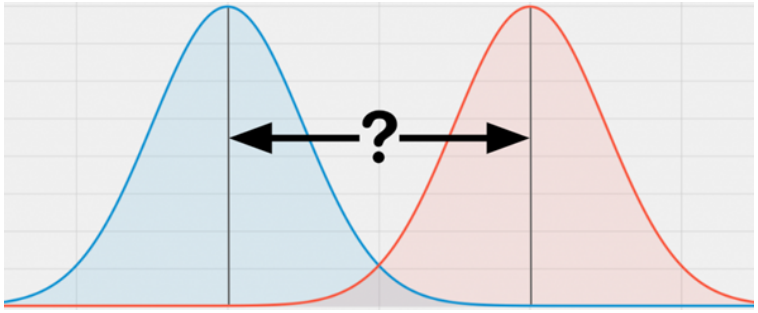
Stat 136 (Bayesian Statistics)

Lesson 2.2: Bayesian Inference for the Difference Between Two Normal Means

Introduction

Oftentimes the objective of a study is comparison of two or more groups or populations based on measures such as the means or variances.

- Treated vs Control
- Independent samples or paired samples



Likelihoods

Let $Y_{11}, Y_{12}, \cdots, Y_{1n_1}$ be a random sample of size n_1 from $N(\mu_1, \sigma^2)$. Also, let $Y_{21}, Y_{22}, \cdots, Y_{2n_2}$ be a random sample of size n_2 from $N(\mu_2, \sigma^2)$. Further, suppose that the 2 random samples are independent and the common variance, σ^2, is known. Then,

$$L(y_{11}, y_{12}, \cdots, y_{1n_1}|\mu_1, \sigma^2) \propto exp\left(-\frac{n_1(\bar{y}_1 - \mu_1)^2}{2\sigma^2}\right)$$

and

$$L(y_{21}, y_{22}, \cdots, y_{2n_2}|\mu_2, \sigma^2) \propto exp\left(-\frac{n_2(\bar{y}_2 - \mu_2)^2}{2\sigma^2}\right)$$

Normal priors

Let the priors be: $N(\mu_{1p}, \sigma_{1p}^2)$ and $N(\mu_{2p}, \sigma_{2p}^2)$. That is,

$$\pi(\mu_1) \propto exp\left(-\frac{(\mu_1 - \mu_{1p})^2}{2\sigma_{1p}^2}\right)$$

and

$$\pi(\mu_2) \propto exp\left(-\frac{(\mu_2 - \mu_{2p})^2}{2\sigma_{2p}^2}\right)$$

Posterior densities

Based on the previous lesson, it can be easily shown that the posterior densities for μ_1 and μ_2 are, respectively

$$\pi(\mu_1|y_{11}, y_{12}, \cdots, y_{1n_1}) \propto exp\left[-\frac{1}{2\left(\frac{\sigma_{1p}^2\sigma^2}{\sigma^2+n_1\sigma_{1p}^2}\right)}\left(\mu_1 - \frac{\sigma^2\mu_{1p} + n_1\sigma_{1p}^2\bar{y}_1}{\sigma^2 + n_1\sigma_{1p}^2}\right)\right]$$

and

$$\pi(\mu_2|y_{21}, y_{22}, \cdots, y_{2n_2}) \propto exp\left[-\frac{1}{2\left(\frac{\sigma_{2p}^2\sigma^2}{\sigma^2+n_2\sigma_{2p}^2}\right)}\left(\mu_2 - \frac{\sigma^2\mu_{2p} + n_2\sigma_{2p}^2\bar{y}_2}{\sigma^2 + n_2\sigma_{2p}^2}\right)\right]$$

Thus, the posterior densities for μ_1 and μ_2 are, respectively, $N(\mu_1^{'}, \sigma_1^{2'})$ and $N(\mu_2^{'}, \sigma_2^{2'})$, where

$$\mu_1^{'} = \frac{\sigma^2\mu_{1p} + n_1\sigma_{1p}^2\bar{y}_1}{\sigma^2 + n_1\sigma_{1p}^2},$$

$$\sigma_1^{2'} = \frac{\sigma^2\sigma_{1p}^2}{\sigma^2 + n_1\sigma_{1p}^2},$$

$$\mu_2^{'} = \frac{\sigma^2\mu_{2p} + n_2\sigma_{2p}^2\bar{y}_2}{\sigma^2 + n_2\sigma_{2p}^2},$$

and

$$\sigma_2^{2'} = \frac{\sigma^2\sigma_{2p}^2}{\sigma^2 + n_2\sigma_{2p}^2}$$

Posterior density of the mean difference μ_d

Recall that if the sampling distributions (likelihood) and the prior distributions are independent, then the posterior distributions are independent as well.

