Lab4

Andrea Avenia

2023-01-30

```r
library(ggplot2)
```

Review

Let

$$\begin{aligned}\{X_j\}_{j=1}^n|\mu,\lambda &\overset{ind}{\sim} \mathcal{N}(\mu, 1/\lambda)\\ \mu|\lambda &\sim \mathcal{N}(\mu_0, 1/(n_0\lambda))\\ \lambda &\sim \text{Ga}(n_1/2, n_1/(2\lambda_1))\end{aligned}$$

Then, we say that $(\mu,\lambda) \sim \text{NG}(\mu_0, n_0, n_1/2, n_1/(2\lambda_1))$, and then

$$(\mu,\lambda)|\{X_j\}_{j=1}^n \sim \text{NG}\left(\frac{n\bar{X}+n_0\mu_0}{n+n_0}, n+n_0, \frac{n+n_1}{2}, \frac{1}{2}\left(\frac{n_1}{\lambda_1} + \frac{nn_0}{n+n_0}(\bar{x}-\mu_0)^2 + \sum_{j=1}^n (x_j-\bar{x})^2\right)\right)$$

For the first two parameters, we pass from (μ_0, n_0) to $(\frac{n\bar{X}+n_0\mu_0}{n+n_0}, n+n_0)$, thus, if we interpret μ_0 and n_0 as the mean and the sample size of some pseudo-observations, then, we update them by simply adding the real observations to our pseudo-observations.\ For the other two parameters, we go from $(n_1, n_1/\lambda_1)/2$ to $(n+n_1, n_1/\lambda_1 + \frac{nn_0}{n+n_0}(\bar{x}-\mu_0)^2 + \sum_{j=1}^n (x_j-\bar{x})^2)/2$. In this case we imagine of having a different set of pseudo-observations, of sample size n_1 and with sample precision to λ_1. We update these parameters, by going from n_1 to n_1+n and by updating in a more complicated way the second parameter.\ This prior is proper as long as n_0, n_1 and λ_1 are positive. When n_0 is close tho zero, we are very unsure about the mean, and when n_1 is close to zero, we are unsure about the precision. In this formulation, μ_0 and λ_1 are our prior guess on the mean and on the precision (the reciprocal of the variance).

Usually, we call $a = n_1/2$ and $b = n_1/(2\lambda_1)$, so that $2a = n_1$ and $a/b = \lambda_1$. Here how to generate samples from the prior

```r
rNG=function(k, m0, n0, a, b){
  lambda=rgamma(k, a, b)
  mu=rnorm(k,m0,sqrt(1/(n0*lambda)))
  return(rbind(mu, lambda))
}
rNG(3, 0, 1, 1, 1)
```

```
##             [,1]       [,2]       [,3]
## mu     0.1499295 0.02965108 -0.7936861
## lambda 1.3093738 4.31668030  0.9702188
```

Here how to update the parameters

```r
NG.update=function(m0,n0,a,b, X){
  n=length(X)
  mX=sum(X)/n
  ssX=sum((X-mX)^2)
  m0.post=(n*mX+n0*m0)/(n+n0)
  n0.post=n+n0
  a.post=a+n/2
  b.post=b+(ssX+(n*n0)/(n+n0)*(mX-m0)^2)/2
  return(list(m0.post=m0.post,n0.post=n0.post,a.post=a.post,b.post=b.post))
}
```

If now the true mean is 1 and the true precision is 3, and our prior guess are 0 for the mean and 1 for the precision with psudo samples sizes equal to 1, let's see if we can recover the truth with different sample sizes

```
N=1000

mu.true=1
lambda.true=3
m0=0
n0=1
lambda1=1
n1=1
a=n1/2
b=n1/(2*lambda1)

X = rnorm(N, mu.true, sd=sqrt(1/lambda.true))

update=NG.update(m0, n0, a, b, X)

mu.post=update$m0.post
lambda.post=update$a.post/update$b.post

c(mu.true,mu.post)
```

```
## [1] 1.000000 1.020633
```

```
c(lambda.true,lambda.post)
```

```
## [1] 3.000000 2.939667
```

