Lab 4: Do a teacher's expectations influence student achievement?

Lei Qian and Rebecca C. Steorts

July 28, 2020

Let's first load packages that we'll need in this assignment and also load the data.

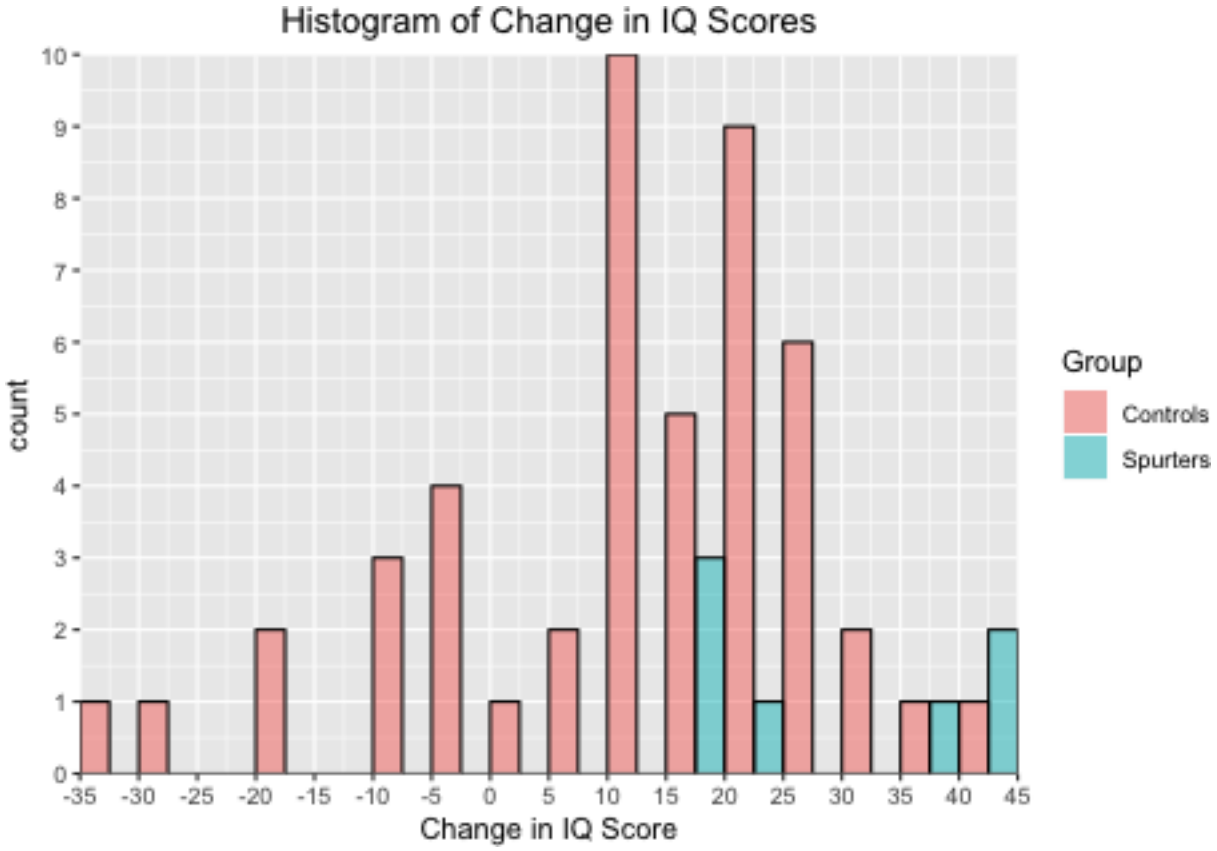
```
# input data
# spurters
x = c(18, 40, 15, 17, 20, 44, 38)
# control group
y = c(-4, 0, -19, 24, 19, 10, 5, 10,
      29, 13, -9, -8, 20, -1, 12, 21,
      -7, 14, 13, 20, 11, 16, 15, 27,
      23, 36, -33, 34, 13, 11, -19, 21,
      6, 25, 30,22, -28, 15, 26, -1, -2,
      43, 23, 22, 25, 16, 10, 29)
# store data in data frame
iqData = data.frame(Treatment = c(rep("Spurters", length(x)),
                                  rep("Controls", length(y))),
                                  Gain = c(x, y))
```

