# R Assignment Two

Katherine Wolf

February 11, 2020

## Question One

### State the point estimate and 95% confidence interval (CI) for the estimated risk difference (RD) and risk ratio (RR) for this analysis. (20 points)

The point estimate for the estimated risk (cumulative incidence) difference derived using model-based standardization,[1] which compares the modeled cumulative incidence of hypertension over the study period in a hypothetical study population in which everyone is obese to that in a hypothetical study population in which everyone is "ideal" weight, is 0.234 (95% CI: 0.169, 0.292).[2]

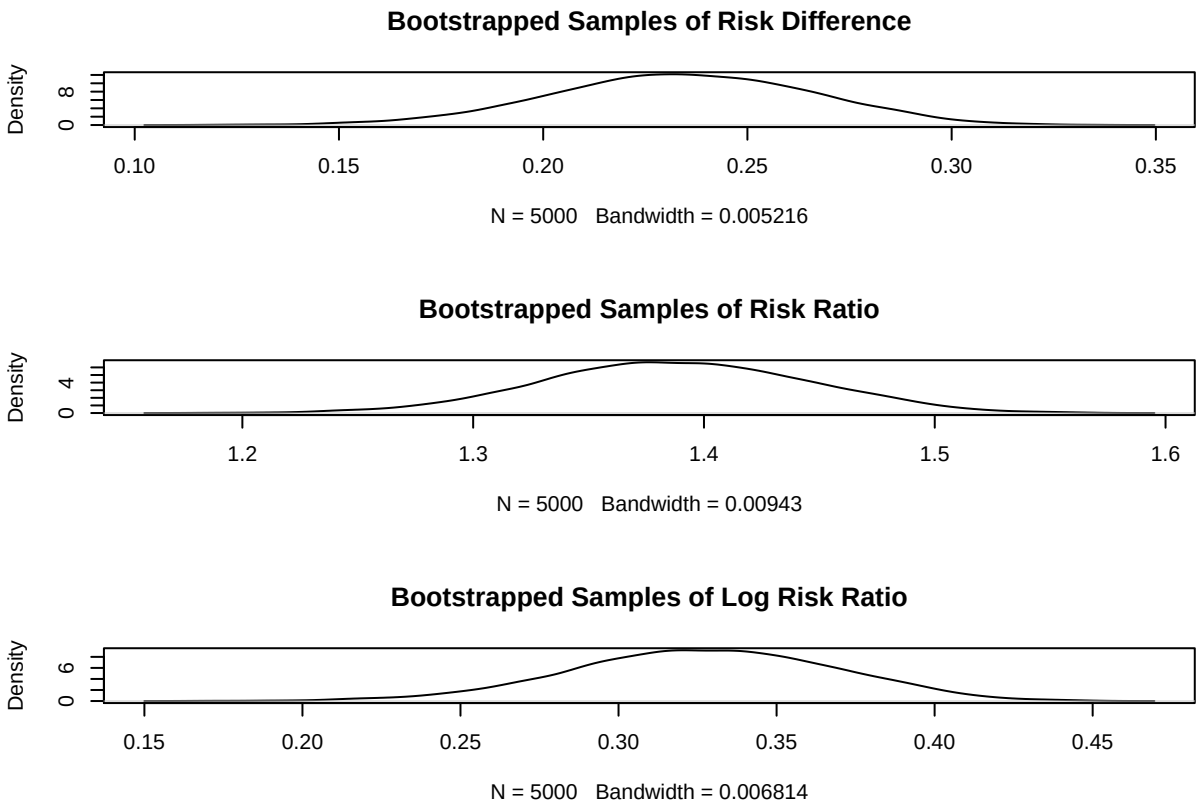
The point estimate for the estimated risk (cumulative incidence) ratio (RR) derived using model-based standardization, which compares the modeled cumulative incidence of hypertension over the study period in a hypothetical study population in which everyone is obese to that in a hypothetical study population in which everyone is "ideal" weight, is 1.388 (95% CI: 1.273, 1.497).

[1] Greenland S. Model-based Estimation of Relative Risks and Other Epidemiologic Measures in Studies of Common Outcomes and in Case-Control Studies. Am J Epidemiol 2004;160(4):301–5. doi:10.1093/aje/kwh221.