Now a function to sample from the posterior

```
rNG.post=function(k,m0,n0,a,b,X){
  update=NG.update(m0, n0, a, b, X)
  m0.post=update$m0.post
  n0.post=update$n0.post
  a.post=update$a.post
  b.post=update$b.post
  return(rNG(k,m0.post,n0.post,a.post,b.post))
}

rNG.post(10,m0, n0, a, b, X)
```

```
##             [,1]     [,2]      [,3]     [,4]     [,5]     [,6]     [,7]
## mu     0.9957161 1.010535 0.9848515 1.016472 1.019170 1.029487 1.028048
## lambda 2.9295721 3.082456 2.8475105 2.949945 2.986747 3.125025 3.148576
##             [,8]     [,9]     [,10]
## mu     0.9780372 1.018184 0.9684089
## lambda 2.9637176 2.888547 3.1230637
```

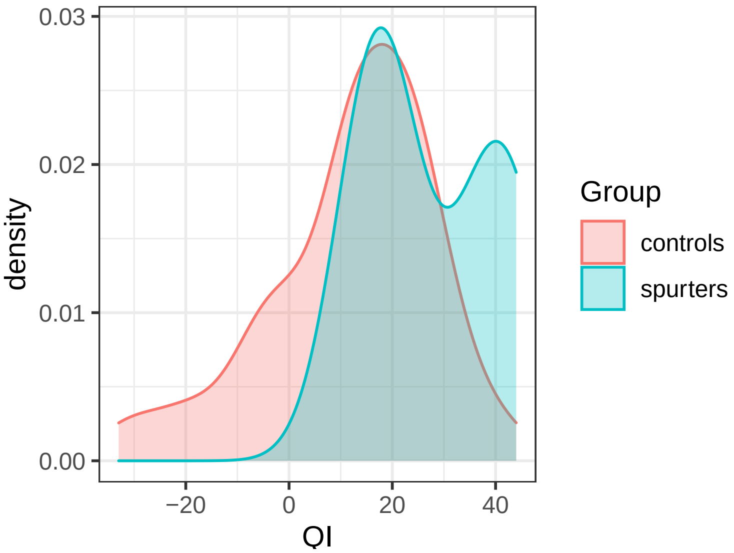
Lab4

Do teacher's expectations influence student achievement? - Students had an IQ test at the begining and end of a year; the data is the difference in IQ score. - 20% of the students were randomly chosen; their teacher was told they were "spurters" (high performers)

```
spurters = c(18, 40, 15, 17, 20, 44, 38)
controls = 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)
```

Task 1: Plot histograms for the change in IQ score for the two groups. Report your findings.

```
data=data.frame(QI=c(spurters,controls),Group=as.factor(c(rep("spurters",7),rep("controls",48))))

ggplot(data,aes(x=QI,col=Group,fill=Group))+
  geom_density(alpha=0.3)+
  theme_bw()
```



Task 2: How strongly does this data support the hypothesis that the teachers expectations caused the spurters to perform better than their classmates?

Let's use a normal model:

$$X_1, \ldots, X_{n_s} \mid \mu_s, \lambda_s^{-1} \overset{iid}{\sim} \text{Normal}(\mu_S, \lambda_S^{-1})$$

$$Y_1, \ldots, Y_{n_C} \mid \mu_c, \lambda_c^{-1} \overset{iid}{\sim} \text{Normal}(\mu_C, \lambda_C^{-1}).$$

We are interested in the difference between the means—in particular, is $\mu_S > \mu_C$?

We can answer this by computing the posterior probability that $\mu_S > \mu_C$:

$$\mathbb{P}[\mu_S > \mu_C | x_{1:n_S}, y_{1:n_C}] = \mathbb{E}[\mathbf{1}_{\mu_S > \mu_C} | x_{1:n_S}, y_{1:n_C}].$$

Let's assume independent Normal-Gamma priors:

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

Subjective choice:

- $\mu_0 = 0$ Don't know whether students will improve or not, on average
- $n_0 = 1$ Weakly informative prior; pseudo sample size equal to 1
- $n_1 = 1$ Weakly informative prior; pseudo sample size equal to 1
- $\lambda_1 = 1/10^2$ We expect the standard deviation to be around 10.
- Thus, $a = 1/2$, $b = 50$.

```
m = 1
c = 1
a = 1
b = 1
```

Now let's sample from the posterior distributions.

```
k=10000
spurters.sampled =
  rNG.post(k, m, c, a, b, spurters)
controls.sampled =
  rNG.post(k, m, c, a, b, controls)
```

