Problem set_ M5 Probability and Hypothesis Testing

Pablo E. Gutiérrez-Fonseca

2023-08-12

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 (Analyze > Distribution). We'll start by just examining some of the descriptive statistics and what they tell us about this data.

1. Based on the summary results in JMP, 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?

Let's see if any outliers are skewing our distribution.

5. Enter the new p-value for the goodness of fit test for the **Dreaming variable** when the outliers are excluded (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 dreaming variable normally distributed once outliers are removed?

Now we will try a log transformations on our data.

7. Enter the new p-values for the goodness of fit test for the **log transformed Dreaming variable**.
8. Is this **log transformed Dreaming variable** normally distributed?
9. Based on these results, would you transform (outlier transform, log transform or no transform) your **dreaming variable**? Assume that outliers removed are incorrect data values and that all values fit the range necessary for a log transform)
10. Now go back and repeat these procedures for the **Non-dreaming variable**. However, instead of entering answers for every one of these steps, just summarize what you would do with this variable: Based on these results, would you transform (outlier transform, log transform or no transform) your non-dreaming variable? Assume that outliers removed are incorrect data values)

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 zscore 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.

11-14. 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** (I've done this for you for the large 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.

15-18. 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?

Now I'm interested in identifying potential outliers in our **non-dreaming data**. Let's start with the standard IQR approach (outliers are any observations that fall more than 1.5 * the interquartile range from the first and third quartiles).

19. How many observations in your data set would be labeled outliers by this IQR definition?
    Some people prefer other methods of identifying outliers. One other option is using probabilities (i.e. outliers are those observations with less than a 5% chance of occurring) and another option is to use standard deviations (i.e. observations that are more than 2 standard deviations from the mean are potential outliers).

20. Armed with your knowledge of what a z score is and how to calculate actual values from given z scores, enter the **upper non-dreaming** value beyond which I may have outliers (values greater than 2 standard deviations from the mean in the positive direction). It may help if you sketch this out so that you can visualize what you are calculating.

21. Do the same for the **lower non-dreaming value** below which I might consider observations to be outliers based on the fact that they are more than 2 standard deviations under the mean.

22. Using this 2-standard deviation threshold, how many potential outliers might be in this nondreaming data set?

Now I'd like to see what outliers might be flagged if instead of the IQR method the JMP uses, or the 2 standard deviation approach that we just calculated, we use a probability threshold.

30 and 31. 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).

32. What is the actual **non-dreaming** value associated with the upper z score identified in the question above.
33. 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!