Homework Exercises Week 06

Alison Lawyer

2024-10-22

```
# keep this chunk in all your RMarkdown scripts
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

```
# List required packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(PresenceAbsence)
library(ggplot2)
library(palmerpenguins)

data("SPDATA")
data("penguins")
```

HOMEWORK EXERCISES

Exercise 1

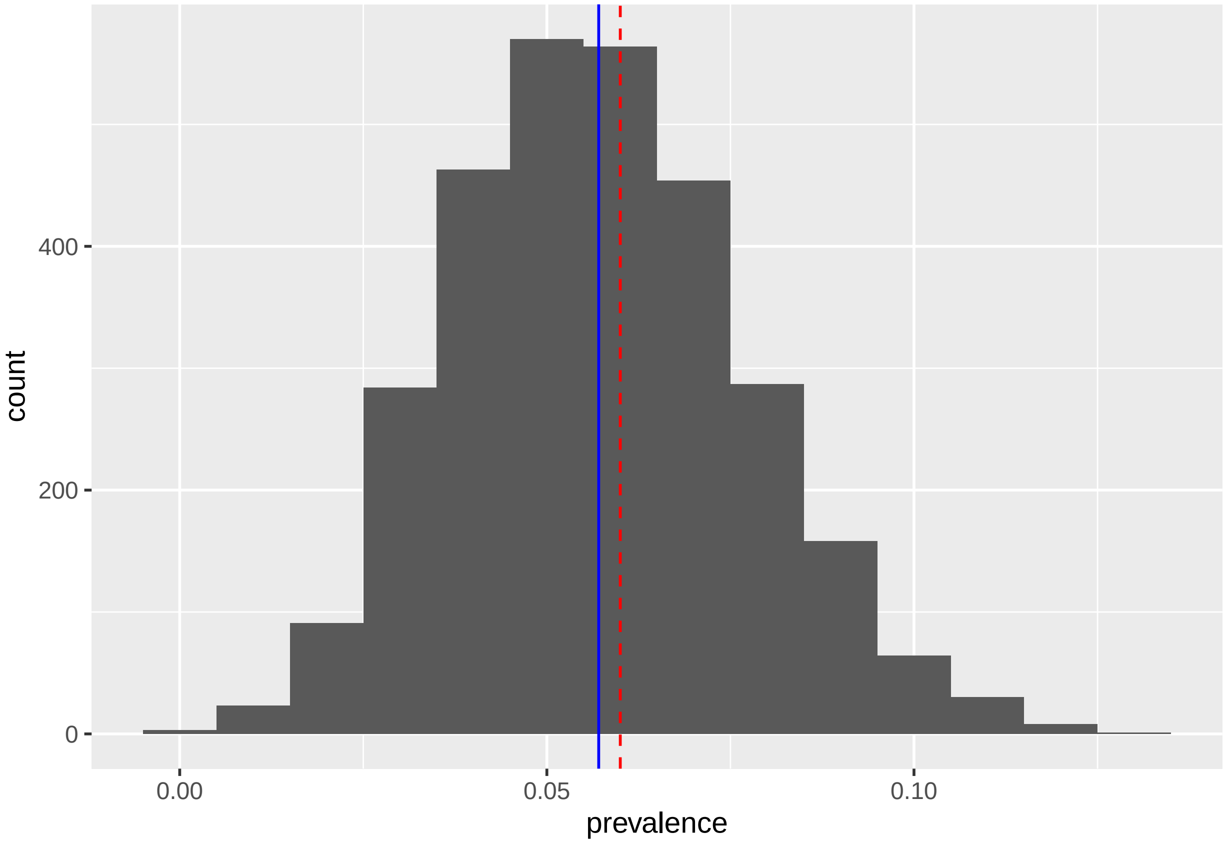
Building on exercise 4 from the lab, at what number of repeated random draws of 100 samples each does the resulting distribution become almost perfectly normal?

While keeping set.seed() the same, iterate through different values for your number of repeated samples, while keeping sample size at 100 sites each time. Keep only your last iteration that you feel best reflects a normal distribution of prevalence values. At this iteration, how does your mean prevalence across samples compare to the population prevalence?

