Lab and Homework Exercises Week 05

Alison Lawyer

2024-10-15

```r
# 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)
```

```r
# List required packages
library(ggplot2)
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 lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- 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 errors
```

Homework Exercises

Exercise 1

In a study of bird populations, you observe that 30% of the birds have a specific plumage coloration trait. You sample 50 birds. Calculate the probability of observing between 12 and 18 birds with this trait.

```r
sum(dbinom(12:18, size = 50, prob = 0.3))
```

```
## [1] 0.7204041
```

```r
# or
pbinom(18, size = 50, prob = 0.3) - pbinom(11, size = 50, prob = 0.3)
```

```
## [1] 0.7204041
```

Exercise 2

In a study of plant growth, you have data on the heights of two different species of plants. Species A has a mean height of 40 cm with a standard deviation of 5 cm, while Species B has a mean height of 35 cm with a standard deviation of 7 cm. Calculate the probability that a randomly selected plant from each species is between 38 cm and 42 cm tall.

```r
mean_A <- 40
sd_A <- 5
mean_B <- 35
sd_B <- 7

pnorm(42, mean = mean_A, sd = sd_A) - pnorm(38, mean = mean_A,
    sd = sd_A)
```

```
## [1] 0.3108435
```

```r
pnorm(42, mean = mean_B, sd = sd_B) - pnorm(38, mean = mean_B,
    sd = sd_B)
```

```
## [1] 0.1754623
```

Exercise 3

Scenario: You are studying the lengths of fish in two different lakes, Lake A and Lake B. You collected data on the lengths of 100 fish from each lake. In Lake A, the mean length is 25 cm with a standard deviation of 3 cm. In Lake B, the mean length is 24.5 cm with a standard deviation of 2.5 cm. Your goal is to explore the difference in mean length of fish between the two population while taking into account their variability.

Step 1: Generate random samples from both populations to use for analysis. Use the function rnorm() to sample from two normal distributions with the parameters for each population given above.

Step 2: Calculate the z-scores for each observation in the sample for both Lake A and Lake B. Save the z-scores, along with the original data, in a dataframe for later use.