Let $\mu_d = \mu_1 - \mu_2$. Then

$$\mu_d \sim N(\mu_d^{'}, \sigma_d^{2'})$$

where

$$\mu_d^{'} = \mu_1^{'} - \mu_2^{'}$$

and

$$\sigma_d^{2'} = \sigma_1^{2'} + \sigma_2^{2'}$$

Credible interval for the mean difference μ_d

A $(1-\alpha) \times 100\%$ credible interval for μ_d is given by

$$\mu_d^{'} \pm z_{\frac{\alpha}{2}} \sigma_d^{'}$$

where:

$$\sigma_d^{'} = \sqrt{\sigma_1^{2'} + \sigma_2^{2'}}$$

Bayesian hypothesis test for μ_d

To test the hypothesis $H_0 : \mu_d \leq 0$ versus $H_1 : \mu_d > 0$ we calculate the posterior probability of the null hypothesis $\implies P(H_0|data) = P(\mu_d \leq 0|data)$

- If this probability is less than some α-level of significance, we reject H_0

To test the hypothesis $H_0 : \mu_d = 0$ versus $H_1 : \mu_d \neq 0$, we construct the $(1-\alpha) \times 100\%$ credible interval for μ_d

- If the $(1-\alpha) \times 100\%$ credible interval for μ_d includes zero, we fail to reject H_0

Example 1

A thermal power station discharges its cooling water into a river. An environmental scientist wants to determine if this has adversely affected the dissolved oxygen level. She takes samples of water one kilometer upstream from the power station, and one kilometer downstream from the power station, and measures the dissolved oxygen level. The data are given below. Is the data sufficient to conclude that dissolved oxygen level is higher upstream than downstream?

Upstream	Downstream
10.1	9.7
10.2	10.3
13.4	6.4
8.2	7.3
9.8	11.7
	8.9

a. Assume that the observations come from $N(\mu_1, 4)$ and $N(\mu_2, 4)$. Use the independent $N(10, 4.5)$ and $N(9, 4)$ as prior distributions for μ_1 and μ_2, respectively. Find the posterior distributions of μ_1 and μ_2, respectively.

b. Find the posterior distribution of $\mu_d = \mu_1 - \mu_2$.

c. Find a 95% credible interval of $\mu_d = \mu_1 - \mu_2$.

d. At the 5% level of significance, perform a Bayesian test of the hypothesis $H_0 : \mu_d \leq 0$ versus $H_1 : \mu_d > 0$

```
u <- c(10.1, 10.2, 13.4, 8.2, 9.8)
d <- c(9.7, 10.3, 6.4, 7.3, 11.7, 8.9)
n1 <- length(u)
ybar1 <- mean(u)
n2 <- length(d)
ybar2 <- mean(d)
print(cbind(n1,ybar1,n2,ybar2))
```

```
##      n1 ybar1 n2 ybar2
## [1,]  5 10.34  6  9.05
```

- Upstream sample: $n_1 = 5$ and $\bar{y}_1 = 10.34$
- Downstream sample: $n_2 = 6$ and $\bar{y}_1 = 9.05$

