CHL5223 A1

Belina Jang

Question 1

(a)

For a test called ANA, we denote it as D. We know that there is a 13% chance that the patient has lupus given several symptoms. This gives us $P(D^+) = 0.13$, which follows $P(D^-) = 0.87$.

We know that the sensitivity is 99% and specificity is 80%. This gives us $P(T^+|D^+) = 0.99$ and $P(T^-|D^-) = 0.80$, which follows $P(T^-|D^+) = 0.01$ and $P(T^+|D^-) = 0.20$.

Q: If the patient has a positive result from this test, what is the new probability of the patient having lupus given this new information?

So we are looking for $P(D^+|T^+)$. Using Bayes' theorem, we get:

$$P(D^+|T^+) = \frac{P(T^+|D^+)P(D^+)}{P(T^+|D^+)P(D^+) + P(T^+|D^-)P(D^-)} = \frac{(0.99) \times (0.13)}{(0.99) \times (0.13) + (0.20) \times (0.87)} \approx 0.4251734$$

$\therefore$ Given the new information, the new probability of the patient having lupus given positive ANA test and several symptoms is approximately 42.5%.

(b)

For the new test called Anti-dsDNA, we denote it as A. We know that $P(A^+|D^+) = 0.73$ and $P(A^-|D^-) = 0.98$, which follows $P(A^-|D^+) = 0.27$ and $P(A^+|D^-) = 0.02$.

We are looking for $P(D^+|A^+)$, which is the probability of the patient having lupus given both tests are positive. Using Bayes' theorem, we get:

$$P(D^+|A^+) = \frac{P(A^+|D^+)P(D^+)}{P(A^+|D^+)P(D^+) + P(A^+|D^-)P(D^-)} = \frac{(0.73) \times (0.425)}{(0.73) \times (0.425) + (0.02) \times (0.575)} \approx 0.964258$$

$\therefore$ Given the new information, the new probability of the patient having lupus given positive Anti-dsDNA test is approximately 96.4%.

(c)

∴ It's important for a screening test to have a high sensitivity even though we might have to sacrifice specificity.

This is because it's important for screening test to capture as much of the true positive cases as possible. If you weight the severity of missing a disease versus falsely diagnosing a disease, you can imagine missing a disease comes with a high price as the patient might not receive timely treatment. On the other hand, false positive (falsely dignosing a disease) from screening test will result in further testing with higher specificity to confirm diagnosis and filter for false positive case from the first screening test. Therefore, it's important for a screening test to have a high sensitivity.

However it's important for the second test (confirmatory test) to have a high specificity. This is because the first screening test is designed to have high sensitivity to capture as many true positive cases as possible, which results in a high number of false positive cases as well. So the focus of the second test is to correctly filter out those false positive cases by capturing as many true negative cases as possible since missing a disease (false negative) is not of a high concern at this point as we already captured as many true positive cases as possible from the first screening test.

So for the right balance, it's important for the screening test to have high sensitivity to capture as many true positive cases as possible and the confirmatory test to have high specificity to correctly filter out the false positive cases.

(d)

Since the general male population prevalence of lupus in male is 1 in 25 000, we have $P(D^+) = \frac{1}{25000}$, which follows $P(D^-) = 1 - \frac{1}{25000} = \frac{24999}{25000}$.

The probability of actually having lupus given a positive ANA test is $P(D^+|T^+)$. Using Bayes' theorem, we get:

$$P(D^+|T^+) = \frac{P(T^+|D^+)P(D^+)}{P(T^+|D^+)P(D^+) + P(T^+|D^-)P(D^-)} = \frac{(0.99) \times (\frac{1}{25000})}{(0.99) \times (\frac{1}{25000}) + (0.20) \times (\frac{24999}{25000})} \approx 0.00019797$$

∴ Given the new information, the probability of having lupus in the general male population is approximately 0.0198%.

Question 2

```r
# set seed
set.seed(0)
# data set
D1=c(53, 49, 63, 72, 55, 65)
D2=c(28, 27, 36, 42, 25, 35)
```

(a) calculate the posterior mean, standard deviation, and a 95% credible region for average height.

