STA 114: Statistics

Lab 7

In this lab we'd consider modeling samples $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_m)$ from two different groups and compare some group characteristics. More precisely, we would look at models $X \sim f_1(x|\theta_1)$, $\theta_1 \in \Theta_1$, $Y \sim f_2(y|\theta_2)$, $\theta_2 \in \Theta_2$, X, Y independent, and look at some quantity $\eta = h(\theta_1, \theta_2)$ that provides a comparison of θ_1 and θ_2. We would look at three examples.

Example (Early vs delayed IV to stab victims)**.** To study whether early injection of IV fluids could be harmful to patients with penetrating injuries to the torso, data were collected at Ben Taub General Hospital in Houston under two treatment settings. In the early resuscitation group, 309 patients were given fluids before they reached the hospital. Another 289 patients in the delayed resuscitation group did not receive any fluid until they reached the operation theater. The study recorded the number X and Y of patients surviving for each group. Our model for these data is $X \sim \mathsf{Binomial}(309, p_1)$, $Y \sim \mathsf{Binomial}(289, p_2)$, X, Y independent, and we are interested in $\eta = p_1 - p_2$, the difference in survival rate under the two treatments.

Example (Duke student food expenditure)**.** We have data $X = (X_1, \cdots, X_n)$ from 2010 and $Y = (Y_1, \cdots, Y_m)$ from 2011 on the weekly food expenditure by randomly selected Duke undergraduate students. We model $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma_1^2)$, $Y_j \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_2, \sigma_2^2)$, X, Y independent, and are interested in $\eta = \mu_1 - \mu_2$, the difference in mean expenditure in the two years.

Example (Tropical cyclone intensity)**.** We have data $X = (X_1, \cdots, X_n)$, $Y = (Y_1, \cdots, Y_m)$ respectively on the maximum wind speeds (minus 35 knots) of all tropical cyclones between 1981-1993 and between 1994-2006. Modeling these as $X_i \overset{\text{IID}}{\sim} \mathsf{Exponential}(\lambda_1)$, $Y_j \overset{\text{IID}}{\sim} \exp(\lambda_2)$, X, Y independent, we want to infer on $\eta = 1/\lambda_1 - 1/\lambda_2$, the difference in mean intensities from the two time periods.

Bayesian approach: General strategy

For this lab, we would look at product priors on (θ_1, θ_2), i.e., the prior pdf is of the form $\xi(\theta_1, \theta_2) = \xi_1(\theta_1)\xi_2(\theta_2)$ for some prior pdfs $\xi_1(\theta_1)$ and $\xi_2(\theta_2)$ on Θ_1 and Θ_2 respectively. By our modeling assumption of independence, the likelihood function also exhibits a product structure:

$$L_{x,y}(\theta_1, \theta_2) = f(x, y|\theta_1, \theta_2) = f_1(x_1|\theta_1) f_2(x_2|\theta_2)$$

and hence the posterior pdf is also of a product form:

$$\begin{aligned} \xi(\theta_1, \theta_2|x, y) &= \text{const.} \times L_{x,y}(\theta_1, \theta_2)\xi(\theta_1, \theta_2) \\ &= \text{const.} \times \{f_1(x|\theta)\xi_1(\theta_1)\} \times \{f_2(y|\theta_2)\xi_2(\theta_2)\} \\ &= \xi_1(\theta_1|x)\xi_2(\theta_2|y) \end{aligned}$$

where $\xi_1(\theta_1|x) = \text{const.} \times f_1(x|\theta_1)\xi_1(\theta_1)$ is the posterior pdf you'd obtain on θ_1 by analyzing only the "x-model": $X \sim f_1(x|\theta_1)$, $\theta_1 \sim \xi_1(\theta_1)$ [and ignoring Y]. Similarly, $\xi_2(\theta_2|y)$ is the posterior pdf you'd obtain on θ_2 by analyzing only the "y-model": $Y \sim f_2(y|\theta_2)$, $\theta_2 \sim \xi_2(\theta_2)$.

You might worry how are we going to make a comparison between θ_1 and θ_2, if the analyses on θ_1 and θ_2 are done separately. Keep in mind, that the analyses are not entirely separate. We still have a joint posterior pdf on (θ_1, θ_2), on that it is of the product form, so under the posterior, θ_1 and θ_2 can be treated as independent. But we can still compare the ranges of their values.

IV fluid example

For the IV fluid analysis, assign a product prior $\xi(p_1, p_2) = \xi_1(p_1)\xi_2(p_2)$ where both $\xi_1(p_1)$ and $\xi_2(p_2)$ are the $\mathsf{Uniform}(0, 1)$ pdf.

TASK 1. Draw samples of (p_1, p_2) from the posterior $\xi(p_1, p_2|x = 193, y = 203)$ and calculate the posterior probability that $p_1 > p_2$. Also find a central 95% posterior credible interval for the difference in log-odds ratios or survival $\eta = \log \frac{p_1}{1-p_1} - \log \frac{p_2}{1-p_2}$.