```
set.seed(1234)
sample_size <- 100
num_samples <- 3000

filtered_dataset <- SPDATA %>%
    select(c(SPECIES, OBSERVED)) %>%
    filter(SPECIES == "ACGR3")
sample_n(filtered_dataset, size = sample_size, replace = FALSE)
```

```
##     SPECIES OBSERVED
## 1     ACGR3        0
## 2     ACGR3        0
## 3     ACGR3        0
## 4     ACGR3        0
## 5     ACGR3        0
## 6     ACGR3        0
## 7     ACGR3        0
## 8     ACGR3        1
## 9     ACGR3        0
## 10    ACGR3        0
## 11    ACGR3        0
## 12    ACGR3        0
## 13    ACGR3        0
## 14    ACGR3        0
## 15    ACGR3        0
## 16    ACGR3        0
## 17    ACGR3        0
## 18    ACGR3        0
## 19    ACGR3        0
## 20    ACGR3        0
## 21    ACGR3        0
## 22    ACGR3        0
## 23    ACGR3        0
## 24    ACGR3        0
## 25    ACGR3        0
## 26    ACGR3        0
## 27    ACGR3        0
## 28    ACGR3        0
## 29    ACGR3        0
## 30    ACGR3        0
## 31    ACGR3        0
## 32    ACGR3        0
## 33    ACGR3        0
## 34    ACGR3        0
## 35    ACGR3        1
## 36    ACGR3        0
## 37    ACGR3        0
## 38    ACGR3        0
## 39    ACGR3        0
## 40    ACGR3        0
## 41    ACGR3        0
## 42    ACGR3        0
## 43    ACGR3        0
## 44    ACGR3        0
## 45    ACGR3        0
## 46    ACGR3        0
## 47    ACGR3        0
## 48    ACGR3        0
## 49    ACGR3        1
## 50    ACGR3        0
## 51    ACGR3        1
## 52    ACGR3        1
## 53    ACGR3        0
## 54    ACGR3        0
## 55    ACGR3        0
## 56    ACGR3        0
## 57    ACGR3        0
## 58    ACGR3        0
## 59    ACGR3        0
## 60    ACGR3        0
## 61    ACGR3        0
## 62    ACGR3        0
## 63    ACGR3        0
## 64    ACGR3        0
## 65    ACGR3        1
## 66    ACGR3        0
## 67    ACGR3        0
## 68    ACGR3        0
## 69    ACGR3        0
## 70    ACGR3        0
## 71    ACGR3        0
## 72    ACGR3        0
## 73    ACGR3        0
## 74    ACGR3        0
## 75    ACGR3        0
## 76    ACGR3        0
## 77    ACGR3        0
## 78    ACGR3        0
## 79    ACGR3        0
## 80    ACGR3        0
## 81    ACGR3        0
## 82    ACGR3        0
## 83    ACGR3        0
## 84    ACGR3        0
## 85    ACGR3        0
## 86    ACGR3        0
## 87    ACGR3        0
## 88    ACGR3        0
## 89    ACGR3        0
## 90    ACGR3        0
## 91    ACGR3        0
## 92    ACGR3        0
## 93    ACGR3        0
## 94    ACGR3        0
## 95    ACGR3        0
## 96    ACGR3        0
## 97    ACGR3        0
## 98    ACGR3        0
## 99    ACGR3        0
## 100   ACGR3        1
```

