

Problem set 8

Your name here

Due 11/19/2021 at 5pm

NOTE: Start with the file `ps8_2021.Rmd` (available from the github repository at <https://github.com/UChicago-pol-methods/IntroQSS-F21/tree/main/assignments>). Modify that file to include your answers. Make sure you can “knit” the file (e.g. in RStudio by clicking on the *Knit* button). Submit both the Rmd file and the knitted Pbb via Canvas.

In this assignment we will return to data from an experiment that measured the effect of constituent names in emails on legislator replies. The published paper is:

Butler, D. M., & Broockman, D. E. (2011). *Do politicians racially discriminate against constituents? A field experiment on state legislators*. AJPS.

The data file is `Butler_Broockman_AJPS_2011_public_csv.csv` and it is found in the `data/legislators_email` directory of the course github repository.

To load the data you can either download and read in the local file, or you can read in the url from github. Note that reading in by the url will only work when you have an internet connection:

```
file <- 'https://raw.githubusercontent.com/UChicago-pol-methods/IntroQSS-F21/main/data/legislators_email.csv'
bb <- read_csv(url(file))
```

Question 1: Inference from a single random variable

(1a) Create an object called `theta_hat` which is the mean of the `reply_atall` variable in the data set.

```
(theta_hat <- mean(bb$reply_atall))
```

```
## [1] 0.5653427
```

(1b) Create an object called `se_hat` which is the estimate of the standard error of the mean of the `reply_atall` variable in the data set, using the formula based on the unbiased sample variance.

```
(se_hat <- sqrt(var(bb$reply_atall)/nrow(bb)))
```

```
## [1] 0.007112145
```

(1c) The formula for the normal approximation-based confidence intervals is below

$$CI_n = \left(\hat{\theta}_n - z_{1-\alpha/2} \times \hat{se}, \hat{\theta}_n + z_{1-\alpha/2} \times \hat{se} \right)$$

z_c describes the c -th quantile of the standard normal distribution. For 95% confidence intervals, $\alpha = 0.05$, so we want to find $z_{1-\alpha/2} = z_{0.975}$. Using `qnorm`, get the 97.5-th quantile of the standard normal distribution.

```
(z975 <- qnorm(0.975))
```

```
## [1] 1.959964
```

(1d) Using `theta_hat`, `se_hat`, and your answer to the previous question, report the 95% normal approximation-based confidence intervals for the estimate of `theta_hat`

```
(CI95 <- c(theta_hat + c(-1,1)*z975*se_hat))
```

```
## [1] 0.5514031 0.5792822
```

(1e) Interpret what the 95% confidence interval means.

The confidence interval is a random interval which, across random samples, will include the true value of the estimand 95% of the time.

(1f) To get the 90% confidence intervals, we will set α as 0.10. So we want to find $z_{1-\alpha/2} = z_{0.95}$. Using `qnorm`, get the 95-th quantile of the standard normal distribution.

```
(z95 <- qnorm(0.95))
```

```
## [1] 1.644854
```

(1g) Using your answer from the question above, report the 90% normal approximation-based confidence intervals for the estimate of `theta_hat`.

```
(CI90 <- c(theta_hat + c(-1,1)*z95*se_hat))
```

```
## [1] 0.5536442 0.5770411
```

(1h) Create a vector of 1000 bootstrapped estimates of the sample mean of `reply_atall`. Save this vector as an object. Report the standard deviation across the estimates. The standard deviation of your bootstrapped estimates should be similar to your answer to 1b above.

Note: This should look very much like your solution to (2e) on hw 7, but you should be sampling with replacement from `bb$reply_atall`.

```
boot_ests <- map(1:1000, # for 1000 times
  # resample w/replacement
  ~ sample(bb$reply_atall, replace = TRUE) %>%
    mean()) # and calculate the resampled mean
```

```
boot_vec <- unlist(boot_ests)
```

```
sd(boot_vec)
```

```
## [1] 0.007266542
```