- Parameters of the normal priors are: $\mu_{1p} = 10$, $\sigma^2_{1p} = 4.5$, $\mu_{2p} = 9$, and $\sigma^2_{2p} = 4.0$
- The posterior distributions of μ_1 and μ_2, respectively, are: $N(\mu_1^{'}, \sigma_1^{2'})$ and $N(\mu_2^{'}, \sigma_2^{2'})$, where

$$\mu_1^{'} = \frac{\sigma^2\mu_{1p} + n_1\sigma^2_{1p}\bar{y}_1}{\sigma^2 + n_1\sigma^2_{1p}} = \frac{4(10) + 5(4.5)(10.34)}{4 + 5(4.5)} \approx 10.2887,$$

$$\sigma_1^{2'} = \frac{\sigma^2\sigma^2_{1p}}{\sigma^2 + n_1\sigma^2_{1p}} = \frac{4(4.5)}{4 + 5(4.5)} \approx 0.6792,$$

$$\mu_2^{'} = \frac{\sigma^2\mu_{2p} + n_2\sigma^2_{2p}\bar{y}_2}{\sigma^2 + n_2\sigma^2_{2p}} = \frac{4(9) + 6(4)(9.05)}{4 + 6(4)} \approx 9.0429,$$

and

$$\sigma_2^{2'} = \frac{\sigma^2\sigma^2_{2p}}{\sigma^2 + n_2\sigma^2_{2p}} = \frac{4(4)}{4 + 6(4)} \approx 0.5714$$

- The posterior distribution for $\mu_d = \mu_1 - \mu_2$ is $N(\mu_d^{'}, \sigma_d^{2'})$, where:

$$\mu_d^{'} = \mu_1^{'} - \mu_2^{'} = 10.2887 - 9.0429 = 1.2458,$$

and

$$\sigma_d^{2'} = \sigma_1^{2'} + \sigma_2^{2'} = 0.6792 + 0.5714 = 1.2506$$

- A 95% credible interval for $\mu_d = \mu_1 - \mu_2$ is

$$\mu_d^{'} \pm z_{\frac{\alpha}{2}}\sigma_d^{'} \implies 1.2458 \pm 1.96(\sqrt{1.2506}) \implies (-0.9461, 3.4377)$$

- To test the hypothesis $H_0 : \mu_d \leq 0$ against $H_1 : \mu_d > 0$, we compute $P(\mu_d \leq 0|data)$

```
round(pnorm(0, 1.2458, sqrt(1.2506)),4)
```

```
## [1] 0.1326
```

- Since this probability is greater than the 5% level of significance, we fail to reject H_0.
- Therefore, the data is not sufficient to conclude that the amount of dissolved oxygen upstream is significantly greater that the amount of dissolved oxygen downstream

When the variance σ^2 is unknown

When the common variance σ^2 is unknown, we estimate it using the pooled sample variance given by

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

Suppose we use independent "flat" priors for μ_1 and μ_2, that is, $\pi(\mu_1) = \pi(\mu_2) = 1$. Just like before, the posterior distributions of μ_1 and μ_2 will be identical to the likelihoods of the sampling distributions. That is, $\mu_1 \sim N(\mu_1^{'}, \sigma_1^{2'})$ and $\mu_2 \sim N(\mu_2^{'}, \sigma_2^{2'})$, where:

$$\begin{aligned}
\mu_1^{'} &= \bar{y}_1 \\
\sigma_1^{2'} &= \frac{s_p^2}{n_1} \\
\mu_2^{'} &= \bar{y}_2 \\
\sigma_2^{2'} &= \frac{s_p^2}{n_2}
\end{aligned}$$

As in the case where σ^2 is known, $\mu_d \sim N(\mu_d{}', \sigma_d^{2'})$, where the parameters $\mu_d{}'$ and $\sigma_d^{2'}$ are defined as before.

A $(1 - \alpha) \times 100\%$ credible interval for μ_d is constructed based on the t distribution, and is given by

$$\mu_d^{'} \pm t_{\frac{\alpha}{2}, df}(\sigma_d^{'}), \text{ where: } df = n_1 + n_2 - 2$$

To test the hypothesis $H_0 : \mu_d \leq 0$ versus $H_1 : \mu_d > 0$, we calculate the posterior probability of the null hypothesis using the t distribution with $n_1 + n_2 - 2$ degrees of freedom.

To test the hypothesis $H_0 : \mu_d = 0$ versus $H_1 : \mu_d \neq 0$, we construct the $(1 - \alpha) \times 100\%$ credible interval for μ_d based on the t distribution with $n_1 + n_2 - 2$ degrees of freedom.

