1 Microbiology

A laboratory wants **to determine if two different methods (A and B) give similar results** for quantifying a particular bacterial species in a particular medium. Under each method, the counts form a random sample. **Assume that the counts follow a Poisson distribution**, since this distribution is a typical model for such data. Let μ_A and μ_B represent the population mean counts for Methods A and B, respectively. Let $\theta = \mu_A - \mu_B$.

1.1 Bootstrap estimate of SE

Use the bootstrap to estimate the standard error of $\hat{\theta} = \mu_A - \mu_B$, where μ_A and μ_B are sample means of counts for methods A and B, respectively. Use $B = 2000$ bootstrap samples. Assume the replicates are **unpaired**.

- Let $A = X$ and $Y = B$. I am changing the labeling so as not to confuse with the number of bootstrap samples, B.
- sampling is unpaired
- Poisson distribution is assumed. Hence, this is parametric bootstrap. Recall:

$$\mathcal{P}(X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}; \quad x = 0, 1, \ldots; \quad 0 \leq \lambda < \infty \tag{1.1}$$

- Following a fact mentioned on page 237 from Efron & Hastie (2017), a sufficient statistic for the parameter of the Poisson distribution λ is the average of the observations $\sum \frac{x}{n}$.

Algorithm:

Step 1: Calculate a sufficient statistic for λ_X (method A) and λ_Y (method B).

$$\begin{aligned} \hat{\lambda}_X &= \frac{\sum_{i=1}^{n} x_i}{n} \\ \hat{\lambda}_Y &= \frac{\sum_{i=1}^{n} y_i}{n} \end{aligned} \tag{1.2}$$

Step 2: Let B be the number of bootstrap samples taken. With $n = 8$ and for each method, sample from the plug-in distribution.

$$\begin{aligned} B^*_{Y,1} &= \left\{y^*_{1,1}, \ldots, y^*_{1,8}\right\} \overset{iid}{\sim} \mathcal{P}\left(\hat{\lambda}_Y\right) \\ B^*_{Y,2} &= \left\{y^*_{2,1}, \ldots, y^*_{2,8}\right\} \overset{iid}{\sim} \mathcal{P}\left(\hat{\lambda}_Y\right) \\ \vdots &= \vdots \\ B^*_{Y,2000} &= \left\{y^*_{2000,1}, \ldots, y^*_{2000,8}\right\} \overset{iid}{\sim} \mathcal{P}\left(\hat{\lambda}_Y\right) \end{aligned} \tag{1.3}$$

$$\begin{aligned} B^*_{X,1} &= \left\{x^*_{1,1}, \ldots, x^*_{1,8}\right\} \overset{iid}{\sim} \mathcal{P}\left(\hat{\lambda}_X\right) \\ B^*_{X,2} &= \left\{x^*_{2,1}, \ldots, x^*_{2,8}\right\} \overset{iid}{\sim} \mathcal{P}\left(\hat{\lambda}_X\right) \\ \vdots &= \vdots \\ B^*_{X,2000} &= \left\{x^*_{2000,1}, \ldots, x^*_{2000,8}\right\} \overset{iid}{\sim} \mathcal{P}\left(\hat{\lambda}_X\right) \end{aligned} \tag{1.4}$$

Step 3: For each of the bootstrap samples, calculate $\hat{\theta}^*_b = \hat{\mu}^*_{X,b} - \hat{\mu}^*_{Y,b}; \quad b = 1, \ldots 2000.$

Step 4: Calculate

$$\widehat{se}\left(\hat{\theta}\right) = \sqrt{\frac{\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \bar{\hat{\theta}}^*\right)^2}{B-1}}; \quad \bar{\hat{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^* \tag{1.5}$$

