ST559 Homework 5

Nick Sun

May 20, 2020

Abstract

We have two problems related to Bayesian ANOVA and Bayesian χ^2 tests.

1 Q1

We have the following data from Politifact:

Name	True	Mostly true	Half true	Mostly false	False	Pants on fire
Paul Ryan	10	15	18	20	7	3
Nancy Pelosi	4	3	11	5	7	3

Let's use Bayesian methods to analyze this data.

assumptions Some assumptions that we will have to make is that Nancy Pelosi and Paul Ryan's statements are independent from one another (this is a stretch) and that both of their statements follow a multinomial distribution where every statement they make has some probability p_i of following into the ith category.

prior distribution A reasonable prior distribution would be the Dirichlet distribution since the Dirichlet is the conjugate prior of the multinomail. We will probably want a noninformative prior, so our prior then for both Nancy and Paul is $Dirichlet(1/2, \dots, 1/2)$ which is a distribution with 6 parameters, one for each Politifact category.

likelihood Our likelihood function for each speaker's statements then is

$$\begin{aligned} p(p_1, p_2, \dots, p_6|data) &\propto p(data|p_1, \dots, p_6)\pi(p_1, \dots, p_6) \\ &\propto p_1^{y_1} \dots p_n^{y_n} \prod_{i=1}^{n} p_i^{a_i - 1} \\ &\propto p_1^{y_1 + a_1 - 1} \dots p_6^{y_6 + a_6 - 1} \end{aligned}$$

where again $a_i = 1/2, \forall i$.

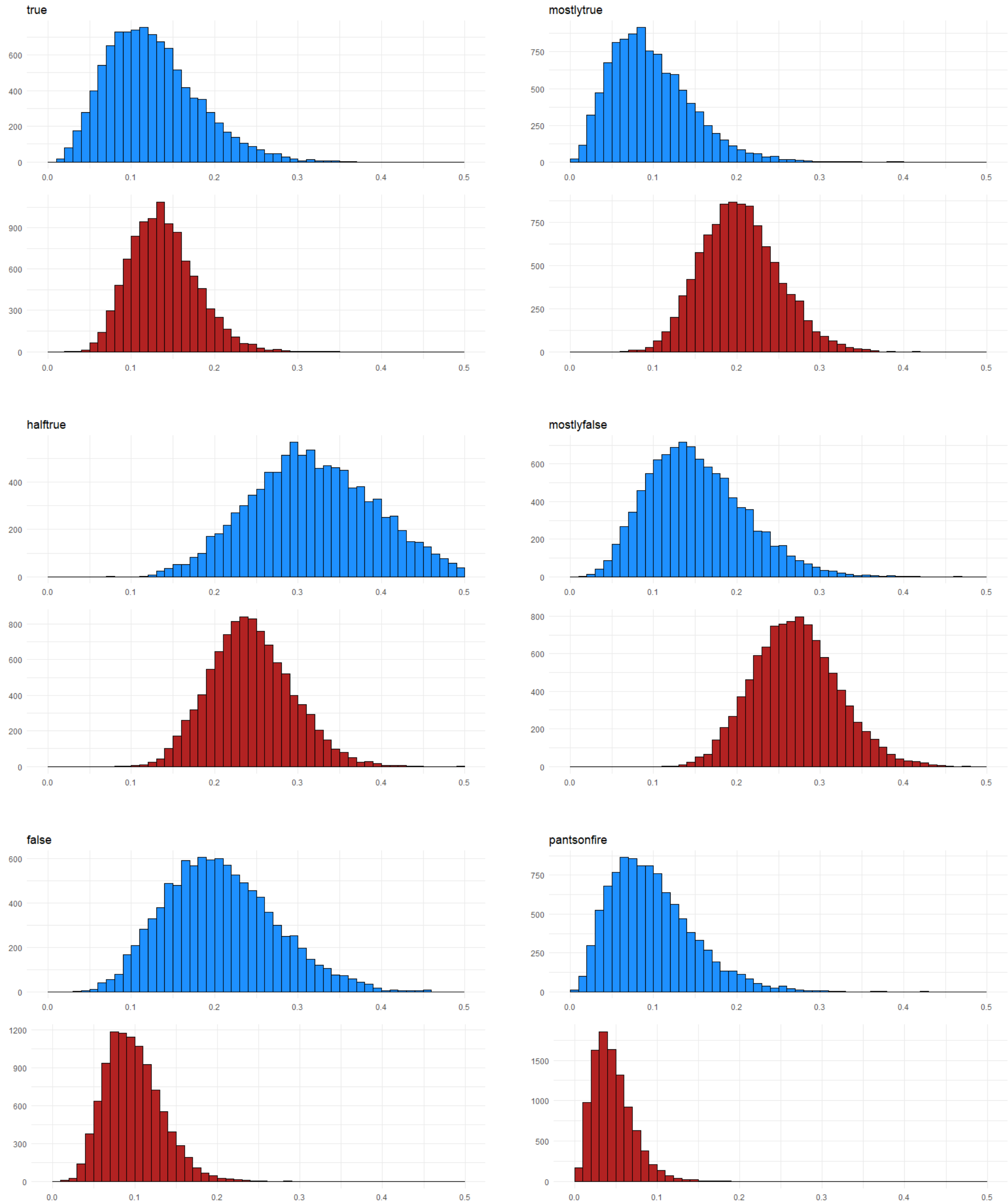
Therefore, the posterior distribution for Nancy Pelosi is