Task 1

Plot histograms for the change in IQ score for the two groups. Report your findings.

```
xLimits = seq(min(iqData$Gain) - (min(iqData$Gain) %% 5),
              max(iqData$Gain) + (max(iqData$Gain) %% 5),
              by = 5)

ggplot(data = iqData, aes(x = Gain, fill = Treatment, colour = I("black"))) +
  geom_histogram(position = "dodge", alpha = 0.5, breaks = xLimits, closed = "left")+
  scale_x_continuous(breaks = xLimits,
                     expand = c(0,0))+
  scale_y_continuous(expand = c(0,0),
breaks = seq(0, 10, by = 1))+
  ggtitle("Histogram of Change in IQ Scores") + labs(x = "Change in IQ Score",
                                                     fill = "Group") +
                                                theme(plot.title = element_text(hjust = 0.5))
```



From the histograms, I know that the randomly selected "spurters" group has a different distribution than the "controls" group. This could indicate that teachers being told that a specific group of students is expected to perform particularly well will pay more attention and time on that group and resulting in more improvement over the year.

Task 2

How strongly does this data support the hypothesis that the teachers? expectations caused the spurters to perform better than their classmates? IQ tests are purposefully calibrated to make the scores normally distributed, so it makes sense to use a normal model here:

$$X_1, \ldots, X_{n_s} \mid \mu_s, \lambda_s^{-1} \overset{iid}{\sim} \text{Normal}(\mu_S, \lambda_S^{-1})$$
$$Y_1, \ldots, Y_{n_C} \mid \mu_c, \lambda_c^{-1} \overset{iid}{\sim} \text{Normal}(\mu_C, \lambda_C^{-1}).$$

We are interested in the difference between the means—in particular, is $\mu_S > \mu_C$? We don't know the standard deviations $\sigma_S = \lambda_S^{-1/2}$ and $\sigma_C = \lambda_C^{-1/2}$, and the sample seems too small to estimate them very well.

It is easy using a Bayesian approach. We just need to compute the posterior probability that $\mu_S > \mu_C$:

$$\Pr(\mu_S > \mu_C \mid x_{1:n_S}, y_{1:n_C}).$$

Let's assume independent Normal-Gamma priors:

$$\text{spurters: } (\mu_S, \lambda_S) \sim \text{NormalGamma}(m, c, a, b)$$
$$\text{controls: } (\mu_C, \lambda_C) \sim \text{NormalGamma}(m, c, a, b)$$

with the following hyperparameter settings, based on subjective prior knowledge:

- $m = 0$ (Don't know whether students will improve or not, on average.)
- $c = 1$ (Unsure about how big the mean change will be—prior certainty in our choice of m assessed to be equivalent to one datapoint.)
- $a = 1/2$ (Unsure about how big the standard deviation of the changes will be.)
- $b = 10^2 a$ (Standard deviation of the changes expected to be around $10 = \sqrt{b/a} = E(\lambda)^{-1/2}$.)

In order to solve this task, let's recall that the likelihoods are gaussians:

Likelihood Functions

$$p(X_1 ... X_{n_s} | \mu_s, \lambda_s^{-1}) = \prod_{i=1}^{n_s} \frac{1}{\sqrt{2{\sigma_s}^2 \pi}} e^{-\frac{(x_i - \mu_s)^2}{2{\sigma_s}^2}} = \prod_{i=1}^{n_s} \frac{1}{\sqrt{2{\lambda_s}^{-1} \pi}} e^{-\frac{\lambda_s (x_i - \mu_s)^2}{2}}$$

$$p(Y_1 ... Y_{n_c} | \mu_c, \lambda_c^{-1}) = \prod_{i=1}^{n_c} \frac{1}{\sqrt{2{\sigma_c}^2 \pi}} e^{-\frac{(y_i - \mu_c)^2}{2{\sigma_c}^2}} = \prod_{i=1}^{n_c} \frac{1}{\sqrt{2{\lambda_c}^{-1} \pi}} e^{-\frac{\lambda_c (y_i - \mu_c)^2}{2}}$$

Priors

Note that the priors are Normal-Gammas.