CODE FILENAME: ../R/01_01_se_estimate.R----------

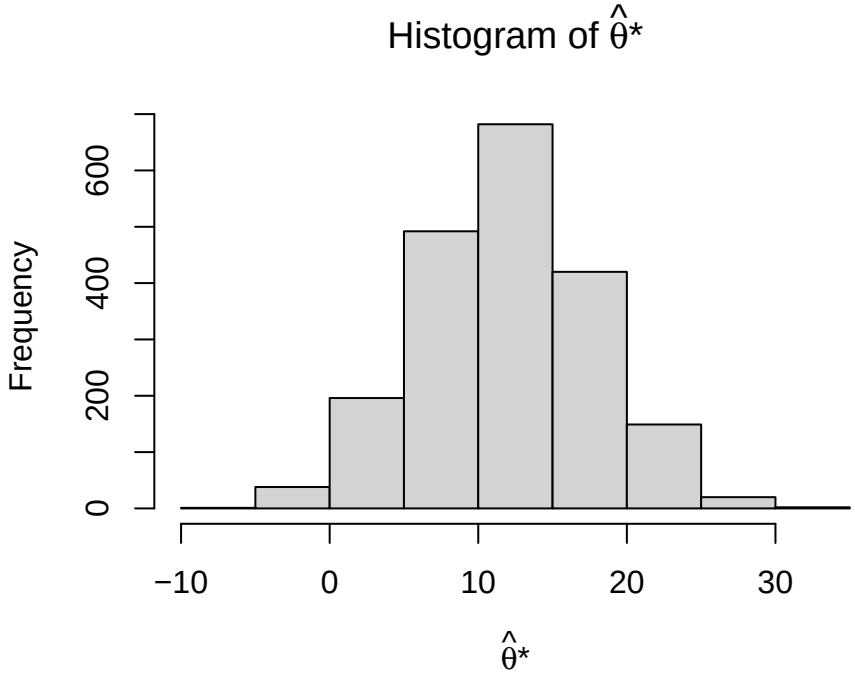
```
get_theta_hat_star <- function(sufficient_stat = 'average') {

  # step 1
  if (sufficient_stat == "average") {
    lambda_X <- mean(microbiology$X)
    lambda_Y <- mean(microbiology$Y)
  } else if (sufficient_stat == "sum")  {
    lambda_X <- sum(microbiology$X)
    lambda_Y <- sum(microbiology$Y)
  }

  # step 2
  set.seed(seed)
  B_starX <- replicate(B,rpois(n, lambda_X))
  set.seed(seed)
  B_starY <- replicate(B,rpois(n, lambda_Y))

  # step 3
  theta_hat_star <- colMeans(B_starX) - colMeans(B_starY)

  # step 4
  return(theta_hat_star)
}

suff_theta_star <- get_theta_hat_star("average")

suff_se <- sd(suff_theta_star)
#hist(
#  suff_theta_star,
#  main = expression(paste("Histogram of ", hat(theta), "*")),
#  xlab = expression(paste(hat(theta), "*")),
#)
suff_ci <- c(
  quantile(suff_theta_star,.025),
  quantile(suff_theta_star,.975)
  )
```

end----------

The bootstrap estimate of the standard error is 5.8341145.

1.2 Histogram

The histogram has a symmetric bell curve shape. This is expected to be centered on the statistic. In this case the statistic $\hat{\theta}$ is 12 while the mean of $\hat{\theta}^*$'s is 12.0179375.



1.3 Confidence Intervals

We are 95% confident that the true value of θ is between 0.625 and 23.375. From my perspective, the interval is broad since the difference ranges from almost zero to a relatively large positive number. However, this gives us an idea that the difference is positive.

2 Fishery

The dataset fishery.csv contains 40 annual counts of the numbers of recruits (R) and spawners (S) in a salmon population. The **units are in thousands of fish**. Recruits are fish that enter the catchable population. Spawners are fish that are laying eggs. Spawners die after laying eggs. The classic model for the relationship between spawners and recruits is

$$R = \frac{1}{\beta_0 + \frac{\beta_1}{S}}; \quad \beta_0 \geq 0 \text{ and } \beta_1 \geq 1$$

We can fit such a model by using a linear regression given by

$$\frac{1}{R_i} = \beta_0 + \beta_1 \frac{1}{S_i} + \epsilon_i; \quad i = 1, \ldots, 40$$

The **S variable can be considered as fixed**. Suppose that fish are iid with mean 0 and finite variance, but their **distribution is unknown**. The total population abundance will only stabilize if $R = S$. Thus, the stable population level is the point where the line $R = S$ intersects the curve relating R and S. The total population will decline if fewer recruits are produced

than the number of spawners who died producing them. If too many recruits are produced, the population will also decline eventually because there is not enough food for them all. Thus, only some middle level of recruits can be sustained indefinitely in a stable population. Let S_0 be the value of S at the stable population level.