[2] I calculated the CIs around the model-standardized point estimates of the RD and the RR using the bias-corrected and accelerated method in the `boot` R package, which corrects for both (1) the potential difference between the mean of all the bootstrap-sample-derived point estimates used to construct the 95% CI and the point estimate calculated from the original dataset, caused by asymmetry in the sampling distribution (the "bias-corrected" piece), as well as (2) changes in skewness in the distribution that vary with the value of the estimated parameter (the "accelerated" piece). See Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine 2000;19(9):1141–1164. doi:10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F. (Page 1154.)

## Question Two

**Turn in the density plots for the risk difference (RD) and risk ratio (RR) measures. Briefly describe the pattern that you see. Does the pattern seem to indicate that the sampling distribution of the RD and RR are approximately normal? (10 points)**




The density plots for the sampling distributions for the RD and the RR drawn from bootstrapped samples of the data look normal and centered close to the point estimates from question one, as does the density plot for the sampling distribution of the natural logarithm of the RR (log RR), although all three are slightly negatively skewed. Moreover, the distribution of bootstrap sample RDs ranges from 0.10 to 0.35, the RRs from a little above 1.1 to 1.6, and the log RRs from 0.15 to 0.48, which means that 95% CIs are unlikely to include invalid values, i.e., RDs $> 1$ or $< -1$, RRs $< 0$, or log RRs $\leq 0$. Thus we can likely use normal approximations to calculate 95% CIs for the RD, the RR, and the log RR. (I didn't quite understand why the original assignment had us look at a density plot for the bootstrap sampling distribution of the RR if we're calculating 95% CIs on the scale of the log RR, so I included a plot of the distribution of the log RRs as well.)

# Question Three

**Showing your work, calculate the 95% confidence interval for the risk difference (RD) using a normal approximation, and report your answer. (Show the formula you used, and the specific values that went into your calculation.) What assumptions does this require? How does this compare to the results from question one? (10 points)**

To simplify notation:

- Let the point estimate of the RD derived via model-based standardization, $\hat{RD}$, be denoted by $\hat{\theta}$;
- Let $n$ denote the number of bootstrapped samples indexed by $i$ (here from $i = 1$ to $i = 5000$); and
- Let each estimate of the model-standardized RD derived from a bootstrap sample of the original data be denoted by $\hat{\theta}_i$.

To calculate the 95% CI for the model-standardized estimate of the RD using the Wald normal approximation

$$\hat{\theta} \pm 1.96 \times SE_{\hat{\theta}_i},$$

**Step One**

I first calculated the mean of the differences between the original model-standardized point estimate of the RD $\hat{\theta} = 0.23442$ and the bootstrapped point estimates of the model-standardized RD $\hat{\theta}_i$, $\overline{\Delta_{\hat{\theta}_i}}$:

$$\overline{\Delta_{\hat{\theta}_i}} = \frac{\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta})}{n} = \frac{\sum_{i=1}^{5000}(\hat{\theta}_i - 0.23442)}{5000} = -6.15 \times 10^{-4}$$

**Step Two**

I used that mean $\overline{\Delta_{\hat{\theta}_i}}$ to calculate the sample standard deviation of the differences between the $\hat{\theta}_i$s and $\hat{\theta}$, in order to use it as the best possible estimate of the standard error (SE) of that estimate, $SE_{\hat{\theta}_i}$:

$$\begin{aligned} SE_{\hat{\theta}_i} \approx sd(\hat{\theta}_i - \hat{\theta}) &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\left(\hat{\theta}_i - \hat{\theta}\right) - \overline{\Delta_{\hat{\theta}_i}}\right)^2} \\ &= \sqrt{\frac{1}{5000-1}\sum_{i=1}^{5000}\left(\left(\hat{\theta}_i - 0.23442\right) - (-6.1 \times 10^{-4})\right)^2} = 0.03183 \end{aligned}$$

(In truth I did both steps one and two using the R one-liner `sd.rd.boot <- sd(rd.samples - stdized.rd)`.)

**Step Three**

I calculated the Wald 95% CI for the RD using the estimated SE above:

$$95\%\ CI_{\hat{\theta}} = \hat{\theta} \pm 1.96 * SE_{\hat{\theta}_i} = 0.23442 \pm 1.96 \times 0.032 = (0.172, 0.297)$$

**Assumptions**

For Wald CIs for the point estimate of the RD to be valid, the following assumptions must hold:

1. The underlying population in the study must have been sampled randomly. (The bootstrapped samples from the study observations must also be drawn randomly, although this assumption is automatically pseudo-met via R's pseudo-random number generator.) The original investigators met or didn't meet this assumption when they collected the data.

