Chapter 8.7 Posterior Predictive Checking

Jim Albert and Monika Hu

Chapter 8 Modeling Measurement and Count Data

Introduction

- The posterior predictive distribution is helpful for assessing the suitability of the Bayesian model.
- The question is whether these observed times to serve for Federer are consistent with replicated data from the posterior predictive distribution.
- Replicated refers to the same sample size as our original sample. If one takes samples of 20 from the posterior predictive distribution, do these replicated datasets resemble the observed sample?

Simulations of replicated data

- Since the population standard deviation is known as $\sigma = 4$ seconds, the sampling distribution of Y is Normal with mean μ and standard deviation σ.
- One simulates replicated data $\tilde{Y}_1, ..., \tilde{Y}_{20}$ from the posterior predictive distribution in two steps:

1. Sample a value of μ from its posterior distribution

$$\mu \sim \mathrm{Normal}\left(\frac{\phi_0\mu_0 + n\phi\bar{y}}{\phi_0 + n\phi}, \sqrt{\frac{1}{\phi_0 + n\phi}}\right).$$

2. Sample $\tilde{Y}_1, ..., \tilde{Y}_{20}$ from the data model

$$\tilde{Y} \sim \mathrm{Normal}(\mu, \sigma).$$

Using R

- Implement method in the following R script to simulate 1000 replicated samples from the posterior predictive distribution.
- The vector `pred_mu_sim` contains draws from the posterior and the matrix `ytilde` contains the simulated predictions where each row of the matrix is a simulated sample of 20 future times.

```
sigma <- 4;  mu_n <- 17.4
sigma_n <- 0.77; S <- 1000
pred_mu_sim <- rnorm(S, mu_n, sigma_n)
sim_ytilde <- function(j){
  rnorm(20, pred_mu_sim[j], sigma)
}
ytilde <- t(sapply(1:S, sim_ytilde))
```

Goodness of fit

- To judge goodness of fit, compare these simulated replicated datasets from the posterior predictive distribution with the observed data.
- One convenient way to implement this comparison is to compute some “testing function”, $T(\tilde{y})$, on each replicated dataset.
- One constructs a graph of these values and overlays the value of the testing function on the observed data $T(y)$.
- If the observed value is in the tail of the posterior predictive distribution of $T(\tilde{y})$, this indicates some misfit of the observed data with the Bayesian model.

Choosing a testing function

To implement this procedure, one needs to choose a testing function $T(\tilde{y})$. Suppose, for example, one decides to use the sample mean $T(\tilde{y}) = \sum \tilde{y}_j / 20$. In the R script, we compute the sample mean on each row of the simulated prediction matrix.

```r
pred_ybar_sim <- apply(ytilde, 1, mean)
```

Predictive distribution

- Show density estimate of the posterior predictive distribution of $\bar{Y}$ and the observed value of the sample mean $\bar{Y} = 17.20$ is displayed as a vertical line.

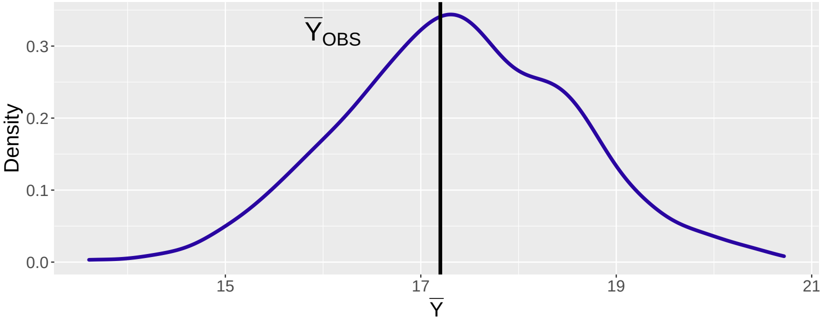

Figure 1: Display of the posterior predictive mean time-to-serve for twenty observations. The observed mean time-to-serve value is displayed by a vertical line.

Conclusion

- Note this observed mean is in the middle of this distribution
- One concludes that this observation is consistent with samples predicted from the Bayesian model.
- But one can consider alternative choices for checking functions.