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 the *Factorization Theorem* (see References), a sufficient statistic for the parameter of the Poisson distribution, λ is $\sum_{i=1}^{n} x_i$. Since it is sufficient to know the average if we know the sum (given the sample), then we can also take the mean as another sufficient statistic.

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

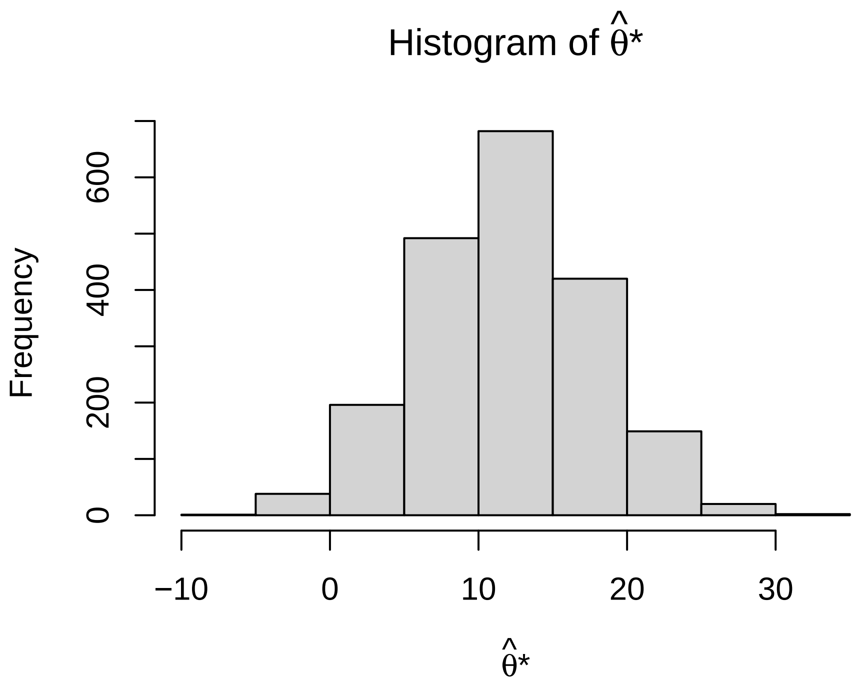
$$\widehat{se}\left(\hat{\theta}\right) = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{\theta}_i^* - \bar{\hat{\theta}}^*\right)^2}{B-1}}; \quad \bar{\hat{\theta}}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^* \tag{1.5}$$

```
[1] 7
```

The bootstrap estimate of the standard error is 5.83411.

1.2 Histogram

The histogram has a symmetric bell curve shape. I recalled in class that this is expected to be centered on the statistic. In this case the statistic is 12 while the mean of $\hat{\theta}^*$'s is 12.01794 This is the reason why I chose the mean for the plug-in distribution. If the plug-in distribution is based on the sum, the histogram will be centered on a different, larger value, far from the statistic. It will also have a higher estimate of the standard error (9.92229).



1.3 Confidence Intervals

We are 95% confident that the true value of θ is between 0.625 and 23.375. If instead we use sum for the sufficient statistic, then the CI will be a lot wider, with the sample statistic, not even making it within the interval 73.24688 and 117.625.

2 Fishery

3 References

Factorization theorem. https://online.stat.psu.edu/stat415/lesson/24/24.2

4 Appendix

4.1 Code to read data for items 1 & 2

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

```
# wd: /home/scientists/sci01/Projects/bootstrap/solutions/ps_01/child

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)
```