COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

Introduction

- We now discuss models for the comparison of means across groups.
- We parameterize the population means by their average and their differences.
- The average group mean and the differences across group means are described by a normal sampling model.
- This model, together with a normal sampling model for variability among units within a group, make up a hierarchical normal model that describes both within-group and between-group variability.
- We also discuss an extension to this normal hierarchical model which allows for across-group heterogeneity in variances in addition to heterogeneity in means.

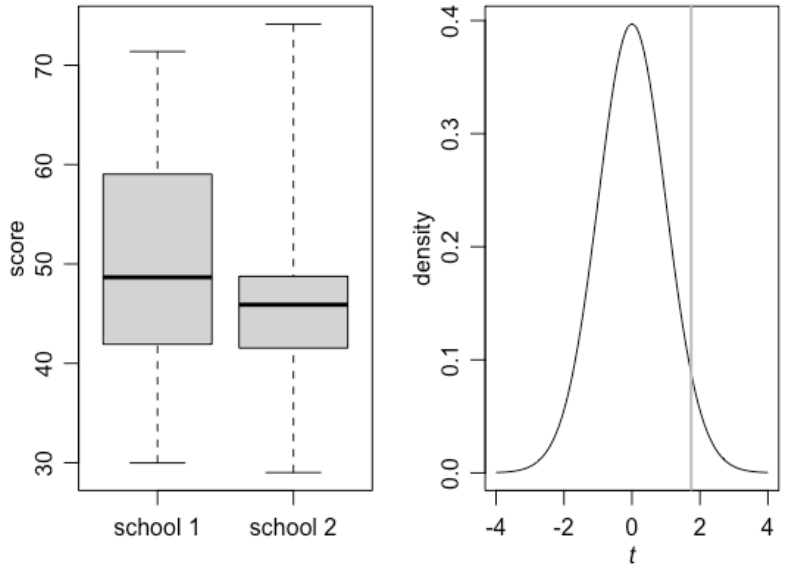
RStan

- Stan is a C++ library for Bayesian modeling and inference.
- Install `rstan` package and all dependents.
- RStan allows a Stan program to be coded in a text file (typically with suffix .stan).
- The first section of a Stan program, the `data` block, specifies the data that is conditioned upon in Bayesian inference.
- The `parameters` block declares the parameters whose posterior distribution is sought.
- Finally, the `model` block looks similar to standard statistical notation.
- Each block is written inside {} braces.
- Each line within a block is separated with ,
- Stan has versions of many of the most useful R functions for statistical modeling, including probability distributions, matrix operations, and various special functions.

Comparing Two Groups

Math scores from a sample of 10th grade students from two public U.S. high schools.

Second graph is the t-statistic value and corresponding distribution.



Example (continued)

Suppose we are interested in estimating θ_1, the average score we would obtain if all 10th graders in school 1 were tested, and possibly comparing it to θ_2, the corresponding average from school 2.

The t-statistic, assuming equal population variances, is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 1.74$$

Is this large enough to reject H_0: $\theta_1 = \theta_2$

See R code.

Example (continued)

- Frequentist approach: since p-value = 0.087 we must reject H_0
- This data analysis procedure results in either treating the two populations as completely distinct or treating them as identical.
- The results tell us to treat the population means of the two groups as being numerically equivalent, although there seems to be some evidence of a difference.
- Instead of using such an extreme procedure, it might make more sense to allow the difference between the groups to vary continuously and have a value that depends on such things as the relative sample sizes n_1 and n_2, the sampling variability σ^2 and our prior information about the similarities of the two populations.

Bayesian Model

Consider the following sampling model for data from the two groups:

$$X_{i1} = \mu + \delta + \varepsilon_{i1}$$
$$X_{i2} = \mu - \delta + \varepsilon_{i2}$$
$$\varepsilon_{ij} \text{ are iid } N(0, \sigma^2)$$

Here δ represents half of the population difference: $\delta = \frac{\theta_1 - \theta_2}{2}$ and μ represents the pooled average: $\mu = \frac{\theta_1 + \theta_2}{2}$.