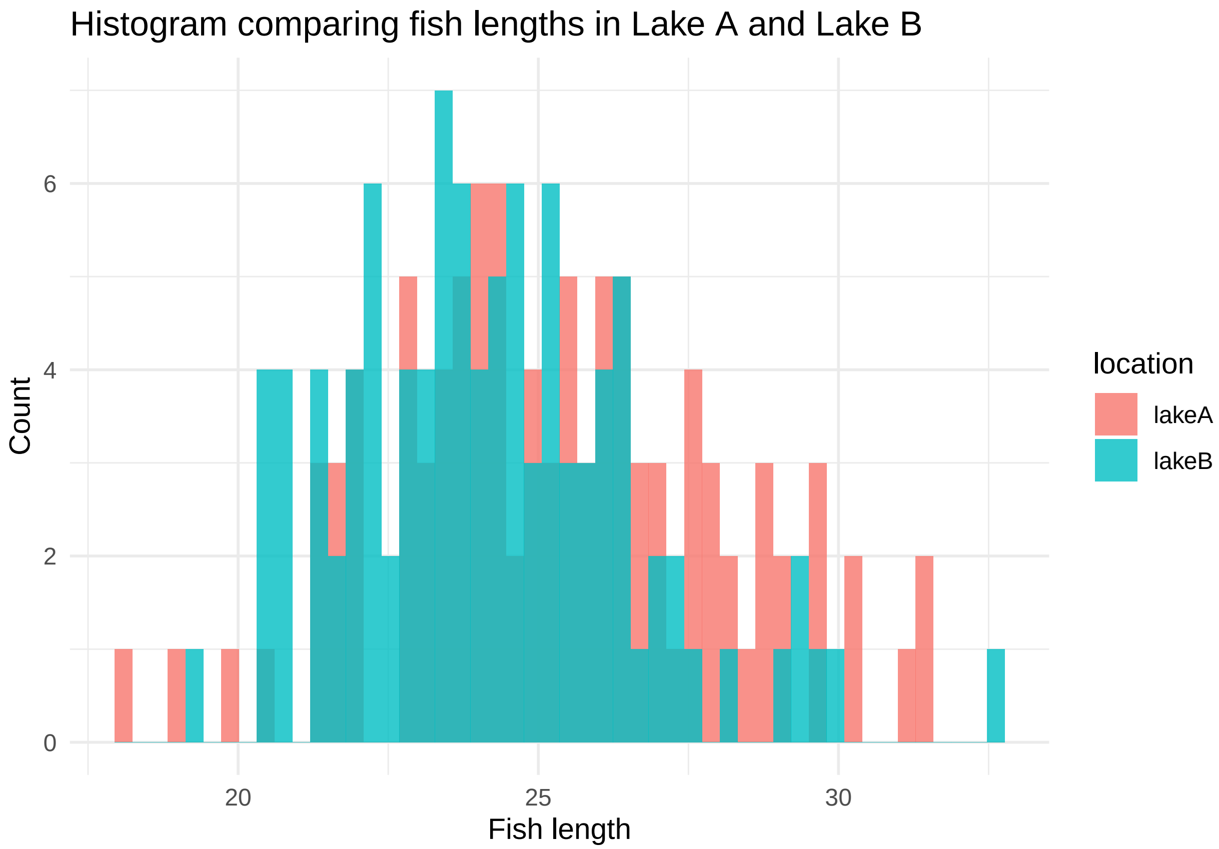
Step 3: Compare the distributions of lengths (original measurements) between the two lakes by plotting histograms. You can overlay both histograms in the same graph or create two different graphs side-by-side. Calculate the mean difference in lengths between Lake A and Lake B.

Step 4: Compare the distributions of z-scores between the two lakes by plotting histograms. You can overlay both histograms in the same graph or create two different graphs side-by-side. Calculate the difference in mean z-scores between Lake A and Lake B.

Step 5: Interpret your findings. How does the difference in mean length compare to the difference in z-scores? What do you think will happen with the z-scores if your sample size goes up?

```r
fishA_mean <- 25
fishA_sd <- 3
fishB_mean <- 24.5
fishB_sd <- 2.5
set.seed(123)

## Step 1:
lakeA_sample <- rnorm(100, mean = fishA_mean, sd = fishA_sd)
```