Prior 1

```r
# priors
prior_prec <- 4/9
prior_mu <- 66

# likelihood precision
lh_prec <- 1/36

posterior1 <- function(prior_mu, prior_prec, lh_prec, data){
  # post precision: prior_prec+length(data)*lh_prec
  post_mu = (prior_prec*prior_mu +
 ↪  length(data)*(lh_prec)*mean(data))/(prior_prec+length(data)*lh_prec)
  post_sd = 1/sqrt(prior_prec + length(data) * lh_prec)
  cil_mu = post_mu - 1.96*post_sd
  ciu_mu = post_mu + 1.96*post_sd

  cat("Posterior mean:", post_mu, "\n")
  cat("Posterior standard deviation:", post_sd, "\n")
  cat("95% credible region1: (", cil_mu,", ",ciu_mu, ")\n", sep="")
  return(list(mu=post_mu, sd=post_sd, ci=c(cil_mu, ciu_mu)))
}

cat("For Data: D1\n"); result1 <- posterior1(prior_mu, prior_prec, lh_prec,
 ↪  D1)
```

```
For Data: D1
```

```
Posterior mean: 64.22727
Posterior standard deviation: 1.279204
95% credible region1: (61.72003, 66.73451)
```

```r
cat("For Data: D2\n"); result2 <- posterior1(prior_mu, prior_prec, lh_prec,
 ↪  D2)
```

```
For Data: D2

Posterior mean: 56.77273
Posterior standard deviation: 1.279204
95% credible region1: (54.26549, 59.27997)
```

Prior 2

```r
# priors
prior_mu <- 66
prior_prec <- 4
prior_alpha <- 1 # shape_parameter
prior_beta <- 36 # rate_parameter

posterior2 <- function(prior_alpha, prior_beta, prior_mu, data, n=10000){
  post_alpha <- prior_alpha + length(data)/2
  post_beta <- prior_beta + 0.5*sum((data-mean(data))^2) +
 ↪  prior_prec*length(data)*(mean(data)-prior_mu)^2/(2*(prior_prec+length(data)))
  post_tau_s <- rgamma(n, shape=post_alpha, rate=post_beta)

  post_mu_s <- rnorm(n, (prior_prec*prior_mu +
 ↪  length(data)*mean(data))/(prior_prec+length(data)),
 ↪  sd=1/sqrt((prior_prec+length(data))*post_tau_s))

  cat("Posterior mean:", mean(post_mu_s), "\n")
  cat("Posterior standard deviation:", sd(post_mu_s), "\n")
  cat("95% credible region2: (", quantile(post_mu_s, 0.025),",
   ↪  ",quantile(post_mu_s, 0.975), ")\n", sep="")
  return(list(mu=mean(post_mu_s),sd=sd(post_mu_s), ci=c(quantile(post_mu_s,
   ↪  0.025), quantile(post_mu_s, 0.975))))
}

cat("For Data: D1\n"); result3 <- posterior2(prior_alpha, prior_beta,
 ↪  prior_mu, D1)
```

```
For Data: D1
```

```
Posterior mean: 62.11175
Posterior standard deviation: 3.012417
95% credible region2: (56.0375, 68.14678)
```

```
cat("For Data: D2\n"); result4 <- posterior2(prior_alpha, prior_beta,
 ↪  prior_mu, D2)
```

```
For Data: D2

Posterior mean: 45.56744
Posterior standard deviation: 7.057562
95% credible region2: (31.51362, 59.47815)
```

prior 3

```
# priors
prior_mu <- 66
prior_prec <- 0.1
prior_alpha <- 0.001
prior_beta <- 0.001

cat("For Data: D1\n"); result5<-posterior2(prior_alpha, prior_beta, prior_mu,
 ↪  D1)
```

