M5 Problem Set Distributions, Probability and z-score for TAs

Pablo E. Gutiérrez-Fonseca

2024-03-07 17:23:22

R practice.

```
library(BSDA)
library(tidyverse)
library(lessR)
library(readxl)
```

Welcome to our first BIG problem set where we begin to explore the concepts and methods of normality, probability and z scores. You will need to download both the instruction document (which you are looking at right now) and the Sleeping Mammals.xls spreadsheet. Take your time and know that this should probably take 2-3 hrs to complete. Once again, I encourage you to work in groups to discuss these questions. However, you must ultimately do the work on your own and turn in your answers and interpretations independently. HAVE FUN!

Consider the data on sleep patterns in various mammalian species in the spreadsheet **Sleeping Mammals.csv**. This dataset contains information on sleep patterns for comparison to factors such as species body weight, brain size, life span, gestation period, and three class measures of risk: predation, danger, and exposure. In this problem set, we will be practicing our transformations and probabilities associated with standard normal distributions.

We are interested in comparing sleep time spent in dreaming and non-dreaming phases for mammals of different sizes. **We hypothesis that large mammals exhibit more time spent in dreaming sleep than small mammals**. Before we can test this hypothesis, we need to examine the distribution of our data.

- Import your data table into R.
- Analyze the distribution of both the Dreaming and Non-Dreaming variables. We'll start by just examining some of the descriptive statistics and what they tell us about this data.

1. Based on the data, mammals spend most of their time in which type of sleep phase (dreaming or non-dreaming)?
2. Which type of sleep phase has the highest variability across the species included here (dreaming or non-dreaming)?

Now we will be testing distributions and various transformations. We will examine one variable at a time. We will start with the Dreaming variable:

- Fit a normal distribution and test for Normality using the Goodness of Fit test.

3. Enter the p-value for the goodness of fit test for the **Dreaming variable**.
(NOTE. . . .just enter the number (no letters, symbols, equal signs etc). . . also be careful of decimal places.

4. Based on this goodness of fit test, is the dreaming variable normally distributed? You can check out this page if you are wondering.

Now with the non-Dreaming variable:

- Fit a normal distribution and test for normality using the shapiro.test().

5. What is the p-value for the goodness of fit test for the non-Dreaming variable (NOTE. . . .just enter the number (no letters, symbols, equal signs etc). . . also be careful of decimal places.

6. Based on this goodness of fit test, is the non-Dreaming variable normally distributed? You can check out this page if you are wondering.

7. Plot the Dreaming variable with a density plot – does it look normal?

8. Plot the non-Dreaming variable with a density plot – does it look normal?

z-score

Since we are interested in comparing sleep phase for large vs. small mammal classes, we need to make sure that each of our species has been assigned to either the small or large class. Note that there are several species for which this data is blank. We will assign these last few species to a class based on how they fit the distribution of the Non-dreaming variable when divided into the two size classes (small and large).

We will calculate the z-score associated with the measured ***non-dreaming value*** *for each of the species we are missing a size class for in our data table to see which mean (0 value) that species z-score is closest to.* This will help us determine which group (large or small) each of these "missing" species best fits when considering their non-dreaming patterns.

- First you will need to determine the non-dreaming mean and standard deviation for both the small and large sub-groups.

9-12. Go back to your notes and using the formula for calculating z scores, **calculate the z-score for each species based on its non-dreaming measured value for the small class**. It might be easiest to just do all of the calculations out (on your paper or in excel) before beginning to enter your results in blackboard..

13-16. Based on your knowledge that higher (absolute value) z-scores are more unusual for a given group's distribution, which class would you assign each species to (small or large)? (hint compare each species z scores for the small and large group. . . . . which group does the species z score fit "best" with. . . ..i.e. for which group is that species z score less extreme?

For 17 and 18. Identify the **upper and lower z-scores** that cumulatively represent $< 2\%$ TOTAL chance of occurring. This doesn't require any calculations, just the determination of two z-scores (upper and lower) from the normal probability table (or excel function).

19. What is the actual **non-dreaming** value associated with the upper z score identified in the question above.

20. How many **non-dreaming** observations in your data set would be labeled as outliers by this zscore definition?

Nice job. . . you've made it through!. . . . . Now take a break!