VE414 Lecture 19

Jing Liu

UM-SJTU Joint Institute

November 14, 2019

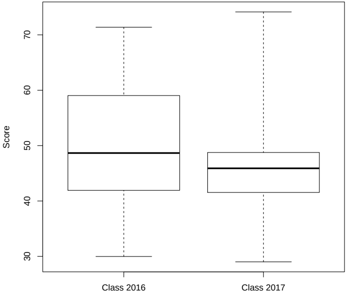
- Let us start with a Bayesian model for comparing and estimating two group means which are covered in any basic probability and statistics course.
- Consider a random sample of 31 students from my freshman course in 2016, and a random sample of 28 students from the same course but in 2017.
- Suppose we are interested in estimating θ_1, the average overall score in the 2016 course, and comparing it to θ_2, the corresponding 2017 value.
- The sample means of the overall scores for the two groups are given below

  $$\bar{x}_1 = 50.81 \qquad \text{and} \qquad \bar{x}_2 = 46.15$$

  which suggest that θ_1 is greater than θ_2.
- However, if different students had been sampled from each of the 2 groups, then perhaps $\bar{x}_2$ would have been bigger than $\bar{x}_1$.
- To assess whether or not the observed difference of $\bar{x}_1 - \bar{x}_2 = 4.66$ is large compared to the sampling variability, a Frequentist would compute

  $$\text{t-statistic} = t\left(\mathbf{x}_1, \mathbf{x}_2\right)$$

- It is the observed difference over an estimate of its standard deviation:



$$t\left(\mathbf{x}_1, \mathbf{x}_2\right) = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}}$$

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

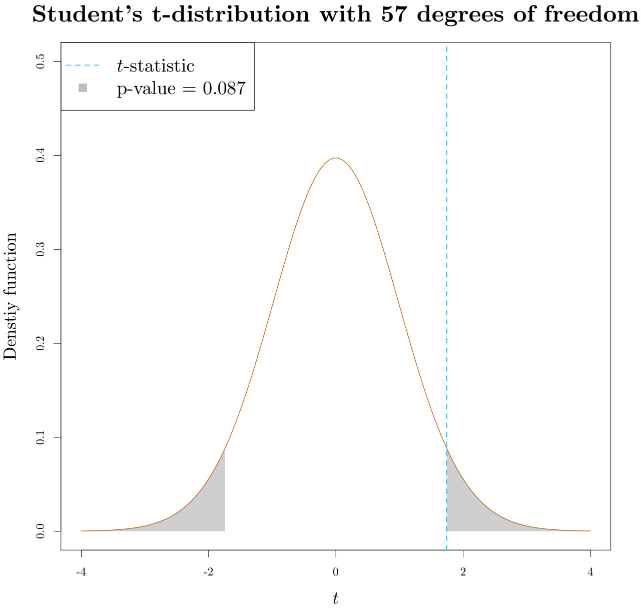
is the pooled estimate of the population variance of the two groups using the sample variance s_1^2 and s_2^2 of the two groups.

Q For our data, the t-statistic is 1.74, how should we judge whether it is big?

- You probably remember that $t(\mathbf{x}_1, \mathbf{x}_2)$ follows a t-distribution with df being

$$n_1 + n_2 - 2 = 57$$

if the scores from the two groups follows the same normal distribution.


Student's t-distribution with 57 degrees of freedom
t-statistic
▨ p-value = 0.087
Denstiy function
0.0
0.1
0.2
0.3
0.4
0.5
-4
-2
0
2
4
t

- A small p-value from the t-test is generally considered by Frequentist as an indication that θ_1 and θ_2 are not the same. Typically, we are advised to

- Reject the model that two groups have the same mean, $\theta_1 = \theta_2$, and use

$$\hat{\theta}_1 = \bar{x}_1 \qquad \text{and} \qquad \hat{\theta}_2 = \bar{x}_2 \qquad \text{provided p-value} < 0.05.$$

- Favour the model that two groups have the same mean, $\theta_1 = \theta_2$, and use

$$\hat{\theta}_1 = \hat{\theta}_2 = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{\sum_{i=1}^{n_1} x_{i1} + \sum_{i=1}^{n_2} x_{i2}}{n_1 + n_2} \quad \text{provided p-value} > 0.05.$$

- This data analysis procedure results in either treating the two population as exactly identical or treating them as completely distinct such that

no information can be retrieved from each other's sample.

