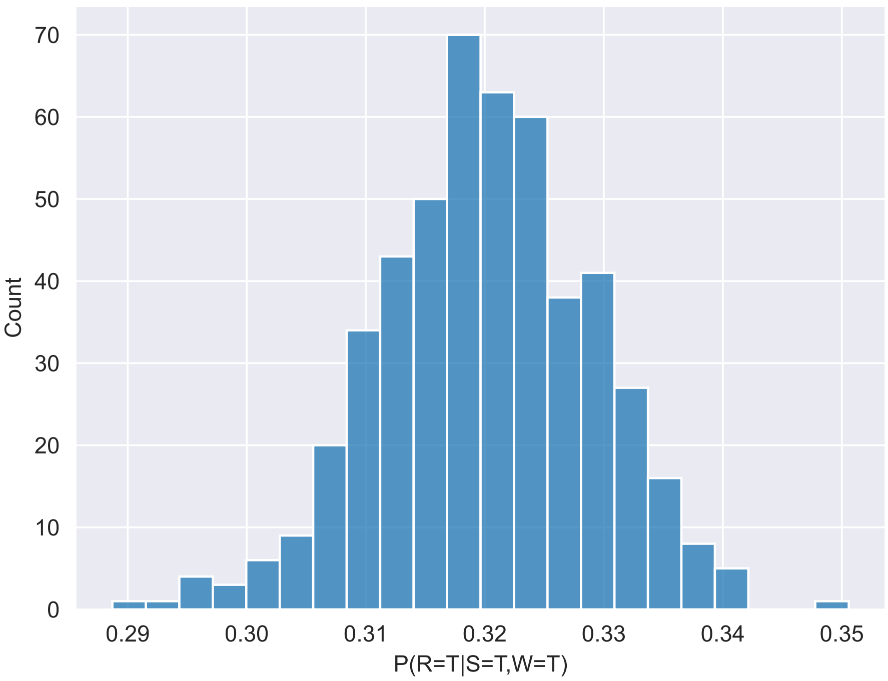

Figure 6: Histogram of values of the conditional probability $P(R = T|S = T, W = T)$ obtained by sampling 50000 values, burning_in and thinning_out from Gibbs sampler 1000 times.

Mean value of the conditional probability is 0.3200 and variance is equal to $7.9845 \cdot 10^{-5}$. The variance is even smaller than in the last run with 100 samples.

5.2 Exact result

We can use values calculated in section 1 (before normalising) to calculate exact value of $P(R = T|S = T, W = T)$. Since

$$P(R = T|C = T, S = T, W = T) \propto 0.0396,$$
$$P(R = F|C = T, S = T, W = T) \propto 0.0090,$$
$$P(R = T|C = F, S = T, W = T) \propto 0.0495,$$
$$P(R = F|C = F, S = T, W = T) \propto 0.1800,$$

we can write that

$$P(R = T|S = T, W = T) = \frac{0.0396 + 0.0495}{0.0396 + 0.0495 + 0.009 + 0.18} \approx 0.32039. \tag{3}$$

In conclusion, result obtained from Gibbs sampler is very close to the exact value.

6 Gelman-Rubin convergence test

Gelman-Rubin convergence test was implemented according to the equation found in Vats and Knudson 2021 [1].

$$\hat{R} = \sqrt{\frac{\hat{\sigma}^2}{s^2}}. \tag{4}$$

Convergence was tested on 100 samples without burning-in and thinning-out and 100 samples after both operations, with parameters found in section 3. Both times, the score was calculated based on $m = 100$ chains. Results for samples without and with burning-in and thinning-out are shown respectively in 5 and 6.

$$\hat{R}_{\text{Cloud}} \approx 1.0341, \quad \hat{R}_{\text{Rain}} \approx 1.0378 \tag{5}$$

$$\hat{R}_{\text{Cloud}} \approx 1.0093, \quad \hat{R}_{\text{Rain}} \approx 1.0095 \tag{6}$$

Assuming one of the most popular cutoffs of $\hat{R} < 1.01$ [1] we see that in the first case, the chains have not reached convergence and for the second case convergence has been reached.

References

[1] Dootika Vats. Christina Knudson. "Revisiting the Gelman–Rubin Diagnostic." Statist. Sci. 36 (4) 518 - 529, November 2021. , *https://doi.org/10.1214/20-STS812*