Using the Monte-Carlo approximation

$$\mathbb{P}(\mu_S > \mu_C \mid x_{1:n_S}, y_{1:n_C}) = \mathbb{E}[\mathbf{1}_{\mu_S > \mu_C} | x_{1:n_S}, y_{1:n_C}] \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\mu_S^{(i)} > \mu_C^{(i)}},$$

we find

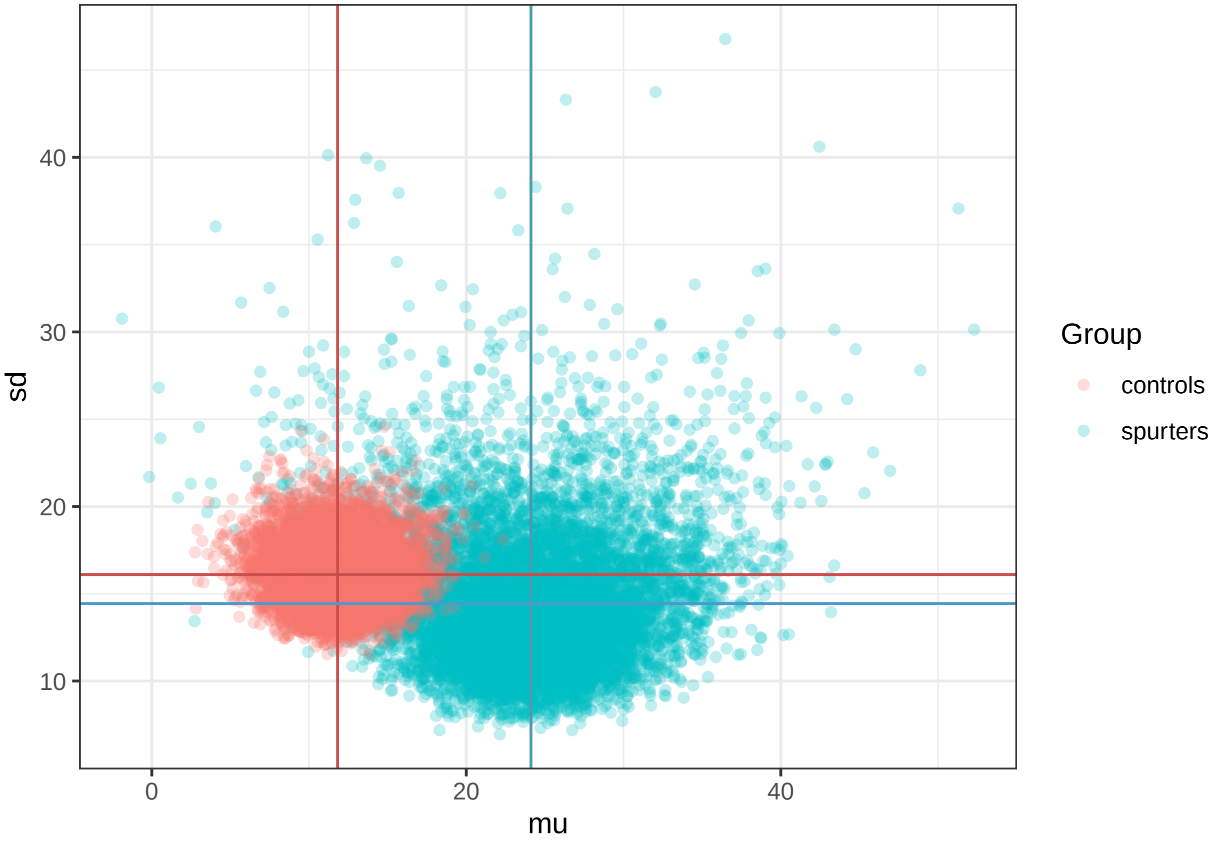
```
mean(spurters.sampled["mu",]>controls.sampled["mu",])
```

```
## [1] 0.9795
```