Conjugate priors:

$$\pi(\mu, \delta, \sigma^2) = \pi(\mu) \times \pi(\delta) \times \pi(\sigma^2)$$
$$\mu \sim N(\mu_0, \gamma_0^2)$$
$$\delta \sim N(\delta_0, \tau_0^2)$$
$$\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

Conditional Posteriors

It can be shown that:

$$\mu|\boldsymbol{x}_1, \boldsymbol{x}_2, \delta, \sigma^2 \sim N(\mu_n, \gamma_n^2)$$
$$\delta|\boldsymbol{x}_1, \boldsymbol{x}_2, \mu, \sigma^2 \sim N(\delta_n, \tau_n^2)$$
$$\sigma^2|\boldsymbol{x}_1, \boldsymbol{x}_2, \mu, \delta \sim IG\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right)$$

where

$$\mu_n = \gamma_n^2 \left[\frac{\mu_0}{\gamma_0^2} + \frac{\sum_{i=1}^{n_1}(y_{i1} - \delta)}{\sigma^2} + \frac{\sum_{i=1}^{n_2}(y_{i2} + \delta)}{\sigma^2}\right]$$

$$\gamma_n^2 = \frac{1}{\frac{1}{\gamma_0^2} + \frac{n_1 + n_2}{\sigma^2}}$$

$$\delta_n = \tau_n^2 \left[\frac{\delta_0}{\tau_0^2} + \frac{\sum_{i=1}^{n_1}(y_{i1} - \mu)}{\sigma^2} - \frac{\sum_{i=1}^{n_2}(y_{i2} - \mu)}{\sigma^2}\right]$$

… (see p. 128)

Introduction to Hierarchical Models

The data in the previous example was part of the 2002 Educational Longitudinal Study (ELS), a survey of students from a large sample of schools across the United States. This dataset includes a population of schools as well as a population of students within each school. Datasets like this, where there is a hierarchy of nested populations, are often called hierarchical or multilevel. Other situations having the same sort of structure:

- patients within several hospitals,
- genes within a group of animals,
- people within counties within regions within countries

The simplest type of multilevel data has two levels, in which one level consists of groups and the other consists of units within groups. In this case we denote $x_{i,j}$ as the data on the ith unit in group j.

Exchangeability

The samples in our datasets that are not completely independent. Samples in hierarchical datasets often form clusters or groups within which some properties are shared. It is reasonable to assume observations within such clusters are exchangeable, or by de Finetti's theorem, they are conditionally independent, given some population feature. That is, we assume:

$$X_1, \ldots, X_n | \phi \sim iid\ f(x|\phi)$$
$$\phi \sim \pi(\phi)$$

For example, in the normal model $\phi = (\theta, \sigma^2)$

Hierarchical Data Structure

Suppose we have m groups and data ($\boldsymbol{X}_1, \ldots \boldsymbol{X}_m$) from each of them, where each $\boldsymbol{X}_j = \left(X_{1j}, \ldots, X_{n_j j}\right)$.

The random variables $X_{1j}, \ldots, X_{n_j j}$ should not be independent, because they all belong to the same cluster (students within each school perform similarly, because they have the same professors, use same textbooks, etc.). Treating them as exchangeable variables makes more sense. That is,

$$X_{1j}, \ldots, X_{n_j j} | \phi_j \sim iid\ f\left(x | \phi_j\right)$$

where ϕ_j is some group-specific parameter.

Hierarchical Data Structure

Key question: how should we represent our information about $\phi_1, \ldots, \phi_m$?

The two extremes are:

- $\phi_1, \ldots, \phi_m$ are independent (aka no pooling of information between clusters). This is the approach taken by classical ANOVA models
- $\phi_1, \ldots, \phi_m$ are the same (aka complete pooling). In this approach we put all observations together and analyze them as if they came from a single cluster.
- $\phi_1, \ldots, \phi_m$ are conditionally independent or exchangeable (partial pooling). That is,

$$\phi_1, \ldots, \phi_m | \psi \sim iid\ \pi(\phi|\psi)$$

Hierarchical Data Structure

Two-level hierarchical data model:

$$X_{1j}, \ldots, X_{n_j j} | \phi_j \sim f\left(x \middle| \phi_j\right), j = 1, \ldots, m$$
$$\phi_j | \psi \sim \pi(\phi | \psi)$$
$$\psi \sim \pi(\psi)$$

- The first line models the *within-group* sampling variability.
- The second line models the *between-group* sampling variability.
- Notice how the parameters are nested within one another.
- The third line is the prior.
- Often the parameter ψ will contain some sort of variation component and if it is estimated to be high, this means the clusters were close to independent, and if it is estimated low, this means the clusters were almost identical.
- Let the data decide how much info to pool!