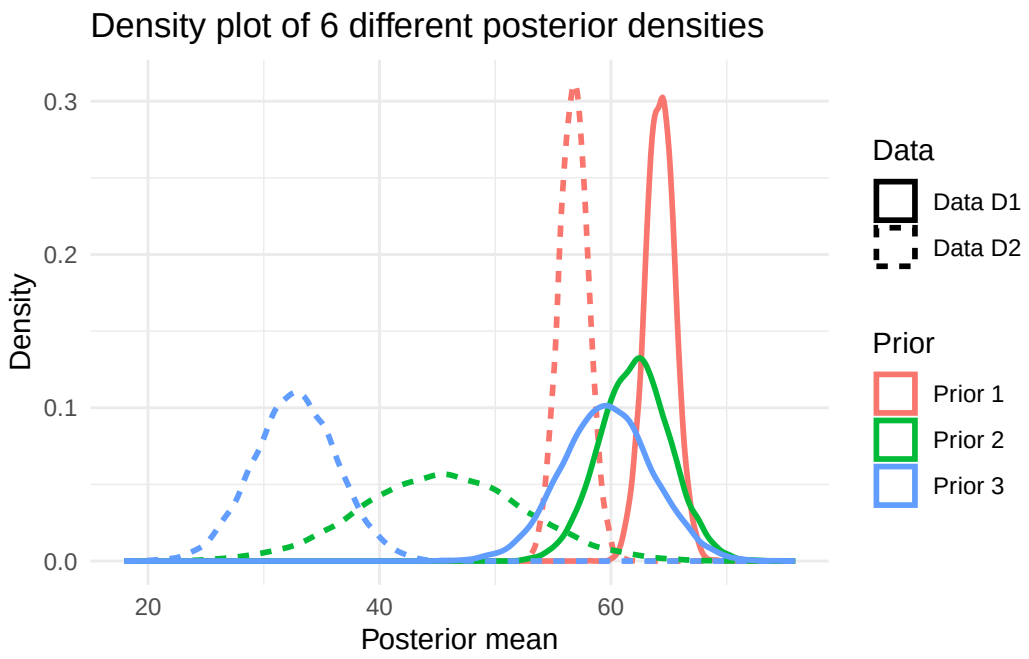
```
For Data: D1

Posterior mean: 59.64123
Posterior standard deviation: 3.909205
95% credible region2: (51.75773, 67.49622)
```

```
cat("For Data: D2\n"); result6<-posterior2(prior_alpha, prior_beta, prior_mu,
 ↪  D2)
```

```
For Data: D2

Posterior mean: 32.75028
Posterior standard deviation: 3.66561
95% credible region2: (25.36984, 40.06923)
```

(b) For the six different posterior densities from the previous question, give the density plots and put all six on the same plot.

```
library(ggplot2)
library(dplyr)
n=10000
density1 <- rnorm(n, result1[[1]], result1[[2]])
density2 <- rnorm(n, result2[[1]], result2[[2]])
density3 <- rnorm(n, result3[[1]], result3[[2]])
density4 <- rnorm(n, result4[[1]], result4[[2]])
density5 <- rnorm(n, result5[[1]], result5[[2]])
density6 <- rnorm(n, result6[[1]], result6[[2]])

df <- data.frame(
value = c(density1, density2, density3, density4, density5, density6),
Prior = rep(c("Prior 1", "Prior 1", "Prior 2", "Prior 2", "Prior 3", "Prior
 ↪  3"),each=n),
Data = rep(c("Data D1", "Data D2", "Data D1", "Data D2", "Data D1", "Data
 ↪  D2"),each=n)
)
q2b_plot <- ggplot(df, aes(x = value, color=Prior, linetype=Data)) +
geom_density(linewidth=1) +
labs(title = "Density plot of 6 different posterior densities", x =
 ↪  "Posterior mean", y = "Density") +
theme_minimal()

q2b_plot
```