$$p(p_1, \dots, p_n|data) = Dirichlet(4.5, 3.5, 11.5, 5.5, 7.5, 3.5)$$

and our posterior distribution for Paul Ryan is

$$p(p_1, \dots, p_n|data) = Dirichlet(10.5, 15.5, 18.5, 20.5, 7.5, 3.5)$$

results After drawing many observations from these Dirichlet posterior distributions, we find the following marginal posterior distributions. The blue histograms are for Nancy Pelosi's posterior distributions and the red histograms are for Paul Ryan.



Given the marginal posterior distributions, we can make the following rough conclusions:

1. Nancy and Paul have roughly the same probability of making a **true** statement (median around 12%)

2. Paul has a greater probability of making **mostly true** statements

3. Nancy has a greater probability of making **half true** statements

4. Paul has a greater probability of making **mostly false** statements

5. Nancy has a greater probability of making **false** statements
6. Nancy has a greater probability of making **pants on fire** statements

This analysis isn't perfect though. If anything about this makes us cautious, it is that Nancy has a lot less statements on Politifact than Paul Ryan did (Paul has 73 statements on Politifact whereas Nancy only has 33). We probably want a more even sample size between the two speakers for future analyses.

2 Q2

We have the following data from *Bayesian Data Analysis, pg 292*:

Machine	Measurements
1	83,92,92,46,67
2	117,109,114,104,87
3	101,93,92,86,67
4	105,119,116,102,116
5	79,97,103,79,92
6	57,92,104,77,100

2.1 part a.

Perform a Bayesian ANOVA for this dataset. Write down the details of the model, prior distributions, and posterior distributions.

answer In this problem, we have 6 machines with 5 observations for each. If we consider $y_{i,j}$ to be the jth measurement from machine i, we can write this as:

$$y_{i,j} \sim N(\mu_i, \sigma^2)$$

where $i = 1, \ldots, 6$ and $j = 1, \ldots, 5$.

Our prior will be

$$\pi(\mu_1, \mu_2, \ldots, \sigma^2) \propto \frac{1}{\sigma^2}$$

Our posterior distributions will be

$$\left.\frac{n(k-1)MS_w}{\sigma^2}\right| data \sim \chi^2_{n(k-1)}$$

$$\left.\frac{\sqrt{k}(\theta_i - \bar{y}_{i.})}{\sqrt{MS_w}}\right| data = \frac{z_i}{\nu}$$