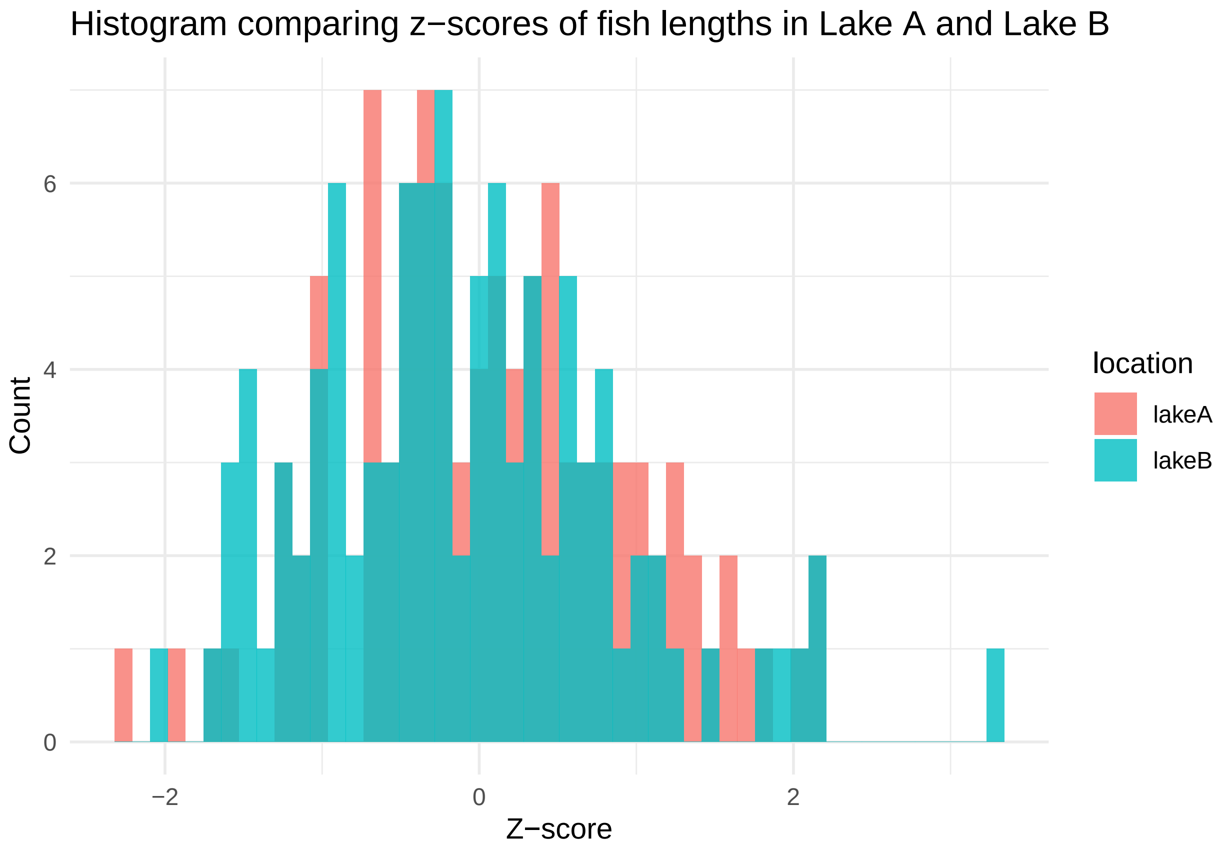
```
lakeB_sample <- rnorm(100, mean = fishB_mean, sd = fishB_sd)

## Step 2:
zscoreA <- (lakeA_sample - fishA_mean)/fishA_sd
zscoreB <- (lakeB_sample - fishB_mean)/fishB_sd

fish_variability <- tibble(sample = c(lakeA_sample, lakeB_sample),
    location = c(rep("lakeA", length(sample)/2), rep("lakeB",
        length(sample)/2)), zscore = c(zscoreA, zscoreB))

## Step 3:
ggplot(fish_variability, aes(sample, fill = location)) + geom_histogram(bins = 50,
    alpha = 0.8, position = "identity") + labs(title = "Histogram comparing fish lengths in Lake A and
    x = "Fish length", y = "Count") + theme_minimal()
```



```
sampleA_mean <- mean(lakeA_sample)
sampleB_mean <- mean(lakeB_sample)
mean_length_differences <- sampleA_mean - sampleB_mean
mean_length_differences
```

```
## [1] 1.040085
```

```
## Step 4:
ggplot(fish_variability, aes(zscore, fill = location)) + geom_histogram(bins = 50,
    alpha = 0.8, position = "identity") + labs(title = "Histogram comparing z-scores of fish lengths in
    x = "Z-score", y = "Count") + theme_minimal()
```



```
mean_zscoreA <- mean(zscoreA)
mean_zscoreB <- mean(zscoreB)
mean_zscore_differences <- mean_zscoreA - mean_zscoreB
mean_zscore_differences
```

```
## [1] 0.1979527
```

Answer: The difference between the mean lengths is larger than the difference between the mean z-scores. This means that the average z-scores of the data points within the data sets are more similar to each other than the average lengths of the fish between the sample locations. As sample sizes increase, the z-scores should become more similar – since z-scores measure how far a data point is away from the mean in terms of standard deviations, a larger sample size should produce a more normal distribution for each location, resulting in less variability in the z-scores.