Load the data and set the needed variables.

```
CODE FILENAME: ../R/02_00_load_data.R----------------------------------------
```

```
    fishery <- (
      read.csv("../../../problems/midterm/datasets/fishery.csv"
               )*1000)^(-1)
    colnames(fishery) <- c("Y", "X")

    n <- nrow(fishery)
    X <- as.matrix(cbind(1,fishery$X))
    y <- as.matrix(fishery$Y)
```

```
end--------------------------------------------------------------------------
```

2.1 Estimate S_0

Fit the regression model and estimate S_0. Call this estimate $\hat{S}_0$.

```
CODE FILENAME: ../R/02_01_stable.R-------------------------------------------
```

```
    # OLS estimate
    beta_hat <- solve(t(X)%*%X)%*%t(X)%*%y

    # stable state values
    R_0 <- 1/(beta_hat[1]/(1-beta_hat[2]))
    S_0 <- 1/(((beta_hat[1]/(1-beta_hat[2]))-beta_hat[1])/beta_hat[2])
```

```
end--------------------------------------------------------------------------
```

The above code yields $\hat{\beta}_0 = 2.0132307 \times 10^{-6}$ and $\hat{\beta}_1 = 0.6978188$.

Stable state: Let $S_0 = R_0 \to \frac{1}{S_0} = \frac{1}{R_0} \leftrightarrow x_0 = y_0$:

$$\begin{aligned} y_0 &= \hat{\beta}_0 + \hat{\beta}_1 x_0 \\ y_0 &= \hat{\beta}_0 + \hat{\beta}_1 y_0 \\ y_0 - \hat{\beta}_1 y_0 &= \hat{\beta}_0 \\ y_0(1 - \hat{\beta}_1) &= \hat{\beta}_0 \\ \hat{y}_0 &= \frac{\hat{\beta}_0}{1 - \hat{\beta}_1} \end{aligned} \tag{2.1}$$

It follows by substitution and more algebra that,

$$\begin{aligned} \frac{\hat{\beta}_0}{1 - \hat{\beta}_1} &= \hat{\beta}_0 + \hat{\beta}_1 x_0 \\ \frac{\frac{\hat{\beta}_0}{1-\hat{\beta}_1} - \hat{\beta}_0}{\hat{\beta}_1} &= \hat{x}_0 \end{aligned} \tag{2.2}$$

Hence,

$$(\hat{x}_0, \hat{y}_0) = \left(\frac{\frac{\hat{\beta}_0}{1-\hat{\beta}_1} - \hat{\beta}_0}{\hat{\beta}_1}, \frac{\hat{\beta}_0}{1-\hat{\beta}_1} \right) \iff$$
$$(\hat{S}_0, \hat{R}_0) = \left(\frac{1}{\frac{\frac{\hat{\beta}_0}{1-\hat{\beta}_1} - \hat{\beta}_0}{\hat{\beta}_1}}, \frac{1}{\frac{\hat{\beta}_0}{1-\hat{\beta}_1}} \right) \tag{2.3}$$

Substituting the values of $\hat{\beta}_0$ and $\hat{\beta}_0$ to equation 2.3, we get $\hat{S}_0 = 1.5009763 \times 10^5$ and $\hat{R}_0 = 1.5009763 \times 10^5$; they are equal, by definition of stable state.

2.2 Bootstrap estimate of SE

Use the bootstrap to obtain an estimate of the standard error of $\hat{S}_0$. Use $B = 2000$ bootstrap samples.

- This is nonparametric bootstrap since it is mentioned that no distribution is assumed.
- It is also mentioned that S variable can be considered as fixed; as a result, we will do SRSWR from the empirical distribution of the residuals.

The algorithm is already given in the regression handout, case 1a. In addition to that, we have to calculate the value $\hat{S}_0^*$, through equation 2.3, but for each bootstrap sample.