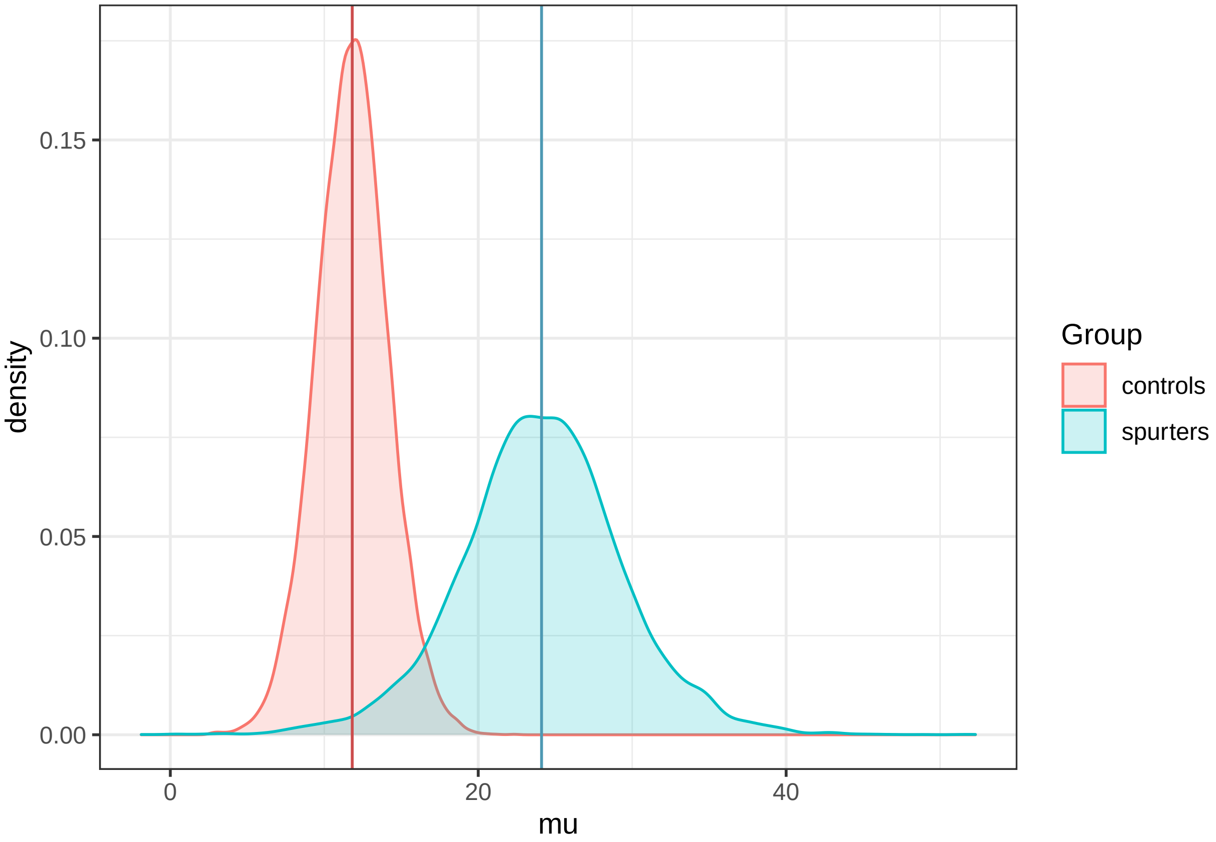
Task 3: Provide a scatterplot of samples from the posterior distributions for the two groups. What are your conclusions?

```
sp.sd.mean=mean(spurters.sampled[2,]^(-1/2))
co.sd.mean=mean(controls.sampled[2,]^(-1/2))
sp.sd.sd=sd(spurters.sampled[2,]^(-1/2))
co.sd.sd=sd(controls.sampled[2,]^(-1/2))

sp.mu.mean=mean(spurters.sampled[1,])
co.mu.mean=mean(controls.sampled[1,])
sp.mu.sd=sd(spurters.sampled[1,])
co.mu.sd=sd(controls.sampled[1,])
```