- So Frequentist's estimate for θ_1 is essentially

$$\hat{\theta}_1 = w\bar{x}_1 + (1 - w)\bar{x}_2 \qquad \text{where} \quad w = \begin{cases} 1 & \text{if p-value} < 0.05, \\ n_1/(n_1 + n_2) & \text{otherwise.} \end{cases}$$

- Instead of either using or not using information from each other's sample, it might make more sense to allow w to vary continuously and depends on the relative sample sizes n_1 and n_2, the sampling variability σ^2, ect.
- A Bayesian model allows for information to be shared across the groups, and prior information about the similarities of the two populations.
- Consider the following model:

$$\begin{aligned}
X_{i1} &= \mu + \delta + \varepsilon_{i1} \\
X_{i2} &= \mu - \delta + \varepsilon_{i2} \\
\varepsilon_{ij} &\sim \text{Normal}\left(0, \sigma^2\right) \\
\mu &\sim \text{Normal}\left(\mu_0, \gamma_0^2\right) \\
\delta &\sim \text{Normal}\left(\delta_0, \tau_0^2\right) \\
\sigma^2 &\sim \text{Scaled Inverse}\, \chi^2\left(\nu_0, \sigma_0^2\right)
\end{aligned}$$

where the priors are assumed to be independent.

Q: What does each of μ and δ represent?

- It can be shown the full set of conditional posteriors are given by

$$\begin{aligned}\mu \mid \{\mathbf{X}_1, \mathbf{X}_2, \delta, \sigma^2\} &\sim \text{Normal}\left(\mu_n, \gamma_n^2\right)\\ \delta \mid \{\mathbf{X}_1, \mathbf{X}_2, \mu, \sigma^2\} &\sim \text{Normal}\left(\delta_n, \tau_n^2\right)\\ \sigma^2 \mid \{\mathbf{X}_1, \mathbf{X}_2, \mu, \delta\ \} &\sim \text{Scaled Inverse}\,\chi^2\left(\nu_n, \sigma_n^2\right)\end{aligned}$$

where the posterior parameters are given by

$$\begin{aligned}\mu_n &= \gamma_n^2\left(\frac{\mu_0}{\gamma_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n_1}(x_{i1}-\delta) + \frac{1}{\sigma^2}\sum_{i=1}^{n_2}(x_{i2}+\delta)\right)\\ \delta_n &= \tau_n^2\left(\frac{\delta_0}{\tau_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n_1}(x_{i1}-\mu) - \frac{1}{\sigma^2}\sum_{i=1}^{n_2}(x_{i2}-\mu)\right)\\ \gamma_n^2 &= \left(\frac{1}{\gamma_0^2} + \frac{n_1+n_2}{\sigma^2}\right)^{-1};\ \tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n_1+n_2}{\sigma^2}\right)^{-1};\ \nu_n = \nu_0 + n_1 + n_2\\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + \sum_{i=1}^{n_1}\big(x_{i1} - (\mu+\delta)\big)^2 + \sum_{i=1}^{n_2}\big(x_{i2} - (\mu-\delta)\big)^2\end{aligned}$$

- If we consider extreme cases for the prior parameters:
- If $\nu_0 = 0$, which leads to using Jeffrey's prior for $\varphi_{\sigma^2} \propto \sigma^{-2}$, then

$$\sigma_n^2 = \frac{1}{\nu_0 + n_1 + n_2}\left(\nu_0\sigma_0^2 + \sum_{i=1}^{n_1}\big(x_{i1} - (\mu+\delta)\big)^2 + \sum_{i=1}^{n_2}\big(x_{i2} - (\mu-\delta)\big)^2\right)$$
$$= \frac{1}{n_1 + n_2}\left(\sum_{i=1}^{n_1}\big(x_{i1} - (\mu+\delta)\big)^2 + \sum_{i=1}^{n_2}\big(x_{i2} - (\mu-\delta)\big)^2\right)$$

which is the MLE of the variance given the values of μ and δ.

- If $\mu_0 = \delta_0 = 0$ and $\gamma_0^2 = \tau_0^2 = \infty$, i.e. uniform priors, $\mu \propto 1$ and $\delta \propto 1$, then

$$\mu_n = \frac{1}{n_1 + n_2}\left(\sum_{i=1}^{n_1}(x_{i1} - \delta) + \sum_{i=1}^{n_2}(x_{i2} + \delta)\right)$$
$$\delta_n = \frac{1}{n_1 + n_2}\left(\sum_{i=1}^{n_1}(x_{i1} - \mu) - \sum_{i=1}^{n_2}(x_{i2} - \mu)\right)$$

which are the corresponding MLEs, respectively.