$$p(\mu_s, \lambda_s | m, c, a, b) = \frac{b^a \sqrt{c}}{\Gamma(a)\sqrt{2\pi}} \lambda_s^{a-0.5} e^{-b\lambda_s} e^{-\frac{c\lambda_s(\mu_s - m)^2}{2}}$$

$$p(\mu_c, \lambda_c | m, c, a, b) = \frac{b^a \sqrt{c}}{\Gamma(a)\sqrt{2\pi}} \lambda_c^{a-0.5} e^{-b\lambda_c} e^{-\frac{c\lambda_c(\mu_c - m)^2}{2}}$$

Posterior Distribution

Recall that from the course notes, that the posterior distribution of the likelihood and prior is an updated Normal Gamma, which has the following form:

$$\begin{aligned}(\mu_s, \lambda_s) | x_{1:n_s} \sim & \text{NormalGamma}\left(m' = \frac{cm + n_s \bar{x}}{c + n_s}, c' = c + n_s, a' = a + \frac{n_s}{2}, b' = b + \frac{1}{2}\sum_{i=1}^{n_s}(x_i - \bar{x})^2 + \frac{n_s c}{c + n_s}\frac{(\bar{x} - m)^2}{2}\right) \\ &= \text{NormalGamma}(24, 8, 4, 855) \\ (\mu_c, \lambda_c) | y_{1:n_c} \sim & \text{NormalGamma}\left(m^* = \frac{cm + n_c \bar{y}}{c + n_c}, c^* = c + n_c, a^* = a + \frac{n_c}{2}, b^* = b + \frac{1}{2}\sum_{i=1}^{n_c}(y_i - \bar{y})^2 + \frac{n_c c}{c + n_c}\frac{(\bar{y} - m)^2}{2}\right) \\ &= \text{NormalGamma}(11.8, 49, 24.5, 6344)\end{aligned}$$

Corresponding Code

```
prior = data.frame(m = 0, c = 1, a = 0.5, b = 50)
findParam = function(prior, data){
  postParam = NULL
  c = prior$c
  m = prior$m
  a = prior$a
  b = prior$b
  n = length(data)
```

```
  postParam = data.frame(m = (c*m + n*mean(data))/(c + n),
                c = c + n,
                a = a + n/2,
                b =  b + 0.5*(sum((data - mean(data))^2)) +
                  (n*c *(mean(data)- m)^2)/(2*(c+n)))
  return(postParam)
}
postS = findParam(prior, x)
postC = findParam(prior, y)
```

% latex table generated in R 3.6.2 by xtable 1.8-4 package % Tue Jul 28 17:17:37 2020

	m	c	a	b
prior	0.00	1.00	0.50	50.00
Spurters Posterior	24.00	8.00	4.00	855.00
Controls Posterior	11.80	49.00	24.50	6343.98

Table 1: Parameters

Task 3

Based on the calculations from the previous task, provide a scatterplot of samples from the posterior distributions for the two groups. What are your conclusions?

```
# sampling from two posteriors

# Number of posterior simulations
sim = 1000

# initialize vectors to store samples
mus = NULL
lambdas = NULL
muc = NULL
lambdac = NULL

# Following formula from the NormalGamma with
# the update paramaters accounted accounted for below

lambdas = rgamma(sim, shape = postS$a, rate = postS$b)
lambdac = rgamma(sim, shape = postC$a, rate = postC$b)


mus = sapply(sqrt(1/(postS$c*lambdas)),rnorm, n = 1, mean = postS$m)
muc = sapply(sqrt(1/(postC$c*lambdac)),rnorm, n = 1, mean = postC$m)

# Store simulations
simDF = data.frame(lambda = c(lambdas, lambdac),
                   mu = c(mus, muc),
                   Treatment = rep(c("Spurters", "Controls"),
                                   each = sim))

simDF$lambda = simDF$lambda^{-0.5}

# Plot the simulations
```

```
ggplot(data = simDF, aes(x = mu, y = lambda, colour = Treatment, shape = Treatment)) +
  geom_point(alpha = 0.2) +
  labs(x = expression(paste(mu, " (Mean Change in IQ Score)")),
       y = expression(paste(lambda^{-1/2}, " (Std. Dev. of Change)")))  +
  ggtitle("Posterior Samples")+
  theme(plot.title = element_text(hjust = 0.5))
```

The simulated scatterplot does look similar to Figure 1 in that the control group is more concentrated with a smaller average mean change in IQ score, while the spurters group has a larger average mean change in IQ score.