When the variance σ^2 is unknown: an example

Let us revisit the previous example. This time we assume that the observations are a random sample from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, where σ^2 is not known

Consider "flat" prior distributions for μ_1 and μ_2.

a. Find the posterior distributions of μ_1 and μ_2, respectively.

b. Find the posterior distribution of $\mu_d = \mu_1 - \mu_2$.

c. Find a 95% credible interval of $\mu_d = \mu_1 - \mu_2$.

d. At the 5% level of significance, perform a Bayesian test of the hypothesis $H_0 : \mu_d \leq 0$ versus $H_1 : \mu_d > 0$

```
sp <- sqrt(((n1-1)*var(u)+(n2-1)*var(d))/(n1+n2-2))
print(sp^2)
```

```
## [1] 3.714111
```

When "flat" priors are used then the posterior distributions for μ_1 and μ_2 are, respectively, $N(\mu_1^{'}, \sigma_1^{2'})$ and $N(\mu_2^{'}, \sigma_2^{2'})$, where

$$\mu_1^{'} = \bar{y}_1 = 10.34$$

$$\sigma_1^{2'} = \frac{s_p^2}{n_1} \approx 0.743$$

$$\mu_2^{'} = \bar{y}_2 = 9.05$$

$$\sigma_2^{2'} = \frac{s_p^2}{n_2} \approx 0.619$$

We also learned that $\mu_d \sim N(\mu_d{}', \sigma_d^{2'})$, where

$$\mu_d^{'} = \bar{y}_1 - \bar{y}_2 = 1.29$$

and

$$\sigma_d^{2'} = \sigma_1^{2'} + \sigma_2^{2'} = 1.362$$

A $(1-\alpha) \times 100\%$ credible interval for μ_d is given by

$$\mu_d{}' \pm t_{\frac{\alpha}{2},df} \sqrt{\sigma_d^{2'} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\implies 1.29 \pm 2.262 \times \sqrt{1.362 \left(\frac{1}{5} + \frac{1}{6}\right)}$$

$$\implies (-0.309, 2.889)$$

To test the hypothesis $H_0 : \mu_d \leq 0$ versus $H_1 : \mu_d > 0$, we compute $P(\mu_d \leq 0 | data)$ using the code chunk below

```r
print(pnorm(0,1.29, sqrt(1.362)))
```

```
## [1] 0.1345032
```

Since this probability is greater than the 5% level of significance, we fail to reject H_0. Therefore, the data is not sufficient to conclude that the amount of dissolved oxygen upstream is significantly greater that the amount of dissolved oxygen downstream

Unequal but known variances

Suppose we draw independent random samples from two normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, where $\sigma_1^2 \neq \sigma_2^2$ but both are known. Then the likelihood of the sample observations are, respectively, given by

$$L(y_{11}, y_{12}, \cdots, y_{1n_1}|\mu_1, \sigma_1^2) \propto exp\left[-\frac{n_1(\bar{y}_1 - \mu_1)^2}{2\sigma_1^2}\right]$$

and

$$L(y_{21}, y_{22}, \cdots, y_{2n_2}|\mu_2, \sigma_2^2) \propto exp\left[-\frac{n_2(\bar{y}_2 - \mu_2)^2}{2\sigma_2^2}\right]$$

Let the priors be $N(\mu_{1p}, \sigma_{1p}^2)$ and $N(\mu_{2p}, \sigma_{2p}^2)$, i. e.

$$\pi(\mu_1) \propto exp\left[-\frac{(\mu_1 - \mu_{1p})^2}{2\sigma_{1p}^2}\right]$$

and

$$\pi(\mu_2) \propto exp\left[-\frac{(\mu_2 - \mu_{2p})^2}{2\sigma_{2p}^2}\right]$$

Therefore, using the Bayes' theorem the posterior distributions of μ_1 and μ_2 are

$$\pi(\mu_1|\bar{y}_1) \propto exp\left[-\frac{1}{2\left(\frac{\sigma_1^2\sigma_{1p^2}}{\sigma_1^2+n_1\sigma_{1p}^2}\right)}\left(\mu_1 - \frac{\sigma_1^2\mu_{1p} + n_1\sigma_{1p}^2\bar{y}_1}{\sigma_1^2 + n_1\sigma_{1p}^2}\right)\right]$$

and

$$\pi(\mu_2|\bar{y}_2) \propto exp\left[-\frac{1}{2\left(\frac{\sigma_2^2\sigma_{2p^2}}{\sigma_2^2+n_2\sigma_{2p}^2}\right)}\left(\mu_2 - \frac{\sigma_2^2\mu_{2p} + n_2\sigma_{2p}^2\bar{y}_2}{\sigma_2^2 + n_2\sigma_{2p}^2}\right)\right]$$