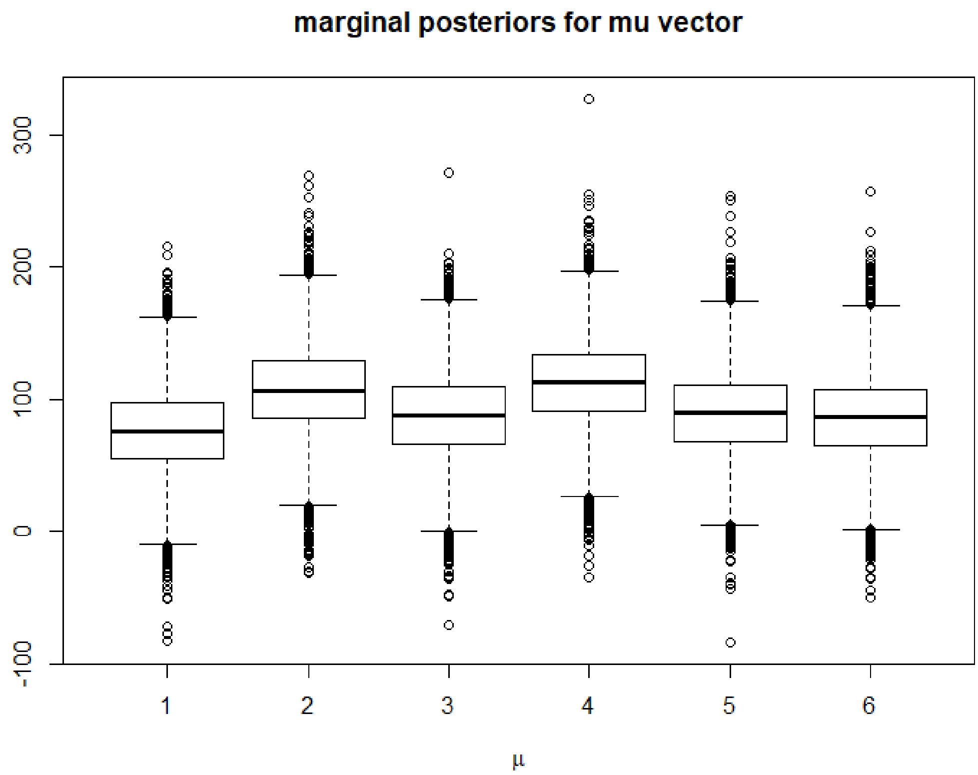
$$\text{where } MS_w = \frac{1}{n(k-1)} \sum_i^n \sum_j^k (y_{ij} - \bar{y}_{i.})^2$$

We calculate $MS_w = 4926.8$, therefore, we have

$$\left.\frac{4926.8}{\sigma^2}\right| \sim \chi^2_{24}$$

For the posteriors of $\mu_1, \ldots, \mu_6$, we see that the marginal distributions will be centered around the $\bar{y}_{i.}$ with a scale factor of $\sqrt{MS_w/k}$. This gives us a multivariate t distribution with $\vec{\mu} = (76, 106.2, 87.8, 111.6, 90.0, 86.0)$ and scale factor of 31.39.

If we randomly draw from this posterior distribution, we get the following marginal posteriors for $\mu_1, \ldots \mu_6$.



We see that there doesn't appear to be any significant difference in the centers of each marginal distribution, suggesting that there is no difference between the machines.

2.2 part b.

Compare these results with classical ANOVA. What conclusions do you draw?