2. The underlying observations in the study population must be independent of each other. (The bootstrapped samples from the study observations must also be independent of each other, although this assumption is also pseudo-met via R's pseudo-random number generator.) The original investigators met or didn't meet this assumption when they collected the data.

3. The original dataset from which the bootstrap samples are drawn must contain sufficiently large numbers of cases and non-cases in both the exposed and unexposed categories.[3] Here we have

    - 150 obese participants with hypertension (exposed cases);
    - 27 obese participants without hypertension (exposed non-cases);
    - 706 "ideal"-weight participants with hypertension (unexposed cases);
    - 481 "ideal"-weight participants without hypertension (unexposed cases); so this assumption also appears met.

4. The distribution of the RDs calculated from the bootstrapped samples, i.e., the sampling distribution, must be asymptotically normal, as confirmed by the plot of the distribution of the RDs in step two. In particular, it must be symmetric.(This does not necessarily mean that the distribution of the original data has to be normal, only that the distribution of the RDs calculated from the bootstrapped samples has to be.)

5. The point estimate of the RD must be (between but) not too close to -1 or 1, particularly in small samples, lest the CIs inaccurately include invalid RDs $< -1$ or $> 1$.[4]

**Comparison of normal and bias-corrected and accelerated (BCA) CI**

The 95% CI for the RD using a Wald normal approximation is (0.172, 0.297), compared to the bias-corrected and accelerated CI of (0.169, 0.292). The Wald 95% CI is very close to the BCA CI, as expected given the large size of the dataset and normal distribution of the RDs calculated from bootstrap samples of the data. Moreover, the Wald CI is slightly higher than the BCA CI due to its assumption of symmetry around the point estimate, which ignores the slight left skew of the data. Thus the Wald CI effectively includes a bit too much of the right "hill" and too little of the left "tail" of the distribution compared to the BCA (which is essentially derived from percentiles of the actual data after correcting for bias and skewness).

[3] Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. International Journal of Epidemiology 2004;33(6):1389–97. doi:10.1093/ije/dyh276. (Page 390.)

[4] Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine 2000;19(9):1141–1164. doi:10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F. (Page 1150.)

# Question Four

**Showing your work, calculate the 95% confidence interval for the risk ratio (RR), and report your answer. (Show the formula you used, and the specific values that went into your calculation.) What assumptions does this require? How does this compare to the results from question one? (10 points)**

To simplify notation:

- Let the point estimate of the RR derived via model-based standardization, $\hat{RR}$, be denoted by $\hat{\theta}$;
- Let $n$ denote the number of bootstrapped samples indexed by $i$ (here from $i = 1$ to $i = 5000$); and
- Let each estimate of the model-standardized RR derived from a bootstrap sample of the original data be denoted by $\hat{\theta}_i$.

To calculate the 95% CI for the model-standardized estimate of the RR using the Wald normal approximation

$$\hat{\theta} \pm 1.96 \times SE_{\hat{\theta}_i},$$

**Step One**

I first calculated the mean of the differences between the natural logarithm of the original model-standardized point estimate of the RR $\ln(\hat{\theta}) = \ln(1.388) = 0.32785$ and the natural logarithms of the bootstrapped point estimates of the model-standardized RR, $\overline{\Delta_{ln(\hat{\theta}_i)}}$:

$$\overline{\Delta_{ln(\hat{\theta}_i)}} = \frac{\sum_{i=1}^{n}(\ln(\hat{\theta}_i) - \ln(\hat{\theta}))}{n} = \frac{\sum_{i=1}^{5000}(\ln(\hat{\theta}_i) - 0.32785)}{5000} = -0.00108$$

**Step Two**

I used that mean $\overline{\Delta_{\ln(\hat{\theta}_i)}}$ to calculate the sample standard deviation of the differences between the $\ln(\hat{\theta}_i)$s and $\ln(\hat{\theta})$, in order to use it as the best possible estimate of the standard error of that estimate, $(SE_{\ln(\hat{\theta}_i)})$:

$$SE_{\ln(\hat{\theta}_i)} \approx sd(\ln(\hat{\theta}_i) - \ln(\hat{\theta})) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\left(\ln(\hat{\theta}_i) - \ln(\hat{\theta})\right) - \overline{\Delta_{\ln(\hat{\theta}_i)}}\right)^2}$$

$$= \sqrt{\frac{1}{5000-1}\sum_{i=1}^{5000}\left(\left(\ln(\hat{\theta}_i) - 0.32785\right) - (-0.00108)\right)^2} = 0.04159$$

(I did both steps one and two using the R one-liner `sd.log.rr.boot <- sd(log(rr.samples) - log(stdized.rr))`.)