Unequal and unknown variances

The unknown variances σ_1^2 and σ_2^2 will be estimated by the sample variances S_1^2 and S_2^2, respectively. Thus, the likelihood of the sample observations become

$$L(y_{11}, y_{12}, \cdots, y_{1n_1}|\mu_1, S_1^2) \propto exp\left[-\frac{n_1(\bar{y}_1 - \mu_1)^2}{2S_1^2}\right]$$

and

$$L(y_{21}, y_{22}, \cdots, y_{2n_2}|\mu_2, S_2^2) \propto exp\left[-\frac{n_2(\bar{y}_2 - \mu_2)^2}{2S_2^2}\right]$$

Assuming the same prior densities for μ_1 and μ_2 then the posterior densities of μ_1 and μ_2 are, respectively:

$$\pi(\mu_1|\bar{y}_1) \propto exp\left[-\frac{1}{2\left(\frac{S_1^2\sigma_{1p}^2}{S_1^2+n_1\sigma_{1p}^2}\right)}\left(\mu_1 - \frac{S_1^2\mu_{1p} + n_1\sigma_{1p}^2\bar{y}_1}{S_1^2 + n_1\sigma_{1p}^2}\right)\right]$$

and

$$\pi(\mu_2|\bar{y}_2) \propto exp\left[-\frac{1}{2\left(\frac{S_2^2\sigma_{2p^2}}{S_2^2+n_2\sigma_{2p}^2}\right)}\left(\mu_2 - \frac{S_2^2\mu_{2p} + n_2\sigma_{2p}^2\bar{y}_2}{S_2^2 + n_2\sigma_{2p}^2}\right)\right]$$

Test of one-sided hypotheses on $\mu_d = \mu_1 - \mu_2$ will be based on the posterior distribution just like before. However, When the variances are estimated by the sample variances, the critical value used in constructing the credible interval is now based on the t distribution with degrees of freedom (Satterthwaite's approximation).

$$df \approx \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$$

Unequal and unknown variances: an example

Independent random samples of ceramic produced by two different processes (P1 & P2) were tested for hardness. The results were:

P1	P2
8.8	9.2

P1	P2
9.6	9.5
8.9	10.2
9.2	9.5
9.9	9.8
9.4	9.5
9.2	9.3
10.1	9.2

```
print(cbind(m1=mean(P1),v1=var(P1),m2=mean(P2), v2=var(P2)))
```

```
##          m1        v1    m2        v2
## [1,] 9.3875 0.2098214 9.525 0.1135714
```

Let these observations be distributed as $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, for Process 1 and Process 2. Further, suppose that $\sigma_1^2 \neq \sigma_2^2$ and are both unknown. Let us use the following priors for μ_1 and μ_2: $N(10, 1)$ and $N(11, 2)$.

Consequently, the posterior distribution of μ_1 is $N(\mu_1^{'}, \sigma_1^{2'})$, where

$$\mu_1^{'} = \frac{S_1^2\mu_{1p} + n_1\sigma_{1p}^2\bar{y}_1}{S_1^2 + n_1\sigma_{1p}^2} = \frac{0.21(10) + 8(1)(9.39)}{0.21 + 8(1)} \approx 9.41$$

and

$$\sigma_1^{2'} = \frac{S_1^2\sigma_{1p}^2}{S_1^2 + n_1\sigma_{1p}^2} = \frac{0.21(1)}{0.21 + 8(1)} \approx 0.026$$

The posterior distribution of μ_2 is $N(\mu_2^{'}, \sigma_2^{2'})$, where

$$\mu_2^{'} = \frac{S_2^2\mu_{2p} + n_2\sigma_{2p}^2\bar{y}_2}{S_2^2 + n_2\sigma_{2p}^2} = \frac{6.32(11) + 8(2)(8.67)}{6.32 + 8(2)} \approx 9.33$$

and

$$\sigma_2^{2'} = \frac{S_2^2\sigma_{2p}^2}{S_2^2 + n_1\sigma_{2p}^2} = \frac{6.32(2)}{6.32 + 8(2)} \approx 0.566$$