```
set.seed(1234)
compiled_random_samples <- replicate(n = num_samples, expr = sample_n(filtered_dataset,
    size = sample_size, replace = FALSE), simplify = FALSE) %>%
    bind_rows(.id = "sample_id")

has_bigtooth_maples <- compiled_random_samples %>%
    group_by(sample_id) %>%
    summarize(prevalence = sum(ifelse(OBSERVED == 1, 1, 0))/n())
# add the total number of observations and divide by total
# number of rows

ggplot(has_bigtooth_maples, aes(x = prevalence)) + geom_histogram(binwidth = 0.01) +
    geom_vline(aes(xintercept = 0.06), color = "red", linetype = "dashed") +
    geom_vline(aes(xintercept = mean(prevalence)), color = "blue",
        linetype = "solid")
```



```
mean(has_bigtooth_maples$prevalence)
```

```
## [1] 0.05706
```

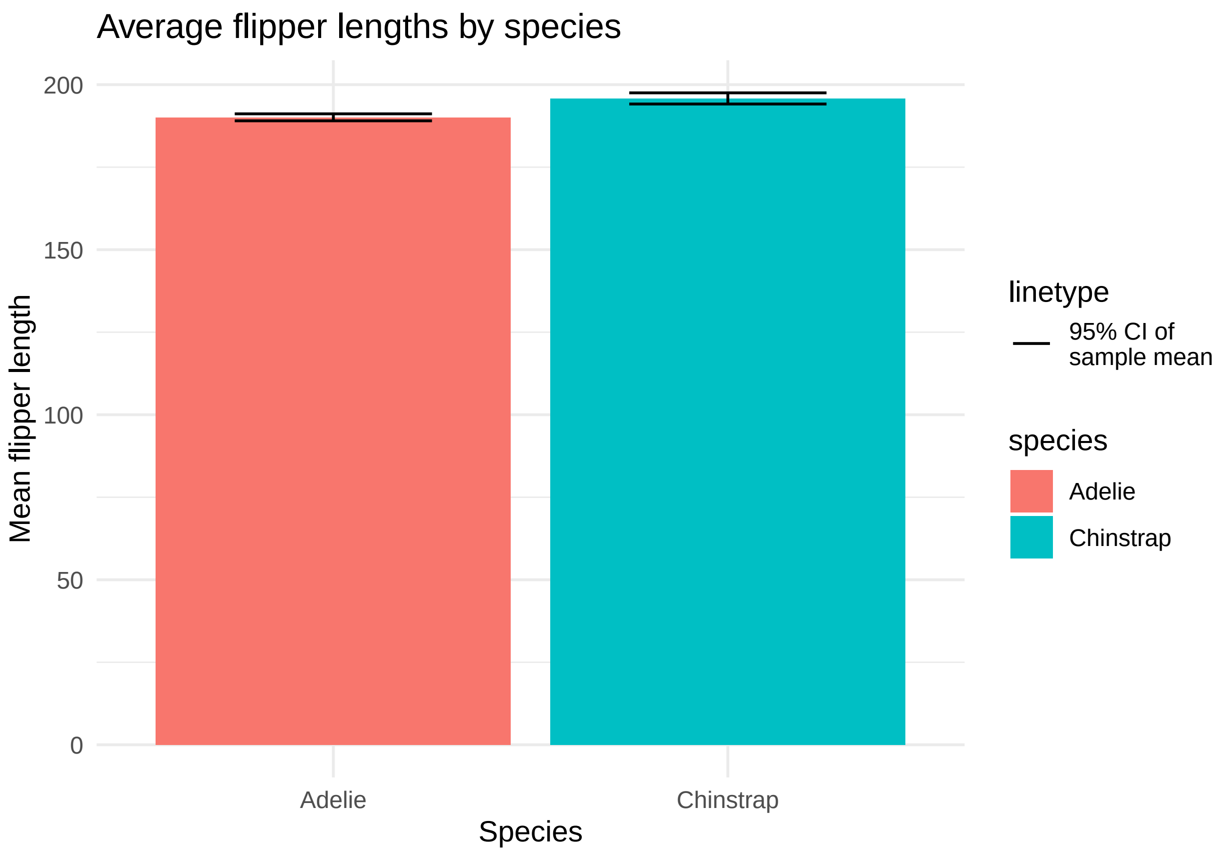
Answer: I got fairly close to a normal distribution at around 3000 repeated samples. The mean prevalence is very close to the population mean at this number of repeated samples, indicating that we can closely approximate the population mean when enough repeated sampling of a given dataset produces a near-normal distribution.

Exercise 2

Using the penguin dataset from package palmerpenguins that you have worked with before (use the cleaned dataframe called "penguins"), assess the difference in flipper length between adelie and chinstrap penguins.

Calculate mean, standard error and 95% confidence interval of flipper length for both species. Plot the result as a bar chart with error bars, where error bars stand for the 95% confidence interval. Label your graph as appropriate. Which species has a larger standard error and why do you think that is? Interpret the result. Be sure to use confidence interval in your interpretation.