CODE FILENAME: ../R/02_02_se_estimate.R----------

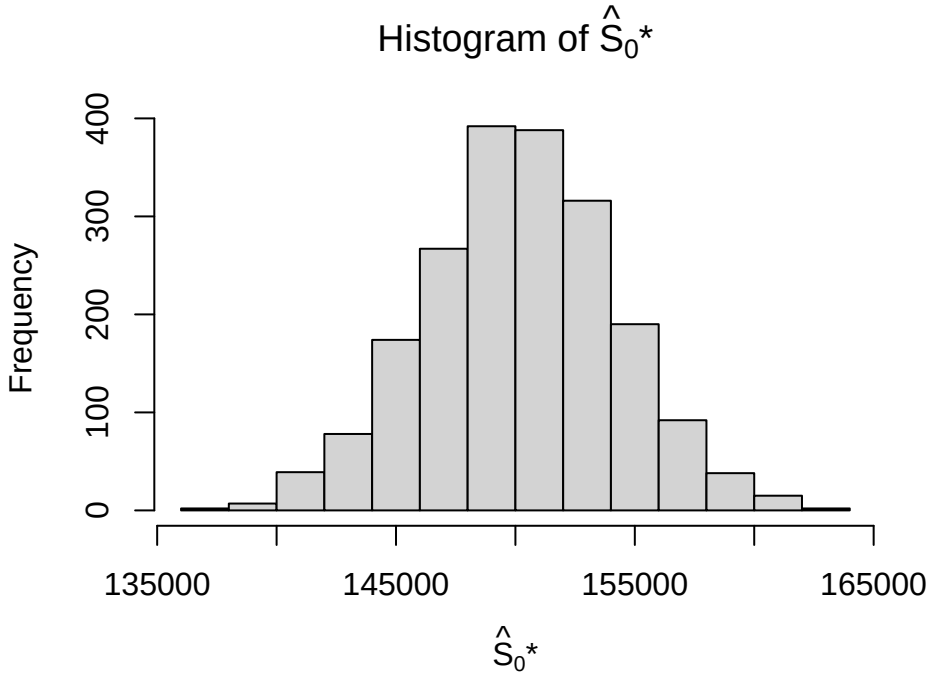
```
y_hat <- X%*%beta_hat
resid <- y - y_hat

S_0_star <- c()
for(b in 1:2000){
  # take bootstrap sample
  resample_star <- sample(resid,n,replace=T)

  # calculate new y from the resample
  y_star <- y_hat + resample_star

  # solve for the OLS estimate
  beta_hat_star <- solve(t(X)%*%X)%*%t(X)%*%y_star

  # use the OLS estimate to calculate the stable state value
  S_0_star[b] <- 1/(
    (
      (beta_hat_star[1]/(1-beta_hat_star[2]))-beta_hat_star[1]
      )/beta_hat_star[2]
    )
}

stable_se <- sd(S_0_star)
#hist(
#  S_0_star,
```

```
      #  main = expression(paste("Histogram of ", hat(S)[0], "*")),
      #  xlab = expression(paste(hat(S)[0], "*")))
      stable_ci <- c(
        quantile(S_0_star,.025),
        quantile(S_0_star,.975)
        )

  end----------------------------------------------------------------------
```

The bootstrap estimate of the standard error is 4033.7155849.

2.3 Histogram

The histogram has a symmetric bell curve shape. This is expected to be centered around $\hat{S}_0 = 1.5009763 \times 10^5$. Indeed, the mean of the distribution below is 1.501892×10^5.



2.4 Confidence Intervals

We are 95% confident that the true value of θ is between 1.4206151×10^5 and 1.5825922×10^5. Since the values being dealt with are large, in my perspective, interval is neither too broad nor too narrow.

3 References

Efron, B., & Hastie, T. (2017). *Computer age statistical inference*. Cambridge University Press.

4 Appendix

4.1 Code to read data for item 1

```
CODE FILENAME: ../R/01_00_load_data.R----------------------------------------

    data <- "../../../problems/midterm/datasets/microbiology.RData"

    if (file.exists(data)) {
      print(paste(c("The file exists; loading", data), collapse = ' '))
      load(data)
    } else {
      paste(c("The file does not exist; creating, loading and saving", data),
            collapse = ' ')
      microbiology <- data.frame(
        'i' = 1:8,
        'X' = c(176, 125, 152, 180, 159, 168, 160, 151),
        'Y' = c(164, 121, 137, 169, 144, 145, 156, 139)
      )
      n <- dim(microbiology)[1]
      seed <- 7
      B <- 2000
      save(microbiology, n, B, seed,file = data)
    }

    rm(data)


end--------------------------------------------------------------------------
```