Finally, the posterior distribution of μ_d is $N(\mu_d^{'}, \sigma_d^{2'})$, where

$$\mu_d^{'} = \mu_1^{'} - \mu_2^{'} = 9.41 - 9.33 = 0.08$$

and

$$\sigma_d^{2'} = \sigma_1^{2'} + \sigma_2^{2'} = 0.026 + 0.566 = 0.592$$

A $(1-\alpha) \times 100\%$ credible interval for μ_d is

$$\mu_d^{'} \pm t_{\frac{0.05}{2},9}\sqrt{\sigma_d^{2'}} \implies (0.08 \pm 2.262\sqrt{0.592}) \implies (-1.66, 1.82)$$

Suppose we would like to test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 > \mu_2$. This is equivalent to testing $H_0 : \mu_d \leq 0$ versus $H_1 : \mu_d > 0$

$$P(\mu_d \leq 0) \approx 0.4586$$

```
pnorm(0, 0.08, sqrt(0.592))
```

```
## [1] 0.4585946
```

We fail to reject H_0, therefore, the data is not sufficient to indicate that $\mu_1 > \mu_2$

Comparing two means based on paired observations

- *Matched-pairs design*: experimental units are matched into pairs of similar units.
  - One of the units from each pair is assigned at random to the 1st treatment, and the other to the 2nd treatment
- *Repeated-measures design*: one set of experimental subjects were observed (data is collected) at two or more time periods.
  - The simplest case is the pre-test-post-test design wherein a pre evaluation is administered before a treatment (e. g. training) is conducted and the same (similar) evaluation is administered after the treatment
- We compute the pairwise differences $d_i = y_{1i} - y_{2i}$ and assume that $d_i \sim N(\mu_d, \sigma_d^2)$ with σ_d^2 known
- We can treat the differences d_i as a sample from a normal distribution and do Bayesian inference using techniques in the previous lessons.

Comparing two means based on paired observations: an example

An experiment was designed to determine whether a mineral supplement was effective in increasing annual yield in milk. Fifteen pairs of identical twin dairy cows were used as the experimental units. One cow from each pair was randomly assigned to the treatment group that received the supplement. The other cow from the pair was assigned to the control group that did not receive the supplement. The data is shown below.

Twin Set	Milk Yield: Control (liters)	Milk Yield: Treatment (liters)
1	3525	3340
2	4321	4279
3	4763	4910
4	4899	4866
5	3234	3125
6	3469	3680
7	3439	3965
8	3658	3849
9	3385	3297
10	3226	3124
11	3671	3218
12	3501	3246
13	3842	4245
14	3998	4186
15	4004	3711

```
ctrl <- c(3525, 4321, 4763, 4899, 3234, 3469, 3439,
          3658, 3385, 3226, 3671, 3501, 3842, 3998, 4004)
trt <- c(3340, 4279, 4910, 4866, 3125, 3680, 3965,
         3849, 3297, 3124, 3218, 3246, 4245, 4186, 3711)
d <- trt - ctrl
print(round(mean(d),3))
```

```
## [1] 7.067
```

Suppose $\sigma_d^2 = 270^2$. Let the normal prior be $N(0, 200^2)$.

The posterior distribution is $N(\mu^{'}, \sigma^{2'})$, where

$$\mu^{'} = \frac{\sigma_d^2\mu_p + n\sigma_p^2\bar{y}_d}{\sigma_d^2 + n\sigma_p^2} = \frac{270^2(0) + 15(200^2)(7.067)}{270^2 + 15(200^2)} \approx 6.30$$

and

$$\sigma^{2'} = \frac{\sigma_d^2\sigma_p^2}{\sigma_d^2 + n\sigma_p^2} = \frac{270^2(200^2)}{270^2 + 15(200^2)} \approx 4333.482$$

NOTE:

When σ_d^2 is unknown, we estimate it using the sample variance of the pairwise differences, consequently, the t distribution ($df = n - 1$) will be used in creating the credible interval.