```r
filtered_penguins <- na.omit(penguins) %>%
    select(flipper_length_mm, species) %>%
    filter(species == "Adelie" | species == "Chinstrap")

summarized <- filtered_penguins %>%
    group_by(species) %>%
    summarize(mean_flipper_length = mean(flipper_length_mm),
        standard_error = sd(flipper_length_mm)/sqrt(n()), lower_ci = mean_flipper_length -
            1.96 * standard_error, upper_ci = mean_flipper_length +
            1.96 * standard_error)

ggplot(summarized, aes(y = mean_flipper_length, x = species,
    fill = species)) + geom_col() + geom_errorbar(aes(ymin = lower_ci,
    ymax = upper_ci, linetype = "95% CI of\nsample mean", width = 0.5)) +
    labs(title = "Average flipper lengths by species", y = "Mean flipper length",
        x = "Species") + theme_minimal()
```



Answer: The Chinstrap species has a larger standard error, likely meaning that the sample is more spread out around the population mean (the sample is not as close to the actual population mean as the sample for the Adelie species). The Adelie species also has a narrower range for the confidence interval, meaning that the sample for Adelie penguins is more stable/less variable than the sample for Chinstrap penguins (the estimate would vary more if we took additional samples for the Chinstrap species).

Exercise 3

Now simulate a sampling distribution of means for flipper length from the data, for the same two species as above, using random sampling with replacement. Collect data from 10 penguins per species in each random sample and collect a total of 100 random samples (so each of the 100 samples includes 10 of each of the two species and their flipper length measures).

Calculate the mean flipper length for each sample so you will end up with 100 means for each species. Calculate the standard deviation of the mean for this sampling distribution. How does it compare to the standard error you calculated in Exercise 2 above? Explain what is going on.

```
filtered_adelie <- na.omit(penguins) %>%
    select(flipper_length_mm, species) %>%
    filter(species == "Adelie")

adelie_samples <- replicate(n = 100, expr = sample_n(filtered_adelie,
    size = 10, replace = TRUE), simplify = FALSE) %>%
    bind_rows(.id = "sample_id")

adelie_means <- adelie_samples %>%
```

```r
    group_by(sample_id) %>%
    summarize(mean = mean(flipper_length_mm))

sd(adelie_means$mean)
```

```
## [1] 1.971509
```

```r
filtered_chinstrap <- na.omit(penguins) %>%
    select(flipper_length_mm, species) %>%
    filter(species == "Chinstrap")

chinstrap_samples <- replicate(n = 100, expr = sample_n(filtered_chinstrap,
    size = 10, replace = TRUE), simplify = FALSE) %>%
    bind_rows(.id = "sample_id")

chinstrap_means <- chinstrap_samples %>%
    group_by(sample_id) %>%
    summarize(mean = mean(flipper_length_mm))

sd(chinstrap_means$mean)
```

```
## [1] 2.248173
```

Answer: The SDs for this distribution seem to align with the SEs from the previous exercise: they seem to indicate that the Chinstrap samples have more variability / less precision than the samples for the Adelie species, due to a larger SD value for the Chinstrap results.