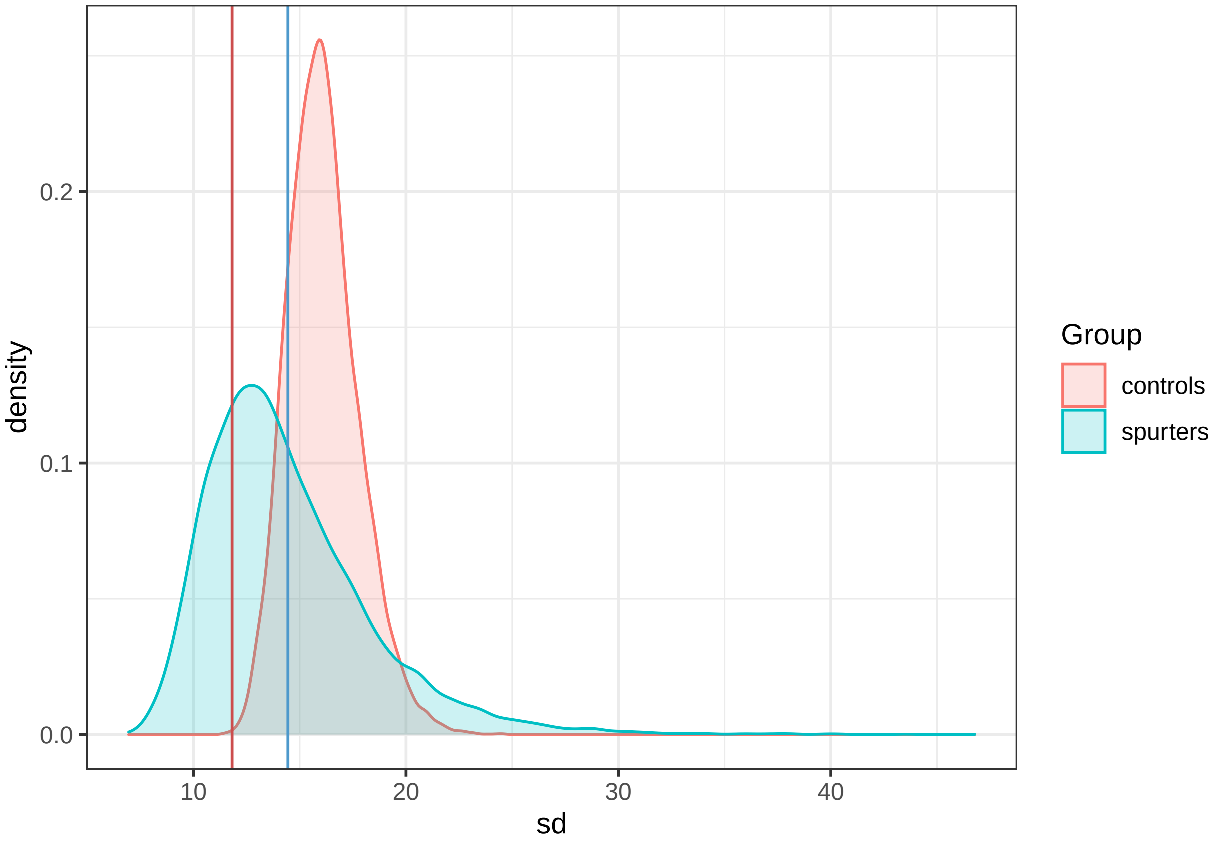
```
dd=data.frame(mu=c(spurters.sampled[1,],controls.sampled[1,]),
              sd=c(spurters.sampled[2,]^(-1/2),controls.sampled[2,]^(-1/2)),
              Group=as.factor(c(rep("spurters",k),rep("controls",k))))

ggplot(data=dd,aes(x=mu,y=sd,col=Group))+
  geom_point(alpha=0.25)+
  geom_vline(xintercept=sp.mu.mean,col=rgb(0.3,0.6,0.7))+
  geom_vline(xintercept=co.mu.mean,col=rgb(0.8,0.3,0.3))+
  geom_hline(yintercept=sp.sd.mean,col=rgb(0.3,0.6,0.8))+
  geom_hline(yintercept=co.sd.mean,col=rgb(0.8,0.3,0.3))+
  #xlim(0,50)+ylim(0,50)+
  theme_bw()
```



```
ggplot(data=dd,aes(x=mu,col=Group,fill=Group))+geom_density(alpha=0.2)+
  geom_vline(xintercept=sp.mu.mean,col=rgb(0.3,0.6,0.7))+
  geom_vline(xintercept=co.mu.mean,col=rgb(0.8,0.3,0.3))+
  theme_bw()
```



```
ggplot(data=dd,aes(x=sd,col=Group,fill=Group))+geom_density(alpha=0.2)+
  geom_vline(xintercept=sp.sd.mean,col=rgb(0.3,0.6,0.8))+
  geom_vline(xintercept=co.mu.mean,col=rgb(0.8,0.3,0.3))+
  theme_bw()
```


density
0.2
0.1
0.0
10
20
30
40
sd
Group
controls
spurters