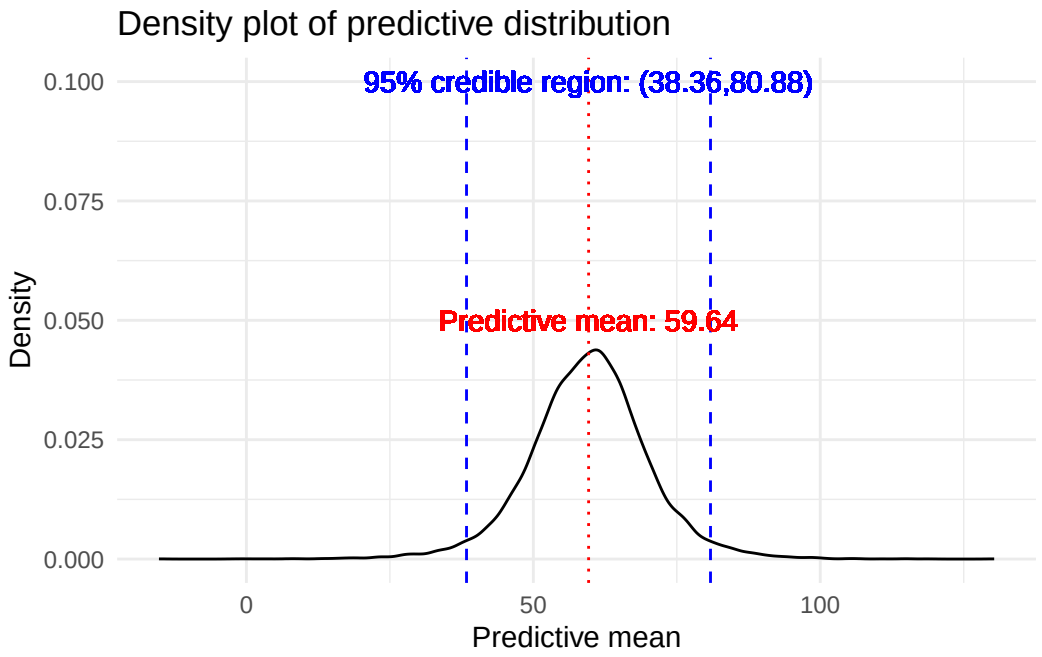
(c) Use data D1 and prior 3, find the predictive distribution for new observations. That is, using the posterior distribution, get the distribution of a new person from this population. To describe this predictive distribution, provide the mean, standard deviation, a 95% credible region and a density plot. These values can be obtained from either analytical methods, from sampling or a combination of analytical and sampling techniques. Don't forget to provide a description as to how you did this. (Either the formula or the sampling algorithm.)

```
set.seed(0)
# prior 3: priors
prior_prec <- 0.1
prior_mu <- 66
prior_alpha <- 0.001
prior_beta <- 0.001

predict_dist <- function(prior_alpha, prior_beta, prior_mu, data, n=10000){
  post_alpha <- prior_alpha + length(data)/2
  post_beta <- prior_beta + 0.5*sum((data-mean(data))^2) +
↪   prior_prec*length(data)*(mean(data)-prior_mu)^2/(2*(prior_prec+length(data)))
  post_tau_s <- rgamma(n, shape=post_alpha, rate=post_beta)
```

```r
  post_mu_s <- rnorm(n, (prior_prec*prior_mu +
 ↪  length(data)*mean(data))/(prior_prec+length(data)),
 ↪  sd=1/sqrt((prior_prec+length(data))*post_tau_s))

  pred <- rnorm(n, post_mu_s, 1/sqrt(post_tau_s))

  cat("Predictive mean:", mean(pred), "\n")
  cat("Predictive standard deviation:", sd(pred), "\n")
  cat("95% credible interval: (", quantile(pred, 0.025),", ",quantile(pred,
   ↪  0.975), ")\n", sep="")
  return(list(pred=pred, mean=mean(pred), sd=sd(pred), cil=quantile(pred,
   ↪  0.025), ciu=quantile(pred, 0.975)))
}
pred_result <- data.frame(predict_dist(prior_alpha, prior_beta, prior_mu,
 ↪  D1))
```

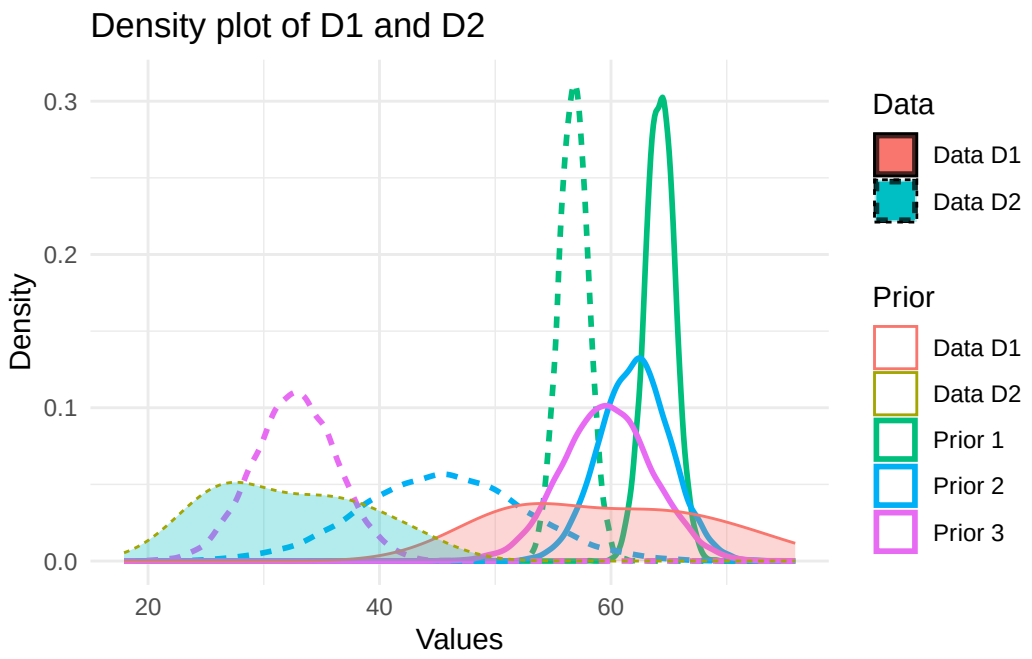
```
Predictive mean: 59.63527
Predictive standard deviation: 10.51416
95% credible interval: (38.35678, 80.87852)
```

```r
# density plot
ggplot(pred_result, aes(x=pred)) +
  geom_density() +
  geom_vline(xintercept = pred_result$mean, color="red", lty=3) +
  geom_vline(xintercept = pred_result$cil, color="blue", lty=2) +
  geom_vline(xintercept = pred_result$ciu, color="blue", lty=2) +
  labs(title = "Density plot of predictive distribution", x = "Predictive
   ↪  mean", y = "Density") +
  theme_minimal() +
  annotate("text", x = pred_result$mean, y = 0.050, label =
   ↪  paste0("Predictive mean: ", signif(pred_result$mean,4)), color="red") +
  annotate("text", x = pred_result$mean, y = 0.1, label = paste0("95%
   ↪  credible region:
   ↪  (",signif(pred_result$cil,4),",",signif(pred_result$ciu,4),")"),
   ↪  color="blue")
```