(1i) We can compare the distribution of the estimator under the bootstrap procedure and under the normal approximation. Using the `quantile()` function and your saved vector of 1000 bootstrapped estimates of the sample mean, report the 2.5th and 97.5th quantiles of the estimates under the bootstrap. These cover 95% of the empirical distribution of the bootstrap. How do they compare to your 95% normal approximation-based confidence intervals in your answer to 1d above?

```
quantile(boot_vec, probs = c(0.025, 0.975))
```

```
##      2.5%      97.5%
## 0.5511422 0.5793425
```

Question 2: Inference from linear models

(2a) Using `lm_robust`, regress `reply_atall` on `treat_deshawn` interacted with `leg_republican`. Print the model object. Save the vector of coefficients as `theta_hats`.

```
(lm_bb <- lm_robust(reply_atall ~ treat_deshawn*leg_republican, data = bb))

##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)    0.52976190 0.01361950 38.8973007 2.385445e-288
## treat_deshawn    0.01596300 0.01923401  0.8299363 4.066156e-01
## leg_republican    0.09949293 0.02000774  4.9727219 6.829449e-07
## treat_deshawn:leg_republican -0.07550407 0.02848382 -2.6507709 8.056896e-03
##              CI Lower   CI Upper   DF
## (Intercept)    0.50306151  0.55646230 4855
## treat_deshawn   -0.02174436  0.05367037 4855
## leg_republican    0.06026870  0.13871715 4855
## treat_deshawn:leg_republican -0.13134525 -0.01966290 4855

theta_hats <- coef(lm_bb)
```

(2b) From the model object above, report and interpret the standard errors and 95% confidence intervals on `treat_deshawn` and `treat_deshawn:leg_republican`. Do the confidence intervals include zero? If so/if not, what does that imply?

The standard error on `treat_deshawn` is 0.019; this is the estimated standard error of the treatment effect among Democrats. We estimate the 95% confidence interval of the estimate to be -0.022 to 0.054. The confidence interval *does* include zero, which means that we can not reject the null hypothesis of no treatment effect at a statistical significance level of $\alpha = 0.05$.

The standard error on `treat_deshawn:leg_republican` is 0.028; this is the estimated standard error of the treatment effect among Republicans. We estimate the 95% confidence interval of the estimate to be -0.131 to -0.02. The confidence interval *does not* include zero, which means that we can reject the null hypothesis of no treatment effect at a statistical significance level of $\alpha = 0.05$.

(2c) Using `map()` and `slice_sample()`, `replace = TRUE`), take 1000 bootstrap re-samples with replacement of the same size as the original data from the `bb` dataset. Save your bootstrapped samples as an object.

```
boot_samples <- map(1:1000, # for 1000 times
  # resample w/replacement
  ~ slice_sample(bb, replace = TRUE, n = nrow(bb)))
```

(2d) Using `map()` again, run the same regression as above on *each* of your bootstrapped samples; extract coefficient estimates; and use `bind_rows()` to create a matrix where each row represents estimates from one of your bootstrap samples, and each column is one of the coefficients.

```
boot_lm <- map(boot_samples,
  ~ lm_robust(reply_atall ~ treat_deshawn*leg_republican, data = .) %>%
    coef()) %>%
  bind_rows()
```

(2e) Report the bootstrapped estimates of the standard errors of each of the coefficients. To do this, get the standard deviations of each of the columns.

```
boot_se_hats <- boot_lm %>%
  map_dbl(sd)
```

(2f) Produce normal approximation-based confidence intervals for each of the coefficients using the bootstrapped standard errors, inserted into the same formula for confidence intervals as

presented in 1c. Compare these to the standard errors from your original `lm_robust()` model object in question 2a.

```
boot_ci <- bind_cols(term = names(theta_hats),
  est = theta_hats,
  boot_se = boot_se_hats) %>%
  mutate(conf_lower = est - qnorm(.975)*boot_se,
    conf_upper = est + qnorm(.975)*boot_se)
```

```
boot_ci
```

```
## # A tibble: 4 x 5
```

##	term	est	boot_se	conf_lower	conf_upper
##	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
## 1	(Intercept)	0.530	0.0136	0.503	0.556
## 2	treat_deshawn	0.0160	0.0190	-0.0213	0.0533
## 3	leg_republican	0.0995	0.0199	0.0606	0.138
## 4	treat_deshawn:leg_republican	-0.0755	0.0283	-0.131	-0.0200