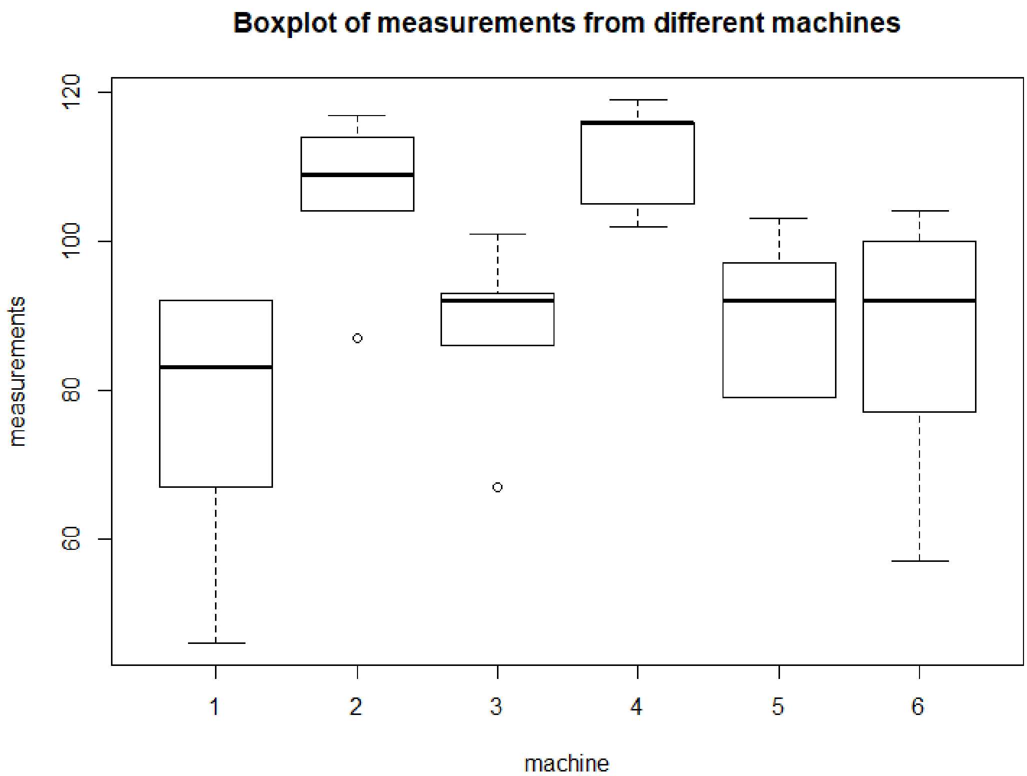
answer Classical ANOVA gives us similar conclusions! We do not have significant evidence against the null hypothesis that $\mu_1 = \ldots = \mu_6$. Here is the console output from the frequentist ANOVA tests.

```
> machines.aov <- aov(measurements ~ machine, data = machines)
> summary(machines.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
machine      1     45    45.4   0.136  0.715
Residuals   28   9353   334.0
```

2.3 part c.

Is the equal variance assumption reasonable for this dataset?

answer I don't think so. If we look at the boxplot for each machine's measurement data, there seems to be some heteroskedasticity.



2.4 part d.

Provide both R and Rstan code for part a.

I didn't end up using Rstan for this problem, I just used code that we discussed in lecture.

```
machine <- c(rep(1, 5), rep(2, 5), rep(3, 5), rep(4, 5), rep(5, 5),
    rep(6, 5))
measurements <- c(83, 92, 92, 46, 67,
                  117, 109, 114, 104, 87,
                  101, 93, 92, 86, 67,
                  105, 119, 116, 102, 116,
                  79, 97, 103, 79, 92,
                  57, 92, 104, 77, 100)
machines <- data.frame(machine, measurements)
boxplot(measurements ~ machine, data = machines,
        main = "Boxplot of measurements from different machines")

# Bayesian methods
machines %>%
    group_by(machine) %>%
    mutate(
        ybar = mean(measurements)
    ) -> machines
```

```
msw <- sum((machines$measurements - machines$ybar)^2)
y.df <- 24

prec <- rchisq(10000, y.df)/y.df/msw
hist(1/prec,
     main = "posterior of sigma^2",
     xlab = expression(sigma^2))

v <- rchisq(10000, y.df)/y.df
theta.i <- matrix(0, 10000, 6)
y.bars <- unique(machines$ybar)

for (i in 1:6) {
  theta.i[,i] = y.bars[i] + sqrt(msw/v/5)*rnorm(10000)
}
boxplot(theta.i,
        main = "marginal posteriors for mu vector",
        xlab = expression(mu)
)

# Frequentist methods

machines.aov <- aov(measurements ~ machine, data = machines)
summary(machines.aov)
tapply(machines$measurements, machines$machine, mean)
```