Food expenditure example

For the food expenditure analysis, assign a product prior $\xi(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \xi_1(\mu_1, \sigma_1^2)\xi_2(\mu_2, \sigma_2^2)$, where both ξ_1 and ξ_2 are $\mathsf{N}\chi^{-2}(125, 0.3, 1, 53)$.

TASK 2. Draw samples of $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ from the posterior for observed data

Year	Expenditures
2010	139 210 75 131 184 140 140 127 100 129 145 145 120 23
2011	125 140 200 200 190 100 140 250 125 180 110 125 120 130 140 150 120 100 95 195 95 130

Calculate the posterior probability of $\mu_1 > \mu_2$.

TASK 3. Now consider "future" observations $X* = X_{n+1}$ and $Y^* = Y_{n+1}$ for each group. Draw samples of (X^*, Y^*) from the posterior predictive (use the samples of parameters you drew earlier) and calculate the posterior predictive probability that $X^* > Y^*$.

Hurricane intensity example

For the hurricane intensity analysis, assign a product prior $\xi(\lambda_1, \lambda_2) = \xi_1(\lambda_1)\xi_2(\lambda_2)$ where each ξ_1 and ξ_2 is the $\mathsf{Gamma}(2, 2/100)$ pdf.

TASK 4. Draw samples of (λ_1, λ_2) from the posterior for observations:

Year	Expenditures
1981-1993	2, 14, 34, 29, 42, 48, 47, 66, 12, 42, 19, 26, 17, 29, 22, 42, 35, 17, 5, 24, 23, 19, 36, 15, 27, 12, 4, 35, 46, 38, 14, 27, 54, 43, 44, 45, 74, 8, 18, 17, 63, 13, 8, 21, 13, 38, 24, 61, 8, 20, 7, 11, 47, 41, 16, 7, 21, 23, 16, 21, 10, 85, 58, 11, 125, 32, 25, 31, 59, 43, 52, 77, 52, 30, 17, 17, 5, 50, 23, 43, 36, 76, 41, 28, 16, 13, 44, 21, 19, 22, 31, 3, 68, 38, 24, 24, 19, 37, 16, 11, 81, 8, 25, 62, 12
1994-2006	32, 29, 38, 7, 21, 49, 39, 37, 10, 40, 7, 54, 74, 42, 48, 72, 22, 11, 100, 45, 33, 68, 20, 55, 2, 34, 46, 47, 28, 42, 72, 22, 61, 36, 45, 18, 53, 43, 16, 6, 39, 10, 27, 79, 41, 3, 27, 65, 25, 38, 33, 32, 56, 3, 40, 42, 41, 49, 88, 69, 11, 43, 120, 58, 12, 88, 79, 14, 48, 53, 6, 71, 122, 22, 16, 19, 24, 55, 31, 22, 109, 32, 77, 29, 36, 41, 24, 7, 30, 22, 35, 59, 55, 14, 44, 42, 26, 32, 9, 60, 13, 30, 14, 7, 11, 35, 6, 21, 40, 13, 55, 22, 50, 95, 13, 31, 46, 26, 52, 73, 28, 16, 109, 51, 75, 37, 35, 31, 15, 65, 31, 39, 36, 23, 95, 26, 9, 114, 116, 49, 54, 15, 15, 11, 13, 13, 23, 51, 82, 33, 35, 18, 64, 16, 109, 24, 75, 47, 34, 48, 72, 62, 17, 23, 15, 69, 22, 38, 16, 23, 49, 13, 28, 11, 11, 17, 18, 46, 52, 58, 82, 32

[copy paste the data on to R, don't copy manually!]
Get a central 95% posterior credible interval for $\eta = 1/\lambda_1 - 1/\lambda_2$.

Classical approach: use software

We'd also look at performing the above tasks from a classical perspective.

TASK 5. For the IV fluid analysis, an ML type approximate confidence interval for $p_1 - p_2$ can be obtained by using the R function `prop.test()` as follows:
`prop.test(x=c(193,203), n=c(309,289), conf.level=.95)$conf.int`
Run this and compare this interval with the credible interval you got earlier. Would you conclude differently for about $p_1 > p_2$ under the two results?

TASK 6. For the food data, run the classic t-test (Welch's test, covered in notes 15, we'd go over this tomorrow), to get a confidence interval for $\mu_1 - \mu_2$:
`t.test(x, y, var.equal=FALSE, conf.level=.95)$conf.int` and compare your conclusions about $\mu_1 > \mu_2$ from this interval with what you obtained before under the Bayesian analysis.

TASK 7. Unfortunately, there is no standard classical confidence interval procedure for the exponential model. So we'd just do what most people do – assume the data are normal and compare the means! Use `t.test` as in the above example to get an interval for the difference in mean intensities and compare with what you had obtained befor.