(d) For this part, please comment on what you have learned about priors and different data sets from this exercise.

```
df <- data.frame(
  value = c(D1, D2),
  Data = rep(c("Data D1", "Data D2"), each = length(D1))
)

# Plot density of raw data
q2b_plot + geom_density(data=df,aes(x = value, fill = Data, color = Data),
 ↪  alpha = 0.3) +
  labs(title = "Density plot of D1 and D2",
       x = "Values",
       y = "Density") +
  theme_minimal()
```



- Prior 1: The prior distribution is a normal distribution with mean 66 and precision 4/9 (SD: sqrt(1/tau)=1.5). The posterior distribution for the average height for data D1 and D2 is approximately normal with mean 64.23 and standard deviation 1.279 for data 1 and mean 56.77 and standard deviation 1.279 for data 2. You can see the mean values didn't change much from the prior mean value of 66. This is because prior 1 has τ_0 fixed therefore the prior has a strong influence on the posterior distribution and the data has a weak influence. Therefore, both mean values are close to the prior mean value of 66 and the standard deviation is also close to the prior standard deviation of 1.5, which caused both distribution to have smaller 95% credible regions. This results in prior 1 not being able to adapt to the data well.
- Prior 2: The prior distribution is a gamma distribution with alpha=1 and beta=36 (SD: 0.02778). The posterior distribution for the average height for data D1 and D2 is approximately normal with mean 62.11 and standard deviation 3.012 for data 1 and mean 45.57 and standard deviation 7.058 for data 2. You can see the mean value didn't deviate from the prior mean value of 66 for data 1 but it deviate significantly for data 2. This is because prior 2 has a weak influence on the posterior distribution and the data has a stronger influence since it vaguely assumes that the population was an adult population, which allows the mean value for data 2 to deviate significantly from the prior mean. You can also see that standard deviation values are higher than the prior 1's posterior standard deviation, which caused both distribution to have larger 95% credible regions. This results in prior 2 being able to adapt to the data well.
- Prior 3: The prior distribution is a gamma distribution with alpha=0.001 and beta=0.001

(SD: 31.62278). The posterior distribution for the average height for data D1 and D2 is approximately normal with mean 59.64 and standard deviation 3.909 for data 1 and mean 32.75 and standard deviation 3.666 for data 2. You can see the mean values didn't change much from the prior mean value of 66 for data 1 while it deviate significantly for data 2. Since data 1 is from adult population as expected, its posterior mean value didn't deviate from the prior mean as much while data 2 is from children population, which caused the mean value to deviate significantly from the prior mean. This was possible because prior 3 has a weak influence on the posterior distribution and the data has a strong influence. You can also see that standard deviation values are higher than the prior 1's posterior standard deviation, which caused both distribution to have larger 95% credible regions. This results in prior 3 being able to adapt to the data well. In fact, you can see the prior 3 seems to have a better fit to the actual data itself than the other two priors.

In summary, it seems like informative priors are more adaptable to the unexpected data (that are significantly different from prior knowledge) than the informative priors while the informative priors do better for the expected data (similar to prior knowledge) than the less informative priors. Therefore, it's important to choose the right prior depends on your confidence in the prior information and the data.

Question 3

Given information:

- X_i: the number of votes cast for the purple party in the ith district.
- θ_i: the probability that someone from district i votes for purple; has an unknown distribution, which you can assume it to be a beta distribution.
- $D_i = 1$ if the purple party is elected in district i and $D_i = 0$ otherwise.
- $T = 1$ if the purple party is elected and $T = 0$ otherwise.

(a)

This probability generating model for the above scenario is:

$$X_i|\theta_i \sim binom(5001, \theta_i)$$

where $\theta_i \sim beta(\alpha_i, \beta_i)$

D_i and T are deterministic functions of Xi.