- Suppose SJTU requires us to adjust the scores reasonably so that the scores have a mean of 50 and a standard deviation of 10 at the university level.
- So we could base our priors on this information, for example,

$$\begin{aligned} \mu &\sim \text{Normal}\left(\mu_0 = 50, \gamma_0^2 = 25^2\right) \\ \sigma^2 &\sim \text{Scaled Inverse}\, \chi^2\left(\nu_0 = 4, \sigma_0^2 = 10^2\right) \end{aligned}$$

- The prior parameters are chosen to match the prior expectation,

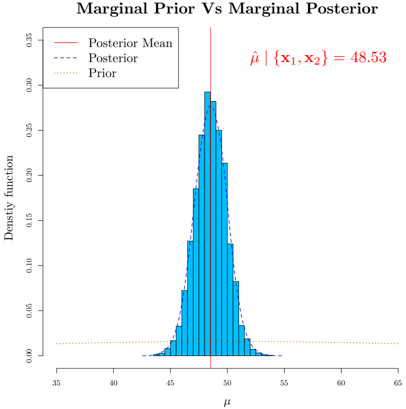
$$\mathbb{E}\left[\mu\right] = \mu_0 = 50 \qquad \text{and} \qquad \mathbb{E}\left[\sigma^2\right] = \frac{\nu_0 \sigma_0^2}{\nu_0 - 2} = 10^2$$

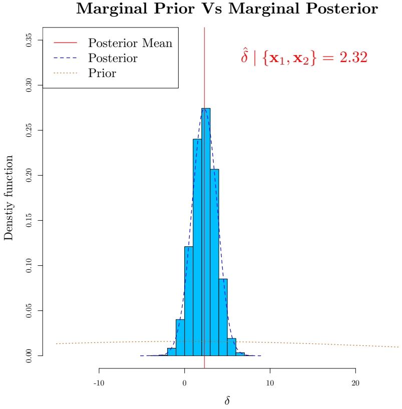
- For the prior distribution δ, the following is chosen

$$\delta \sim \text{Normal}\left(\delta_0 = 0, \tau_0^2 = 25^2\right)$$

  to reflect the prior opinion that $\theta_1 > \theta_2$ and $\theta_2 > \theta_1$ are equally probable.
- γ_0^2 and τ_0^2 are reasonably big to reflect that we are reasonably ignorant.

- Since the full set of conditional posteriors are available, we can construct a Gibbs sampler to obtain the joint posterior and the marginal posteriors.


Marginal Prior Vs Marginal Posterior


Marginal Prior Vs Marginal Posterior

- The marginal posteriors are much more spiky than their corresponding priors.

- Instead of using the following as an estimate of θ_1,

$$\hat{\theta}_1 = w\bar{x}_1 + (1-w)\bar{x}_2 \qquad \text{where} \quad w = \begin{cases} 1 & \text{if p-value} < 0.05, \\ n_1/(n_1+n_2) & \text{otherwise.} \end{cases}$$

  a Bayesian estimate of θ_1 is given by the sum of marginal posterior means

$$\hat{\theta}_1 = \hat{\mu} \mid \{\mathbf{x}_1, \mathbf{x}_2\} + \hat{\delta} \mid \{\mathbf{x}_1, \mathbf{x}_2\} = 48.53 + 2.32 = 50.85$$

  where the marginal posterior means are estimated using sample means.

- The posterior probability

$$\Pr\left(\theta_1 > \theta_2 \mid \mathbf{x}_1, \mathbf{x}_2\right) = \Pr\left(\delta > 0 \mid \mathbf{x}_1, \mathbf{x}_2\right)$$

  can be estimated using our samples, that is,

$$\Pr\left(\theta_1 > \theta_2 \mid \mathbf{x}_1, \mathbf{x}_2\right) \approx \frac{\text{Number of samples such that } \delta^{(t)} > 0}{\text{Total number of samples}} = 0.96$$

  whereas the prior probability $\Pr(\delta > 0) = 0.5$, so we may conclude $\theta_1 > \theta_2$.

- Notice the posterior probability

$$\Pr\left(\theta_1 > \theta_2 \mid \mathbf{x}_1, \mathbf{x}_2\right) \approx 0.96$$

is not the posterior probability that a randomly selected student from class 2016 has a higher score than one taken from class 2017.

$$\Pr\left(\theta_1 > \theta_2 \mid \mathbf{x}_1, \mathbf{x}_2\right) \neq \Pr\left(X_1 > X_2 \mid \mathbf{x}_1, \mathbf{x}_2\right)$$

Q: How can we obtained this latter probability given our samples?

- It can be obtained from the joint posterior predictive distribution,

$$\Pr\left(X_1 > X_2 \mid \mathbf{x}_1, \mathbf{x}_2\right) \approx 0.62$$

which is computed by first generating a sample of $\{X_1, X_2\}$ values, one pair for each set of $\{\mu^{(t)}, \delta^{(t)}, \sigma^{2^{(t)}}\}$ in the sample generated by Gibbs, then using

$$\frac{\text{Number of samples such that } X_1^{(t)} > X_2^{(t)}}{\text{Total number of samples}} = 0.62$$