**Step Three**

I calculated the Wald 95% CI for the natural logarithm of the RR using the estimated standard error above:

$$95\%CI_{ln(\hat{\theta})} = \ln(\hat{\theta}) \pm 1.96 * SE_{\ln(\hat{\theta}_i)} = 0.32785 \pm 1.96 \times 0.04159 = (0.246, 0.409)$$

**Step Four**

I then exponentiated the CI to get it on the scale of the original RR:

$$95\%\ CI_{\hat{\theta}} = e^{\ln(\hat{\theta}) \pm 1.96 * SE_{\ln(\hat{\theta}_i)}} = (e^{0.246}, e^{0.409}) = (1.279, 1.506)$$

**Assumptions**

For Wald CIs for the point estimate of the RR to be valid, the following assumptions must hold:

1. Assumption 1 from Question 3.
2. Assumption 2 from Question 3.
3. Assumption 3 from Question 3.
4. The distribution of the natural logarithms of the RRs calculated from the bootstrapped samples, i.e., the sampling distribution, must be asymptotically normal, as confirmed by the plot of the distribution of the log RRs in step two. In particular, it must be symmetric. (This does not necessarily mean that the distribution of the original data has to be normal, only that the distribution of the log RRs calculated from the bootstrapped samples has to be.)[5]
5. The RR cannot be too close to zero, particularly in small samples, lest the CIs inaccurately include invalid negative RRs.[6]

**Comparison of normal and BCA CIs**

The 95% CI for the RR using a Wald normal approximation is (1.279, 1.506), compared to the bias-corrected and accelerated CI of (1.273, 1.497). The Wald 95% CI for the RR is again very close to the BCA CI, as expected given the large size of the dataset and normality of the RR bootstrap sampling distribution. Moreover, the Wald CI is again slightly higher than the BCA CI due to its assumption of symmetry around the point estimate, which ignores the slight left skew of the data. Thus the Wald CI effectively includes a bit too much of the right "hill" and too little of the left "tail" of the distribution compared to the BCA (which is essentially derived from percentiles of the actual data after correcting for bias and skewness).

[5] Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. International Journal of Epidemiology 2004;33(6):1389–97. doi:10.1093/ije/dyh276. (Page 390.)