- $D_i = 1$ if $X_i > 2500$ and $D_i = 0$ otherwise.
- $T = 1$ if $\sum_{i=1}^{3} D_i \geq 2$ and $T = 0$ otherwise.

(b)

Non-informative prior

For the non-informative prior, we need a prior with equal probability for all θ_i, which is a uniform distribution. So we can choose $beta(1, 1)$.

Informative prior

For the informative prior, we need to find a prior with approximately 95% of its mass between 0.40 and 0.60. So we are looking for $\theta_i \sim beta(\alpha, \beta)$ such that $P(0.40 < \theta_i < 0.60) = 0.95$. Which means that the centre of the distribution will be approximately 0.50, which follows $\mu = \frac{\alpha}{\alpha+\beta} = 0.50$. This gives us $\alpha = \beta$.

So we basically want to find α such that cumulative probabilities of $beta(\alpha, \alpha)$ is 0.975 at 60% quantile and 0.025 at 40% quantile, which is the same as finding `alpha` such that `pbeta(0.60,alpha,alpha)=0.975` and `pbeta(0.40,alpha,alpha)=0.025`. One will follow the other automatically. So we only need to find `alpha` such that `pbeta(0.60,alpha,alpha)=0.975`.

```r
# set seed
set.seed(0)

alpha_beta <- function(alpha){
  # cumulative prob at 0.60
  prob <- pbeta(0.60, alpha, alpha)

  # difference between calculated prob and 0.975
  return(0.975-prob)
}

alpha = uniroot(alpha_beta, c(0, 1000))$root

cat("alpha =", alpha)
```

```
alpha = 47.29982
```

$\therefore \alpha = \beta = 47.300$.

(c)

From simple random sampling, we got:

district 1: 53 said purple and 45 said brown which gives $x_1 = 53,\ n_1 - x_1 = 45$

district 2: 72 said purple and 78 said brown which gives $x_2 = 72,\ n_2 - x_2 = 78$

district 3: 18 said purple and 22 said brown which gives $x_3 = 18,\ n_3 - x_3 = 22$

Given above information, we get:

$$\theta_i \sim beta(\alpha + x_i, \beta + (n_i - x_i))$$

Then we can calculate the posterior distribution for θ_i for each district.

For non-informative prior (previously defined as $beta(1, 1)$):

$\theta_1 \sim beta(1 + 53, 1 + 45)$

$\theta_2 \sim beta(1 + 72, 1 + 78)$

$\theta_3 \sim beta(1 + 18, 1 + 22)$

For informative prior (previously defined as $beta(47.300, 47.300)$):

$\theta_1 \sim beta(47.300 + 53, 47.300 + 45)$

$\theta_2 \sim beta(47.300 + 72, 47.300 + 78)$

$\theta_3 \sim beta(47.300 + 18, 47.300 + 22)$

Q: Calculate the posterior probability for the percent who will vote for purple in each district.

```
# set seed
set.seed(0)

# prior 1 (non-informative) (alpha,beta)
p1_ab <- list(c(54, 46), c(73, 79), c(19, 23))

# prior 2 (informative) (alpha,beta)
p2_ab <- list(c(47.300+54, 47.300+46), c(47.300+73, 47.300+79), c(47.300+19,
 ↪  47.300+23))

posterior <- function(ab){
  result <- list()
  for (i in 1:length(ab)){
    alpha <- ab[[i]][1]
    beta <- ab[[i]][2]
```

```
    theta <- rbeta(n, alpha, beta)
    cat("District", i, "\n")
    cat("Posterior distribution: beta(",alpha,", ",beta,")\n", sep="")
    cat("Posterior mean:", mean(theta), "\n")
    cat("Posterior standard deviation:", sd(theta), "\n")
    cat("95% credible interval: (", quantile(theta, 0.025),",
     ↪  ",quantile(theta, 0.975), ")\n \n", sep="")
    result = append(result,
 ↪  list(list(district=i,post_dist=paste0("beta(",alpha,",",beta,")"),mean=mean(theta),
 ↪  sd=sd(theta), ci=c(quantile(theta, 0.025), quantile(theta, 0.975)))))
  }
  return(result)
}
```

```
cat("Using non-informative prior: beta(1, 1)"); posterior1 <-
 ↪  posterior(p1_ab)
```

```
Using non-informative prior: beta(1, 1)

District 1
Posterior distribution: beta(54, 46)
Posterior mean: 0.5399422
Posterior standard deviation: 0.05000965
95% credible interval: (0.4416242, 0.6365645)

District 2
Posterior distribution: beta(73, 79)
Posterior mean: 0.480555
Posterior standard deviation: 0.04021512
95% credible interval: (0.4023104, 0.5593895)

District 3
Posterior distribution: beta(19, 23)
Posterior mean: 0.4519909
Posterior standard deviation: 0.07577181
95% credible interval: (0.3054493, 0.6003805)
```

```
cat("Using informative prior: beta(47.300, 47.300)"); posterior2 <-
 ↪  posterior(p2_ab)
```

```
Using informative prior: beta(47.300, 47.300)
```

```
District 1
Posterior distribution: beta(101.3, 93.3)
Posterior mean: 0.5208666
Posterior standard deviation: 0.03634897
95% credible interval: (0.4492881, 0.5920399)

District 2
Posterior distribution: beta(120.3, 126.3)
Posterior mean: 0.4877466
Posterior standard deviation: 0.0311843
95% credible interval: (0.4260986, 0.5473116)

District 3
Posterior distribution: beta(66.3, 70.3)
Posterior mean: 0.4854993
Posterior standard deviation: 0.04273399
95% credible interval: (0.4012432, 0.5693055)
```