The Hierarchical Normal Model

Two-level hierarchical normal data model with $\phi_j = (\theta_j, \sigma^2)$ and $\psi = (\mu, \tau^2)$

$$X_{1j}, \ldots, X_{n_j j} | \phi_j \sim N(\theta_j, \sigma^2)$$
$$\theta_j | \psi \sim N(\mu, \tau^2)$$
$$\psi \sim \pi(\psi)$$

In the school example, notice the meaning of these parameters:

θ_j is the average performance of all students withing school j (aka random effects)

σ^2 measures the within-cluster variability (ie, student level)

μ is the average performance of all students in all schools (aka fixed effect)

τ^2 measures the between-cluster variability (ie, school level)

Priors

Two-level hierarchical normal data model with $\phi_j = (\theta_j, \sigma^2)$ and $\psi = (\mu, \tau^2)$ is:

$$X_{1j}, \ldots, X_{n_j j} | \phi_j \sim N(\theta_j, \sigma^2)$$

$$\theta_j | \psi \sim N(\mu, \tau^2)$$

$$\psi \sim \pi(\psi)$$

We will use the following semiconjugate priors:

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

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

$$\mu \sim N(\mu_0, \gamma_0^2)$$

Posterior Inference

We need to sample (or approximate) the full posterior distribution $f(\theta_1, \ldots, \theta_m, \mu, \tau^2, \sigma^2 | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$.

We will use Gibbs sampler approach. First note that:

$$
\begin{aligned}
& f(\theta_1, \ldots, \theta_m, \mu, \tau^2, \sigma^2 | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) \\
& \propto \pi(\mu, \tau^2, \sigma^2) \times \pi(\theta_1, \ldots \theta_m | \mu, \tau^2, \sigma^2) \\
& \times f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m | \theta_1, \ldots \theta_m, \mu, \tau^2, \sigma^2) \\
= \pi(\mu)\pi(\tau^2)\pi(\sigma^2) & \left[\prod_{j=1}^{m} \pi\left(\theta_j | \mu, \tau^2\right) \right] \left[\prod_{j=1}^{m} \prod_{i=1}^{n_j} f\left(x_{ij} | \theta_j, \sigma^2\right) \right]
\end{aligned}
$$

From here we must derive each of the conditional posterior distributions.

Conditional posteriors of μ and τ^2

The part of the joint posterior that depends on μ and τ^2 is

$$\pi(\mu)\pi(\tau^2)\left[\prod_{j=1}^{m}\pi\left(\theta_j|\mu,\tau^2\right)\right]$$

This means that:

$$f(\mu|\theta_1,\dots,\theta_m,\tau^2,\sigma^2,\boldsymbol{x}_1,\dots,\boldsymbol{x}_m)\propto\pi(\mu)\prod_{j=1}^{m}\pi\left(\theta_j|\mu,\tau^2\right)$$

$$f(\tau^2|\theta_1,\dots,\theta_m,\mu,\sigma^2,\boldsymbol{x}_1,\dots,\boldsymbol{x}_m)\propto\pi(\tau^2)\prod_{j=1}^{m}\pi\left(\theta_j|\mu,\tau^2\right)$$

and it can be shown that these are normal and IG distributions

Conditional posteriors of θ_j and σ^2

It can be shown that $\boldsymbol{\theta_j}$s are conditionally independent and

$$f\left(\theta_j \middle| \mu, \tau^2, \sigma^2, \boldsymbol{x}_1, \dots, \boldsymbol{x}_m\right) \propto \pi\left(\theta_j \middle| \mu, \tau^2\right) \prod_{i=1}^{n_j} f\left(x_{ij} \middle| \theta_j, \sigma^2\right)$$

and

$$\theta_j | \mu, \tau^2, \sigma^2, \boldsymbol{x}_1, \dots, \boldsymbol{x}_m \sim N\left(\frac{\frac{n_j \bar{x}_j}{\sigma^2} + \frac{1}{\tau^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

For the distribution of

$$\sigma^2 | \theta_1, \dots \theta_m, \boldsymbol{x}_1, \dots, \boldsymbol{x}_m \sim IG$$

see the details on p. 135.