[6] Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine 2000;19(9):1141–1164. doi:10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F. (Page 1150.)

```r
knitr::opts_chunk$set(echo = FALSE,
                      warning = FALSE,
                      message = FALSE)


library(foreign)
library(boot)
load("frmgham_recoded.wide.Rdata")


# Make 2nd category (BMI 18.5-24.9, ideal weight) the referent group:
frmgham_recoded.wide$bmi_cat <-
  relevel(as.factor(frmgham_recoded.wide$bmi_cat),"2")


standardized.measures <- function(dataset, index){
  # Create resampled version of dataset using index vector:
  dataset.resampled <- dataset[index,];

  # Fit logistic regression from the resampled dataset:
  # Outcome: hyperten
  # Covariates: bmi_cat, cursmoke, age, sex, educ
  # Store results in object called logistic.frmgham

  logistic.frmgham <- ##### COMPLETE ON OWN
    glm(hyperten ~ bmi_cat + cursmoke + age + sex + educ,
        data = dataset.resampled,
        family = binomial(link = "logit"))

  ##### end complete on own

  ### To calculate the measures of association:

  # STEP 1: Create two new versions of the *original* dataset, called
  # frmgham_recoded.wide.obese and frmgham_recoded.wide.ideal
  frmgham_recoded.wide.obese <-
    frmgham_recoded.wide.ideal <-
    frmgham_recoded.wide

  # Set BMI to obese (category 4) in population 1 [exposed]
  # and ideal weight (category 2) in population 0 [unexposed]:
  frmgham_recoded.wide.obese$bmi_cat <- "4" # population w/ all obese
  frmgham_recoded.wide.ideal$bmi_cat <- "2" # population w/ all ideal weight

  # STEP 2: Obtain predicted individual risk of hypertension under each new dataset:
  rhat.obese <- predict(logistic.frmgham, type = "response",
                        newdata = frmgham_recoded.wide.obese)
  rhat.ideal <- predict(logistic.frmgham, type = "response",
                        newdata=frmgham_recoded.wide.ideal)

  # STEP 3: Calculate the average risk (proportion) of hypertension
  # in each hypothetical population:
  risk.obese <- mean(rhat.obese)
```

```
  risk.ideal <- mean(rhat.ideal)

  # Calculate risk difference and risk ratio using ideal weight as reference:
  rd <- risk.obese - risk.ideal
  rr <- risk.obese/risk.ideal

  # STEP 4: Return these estimates:
  return(c(rd, rr))
}


n.frmgham <- nrow(frmgham_recoded.wide)
stdized.measures <- standardized.measures(frmgham_recoded.wide,
                                          index=seq(1:n.frmgham))
stdized.rd <- stdized.measures[1] # Risk difference is 1st element
stdized.rr <- stdized.measures[2] # Risk ratio is 2nd element


set.seed(123)
# Put the bootstrapped sample results into object called bs.standardized
bs.standardized <- boot(frmgham_recoded.wide, standardized.measures, 5000)

if(!exists("ci.rd.bca")) {
  ci.rd.bca <- boot.ci(bs.standardized, type= "bca", index=1)
}

if(!exists("ci.rr.bca")) {
  ci.rr.bca <- boot.ci(bs.standardized, type= "bca", index=2)
}


rd.samples <- bs.standardized$t[,1]  # risk difference
rr.samples <- bs.standardized$t[,2]  # risk ratio


# Calculate the standard deviation of each of the series of bootstrapped samples obtained in the above

sd.rd.boot <- sd(rd.samples - stdized.rd)

sd.log.rr.boot <- sd(log(rr.samples) - log(stdized.rr))

# Use these estimates of standard deviation to calculate 95 confidence intervals for the RD and RR usin

rd.ci.norm.apx <- c(stdized.rd - 1.96*sd.rd.boot,
                    stdized.rd + 1.96*sd.rd.boot)

log.rr.ci.norm.apx <- c(log(stdized.rr) - 1.96*sd.log.rr.boot,
                        log(stdized.rr) + 1.96*sd.log.rr.boot)

rr.ci.norm.apx <- exp(log.rr.ci.norm.apx)


par(mfrow=c(3,1))
```

```r
plot(density(rd.samples), main = "Bootstrapped Samples of Risk Difference")
plot(density(rr.samples), main = "Bootstrapped Samples of Risk Ratio")
plot(density(log(rr.samples)), main = "Bootstrapped Samples of Log Risk Ratio")
par(mfrow=c(1,1))


# risk difference comparisons

# true value 0.234

# bca ci 0.169, 0.292

# normal ci 0.172, 0.297

# bca middle
(0.169 + 0.292)/2

# bca size
0.292 - 0.169

# normal ci middle
(0.172 + 0.297)/2

# normal size
0.297 - 0.172

# risk ratio comparisons

# true value 1.388

# bca ci 1.273, 1.497

# normal ci 1.279, 1.506

# bca middle
(1.273 + 1.497)/2
# bca size
1.497 - 1.273

# normal ci middle
(1.279 + 1.506)/2

# normal ci size
1.506 - 1.279


library(tidyverse)

obese_hyper_count <-
  frmgham_recoded.wide %>%
  filter(bmi_cat == "4" & hyperten == 1) %>%
  nrow()
```

```r
obese_no_hyper_count <-
  frmgham_recoded.wide %>%
  filter(bmi_cat == "4" & hyperten == 0) %>%
  nrow()

ideal_hyper_count <-
  frmgham_recoded.wide %>%
  filter(bmi_cat == "2" & hyperten == 1) %>%
  nrow()

ideal_no_hyper_count <-
  frmgham_recoded.wide %>%
  filter(bmi_cat == "2" & hyperten == 0) %>%
  nrow()

obese_hyper_count
obese_no_hyper_count
ideal_hyper_count
ideal_no_hyper_count


log(stdized.rr)

mean(log(rr.samples) - log(stdized.rr))
```