(d)

```r
# set seed
set.seed(0)

library(hash)
party <- list("Purple", "Brown")
district <- list("District 1", "District 2", "District 3")

n = 10000

# simulation
sim <- function(ab){

  # number of wins in town council
  t <- hash()
  t[["Purple"]] <- 0
  t[["Brown"]] <- 0

  # number of purple wins per districts
  d <- hash()
  d[["District 1"]] <- 0
```

```
  d[["District 2"]] <- 0
  d[["District 3"]] <- 0

  # each simulated election
  for (j in 1:n){
    di<-0
    # for each district i
    for (i in 1:length(ab)){
      alpha <- ab[[i]][1]
      beta <- ab[[i]][2]
      # probability of a citizen voting for the purple party in district i
      theta_i <- rbeta(1, alpha, beta)
      # number of citizens who voted for purple party in district i
      x <- rbinom(1, 5001, theta_i)
      # counts the number of purple wins
      di = di + ifelse(x>2500, 1, 0)
      # save the result
      d[[district[[i]]]] <- d[[district[[i]]]] + ifelse(x>2500, 1, 0)
    }
    # final result from each election
    if (di>=2){
      t[["Purple"]] <- t[["Purple"]] + 1
    } else {
      t[["Brown"]] <- t[["Brown"]] + 1
    }
  }

  # probablity of purple party having a majority in town council
  cat("Probability of purple party having a majority in town council:",
   ↪  t[["Purple"]]/n)
  # probablities of purple party winning in each district
  cat("\nProbabilities of purple party winning in each district:\n")
  for (i in 1:length(d)){
    cat(district[[i]], ": ", d[[district[[i]]]]/n, "\n", sep = "")
  }
  return(list("t"=t, "d"=d))
}
```

```
cat("Using non-informative prior: beta(1, 1)"); sim1 <- sim(p1_ab)
```

```
Using non-informative prior: beta(1, 1)
```

```
Probability of purple party having a majority in town council: 0.4178
Probabilities of purple party winning in each district:
District 1: 0.7881
District 2: 0.325
District 3: 0.263
```

```
cat("Using informative prior: beta(47.300, 47.300)"); sim2 <- sim(p2_ab)
```

```
Using informative prior: beta(47.300, 47.300)
```

```
Probability of purple party having a majority in town council: 0.4635
Probabilities of purple party winning in each district:
District 1: 0.7188
District 2: 0.3548
District 3: 0.3748
```

Question 4

(a) What is your (posterior) belief in θ which is the average amount that the drug lowers the value of HbA1c? (This should be expressed as a distribution.)

given information:

$n = 209$, $\bar{Y_1} = -1.82$, $S_1 = 0.21$

Then the posterior distribution for θ is:

$$\theta|\bar{Y_1}, S_1 \sim N(\bar{Y_1}, S_1^2) = N(\mu = -1.82, \sigma^2 = 0.21^2)$$

(b) Starting with the prior belief that you had from your own clinical trial, update your belief using the new data. (That is, use the information from the previous question as your prior distribution and then use this new information as data to get a new posterior distribution.) What is your new posterior belief on the values of θ?

Given information:$n = 79$,$\bar{Y_2} = -1.02$,$S_2 = 0.28$

Then we get:

$$\mu' = \frac{\tau_0\mu_0 + n\tau\bar{x}}{\tau_0 + n\tau} = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{\frac{\bar{Y_1}}{S_1^2} + \frac{\bar{Y_2}}{S_2^2}}{\frac{1}{S_1^2} + \frac{1}{S_2^2}} = \frac{\frac{-1.82}{0.21^2} + \frac{-1.02}{0.28^2}}{\frac{1}{0.21^2} + \frac{1}{0.28^2}} = -1.532$$

and

$$\sigma^2 = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{1}{\frac{1}{0.21^2} + \frac{1}{0.28^2}} = 0.028224$$

Then the new posterior distribution for θ is:

$$\theta|\bar{Y_1}, \bar{Y_2}, S_1, S_2 \sim N(\mu = -1.532, \sigma^2 = 0.028224)$$

(c) Now consider the problem from your colleagues point of view. When she first collects her data without seeing your data, what is her posterior belief for the value of θ? After she finds out about your data and she updates her belief, what is her new belief in the value of θ?

Given information: $n = 79$, $\bar{Y_2} = -1.02$, $S_2 = 0.28$

Then without seeing my data, the posterior distribution for θ is:

$$\theta|\bar{Y_2}, S_2 \sim N(\bar{Y_2}, S_2^2) = N(\mu = -1.02, \sigma^2 = 0.28^2)$$

Then after she finds out about my data, we get:

$$\mu' = \frac{\tau_0\mu_0 + n\tau\bar{x}}{\tau_0 + n\tau} = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{\frac{\bar{Y_1}}{S_1^2} + \frac{\bar{Y_2}}{S_2^2}}{\frac{1}{S_1^2} + \frac{1}{S_2^2}} = \frac{\frac{-1.82}{0.21^2} + \frac{-1.02}{0.28^2}}{\frac{1}{0.21^2} + \frac{1}{0.28^2}} = -1.532$$

and

$$\sigma^2 = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{1}{\frac{1}{0.21^2} + \frac{1}{0.28^2}} = 0.028224$$

Then the new posterior distribution for θ is:

$$\theta|\bar{Y_1}, \bar{Y_2}, S_1, S_2 \sim N(\mu = -1.532, \sigma^2 = 0.02822)$$

which is the same distribution as in part (b).

(d) Pooling all this information, what is your new belief in the value of θ? Provide the general formula for combining this information. (Note: since you can easily get the formula by googling up "fixed effect meta analysis", there will be little weight to the actual formula. You will be mostly graded on explaining how the formula follows from the principles/formulas presented in the class material.)

Given information:

$n_1 = 209, \bar{Y_1} = -1.82, S_1 = 0.21$ which gives $\tau_1 = \frac{1}{S_1^2} = \frac{1}{0.21^2}$

$n_2 = 79, \bar{Y_2} = -1.02, S_2 = 0.28$ which gives $\tau_2 = \frac{1}{S_2^2} = \frac{1}{0.28^2}$

$n_3 = 19, \bar{Y}_3 = -1.9, S_3 = 0.945$ which gives $\tau_3 = \frac{1}{S_3^2} = \frac{1}{0.945^2}$

$n_4 = 100, \bar{Y}_4 = -2.00, S_4 = 0.285$ which gives $\tau_4 = \frac{1}{S_4^2} = \frac{1}{0.285^2}$

$n_5 = 20, \bar{Y}_5 = -1.21, S_5 = 0.545$ which gives $\tau_5 = \frac{1}{S_5^2} = \frac{1}{0.545^2}$

Then the new posterior mean is:

$$\mu' = \frac{\tau_0\mu_0 + \sum_i \tau_i \bar{Y}_i}{\tau_0 + \sum_i \tau_i} = \frac{0 + \frac{-1.82}{0.21^2} + \frac{-1.02}{0.28^2} + \frac{-1.9}{0.945^2} + \frac{-2.00}{0.285^2} + \frac{-1.21}{0.545^2}}{0 + \frac{1}{0.21^2} + \frac{1}{0.28^2} + \frac{1}{0.945^2} + \frac{1}{0.285^2} + \frac{1}{0.545^2}} \approx -1.62945$$

and

$$\sigma^2 = \frac{1}{\sum_i \tau_i} = \frac{1}{\frac{1}{0.21^2} + \frac{1}{0.28^2} + \frac{1}{0.945^2} + \frac{1}{0.285^2} + \frac{1}{0.545^2}} \approx 0.01915$$

Then the new posterior distribution for θ is:

$$\theta|\bar{Y}_1, \bar{Y}_2, \bar{Y}_3, \bar{Y}_4, \bar{Y}_5, S_1, S_2, S_3, S_4, S_5 \sim N(\mu = -1.629, \sigma^2 = 0.01915)$$