Exploration 6.2: Do house cats have socio-spatial cognition?

Comparing Two Means: Simulation-Based Approach

LEARNING GOALS

- State the null and the alternative hypotheses in terms of "no association" versus "there is an association" as well as in terms of comparing means (i.e., μ_1 and μ_2) for an explanatory variable with two categories.
- Implement the 3S strategy to compare two means: Find a statistic, simulate, and compute the strength of evidence against observed study results happening by chance alone.
- Describe how to use cards to simulate what outcomes (in terms of difference in means or median) are to be expected in repeated random assignments if there is no association between the two variables.
- Use the **Multiple Means** applet to conduct a simulation of the null hypothesis and be able to read output from the **Multiple Means** applet.
- Find and interpret the standardized statistic and the p-value for a test of two means
- Use the 2SD method to find a 95% confidence interval for the difference in population means for two "treatment" groups and interpret the interval in the context of the study; interpret what it means for the 95% confidence interval for difference in means to contain zero.
- State a complete conclusion about the alternative hypothesis (and null hypothesis) based on the p-value and/or standardized statistic and the study design, including statistical significance, estimation, generalizability, and causation.

STEP 1: Ask a research question. For many animals in the wild, it is important to be able to keep track other animals even without actually seeing them. Some animals do this using their other senses like sound or smell. Keeping track of another animal by holding a mental representation of their location is a form of *socio-spatial cognition*. Researchers in Japan (Takagi et al., 2021) wanted to determine whether common house cats have socio-spatial cognition.

STEP 2: Design a study and collect data. The researchers gave the cats a "teleportation-like" situation where the cat would hear their owner's voice outside a room and then suddenly hear it again on the opposite side of the room. If the cats had socio-spatial cognition, the researchers thought the cats would show surprise by what seemed like a sudden change in the owner's position. The researchers tested 40 cats in familiar surroundings, either at their home or in a cat café where they lived. They placed the first speaker on the opposite side of a door of the room where the cat was located. A second speaker was placed on the opposite side of the room near a door or window that was at least four meters away from the first speaker. The cats would hear their owner's voice from the first speaker. Then 2.5 seconds later, they would hear either hear their owner's voice or another person's voice from the second speaker. After hearing the voice from the second speaker, eight people rated the level of surprise the cat displayed (shifting ears, looking back and forth, moving) on a scale of 0 (no surprise) to 4 (strongly surprised). The raters' scores were averaged to give each cat a surprise score. If common house cats have socio-spatial cognition, the researchers thought cats hearing their owner's voice coming from the second speaker would show a higher level of surprise than those that heard a different voice. The 40 cats were randomly assigned to the two conditions.

1. What are the observational (or experimental) units?

2. Identify the explanatory and response variables in this study. Also classify them as either categorical or quantitative.

3. Was this an experiment or an observational study? Explain how you are deciding.

4. Let μ_{same} be the long-run mean surprise score for cats that hear the same voice (the owner's) on the second speaker and let $\mu_{different}$ denote the long-run mean surprise score for cats that hear a different voice on the second speaker. In words and symbols, state the null and the alternative hypotheses to investigate whether hearing the same voice has an effect on surprise level of the cats. (Note: This is a two-sided test as it is certainly possible that cats could show a higher level of surprise by hearing a different voice for some reason.)

STEP 3: Explore the data.

5. Enter the **Surprise** data into the **Multiple Means** applet. Select Surprise as the response variable and Voice as the explanatory variable.

   a. Notice that the applet creates parallel dotplots, one for each study group. You can also add boxplots to these graphs. Based on these graphs alone, which group (same or different) appears to have had the higher mean surprise score? How are you deciding?

   b. Notice also that the applet also computes numerical summaries of the data, such as the mean and standard deviation (SD) for the surprise scores in each group.

      i. For the *same* group, record the sample size (n), mean, and SD.

      n_{same} = $\bar{x}_{same}$ = SD_{same} =

      ii. For the *different* group, record the sample size (n), mean, and SD.

      $n_{different}$ = $\bar{x}_{different}$ = $SD_{different}$ =

Recall from earlier in the course that the standard deviation is a measure of variability. Relatively speaking, smaller standard deviation values indicate less variability and a distribution whose data values tend to cluster more closely together, compared to a distribution with a larger standard deviation.

   c. Based on the numerical summaries, which group (same or different) had the higher mean surprise score? Was your predict in (a) correct? Which group had the higher variability in surprise scores?

   d. Notice that the applet also reports the observed difference in the mean surprise scores between the two groups. Record this value. Before you conduct an inferential analysis, does this difference in sample means strike you as a meaningful difference? Explain your answer.

$\bar{x}_{same} - \bar{x}_{different}$ =

STEP 4: Draw inferences beyond the data.

6. What are two possible explanations for why we observed the two groups to have different sample mean surprise scores?

7. Describe how you might go about deciding whether the observed difference between the two sample means is statistically significant. (*Hint*: Think about how you assessed whether an observed difference between two sample proportions was statistically significant in comparing two proportions. Use the same strategy, with an appropriate modification for working with means instead of proportions.)

Once again, the key question is how often random assignment alone would produce a difference in the groups at least as extreme as the difference observed in this study if there really was no genuine effect of the same voice on surprise score. Similar questions were addressed when comparing two proportions. The only change is that now the response variable is *quantitative* rather than *categorical*, so the relevant statistic is the difference in group *means* rather than the difference in group *proportions*. Also once again, we can use *simulation* to investigate how often such an extreme difference would occur by chance (random assignment) alone (if the null hypothesis of no difference/no effect/no association were true). In other words, we will again employ the 3S strategy.

1. Statistic

8. A natural statistic for measuring how different the observed group means are from each other is the difference in the mean surprise scores between the two groups. Report the value of this statistic, as you did in #5(d).

2. Simulate*:* You can start by using index cards to perform a tactile simulation of randomly assigning the 40 cats (and their surprise scores) between the two groups, *assuming* that voice condition has no impact on surprise score. Because the null hypothesis asserts that surprise score is not associated with voice condition, we will assume that the 40 cats would have had exactly the same surprise scores as they did, *regardless* of which voice condition group (same or different) the cat had been assigned.

9. To conduct this simulation:

    a. How many cards do you need?

    b. What will you write on each card?

    c. To conduct *one repetition* of this simulation shuffle the stack of 40 cards well and then randomly distribute cards into two stacks: one stack with 21 cards (the same-voice group) and one with 19 (the different-voice group).

        i. Calculate and report the sample means for each rerandomized group.

        ii. Calculate the difference in group means: Same-voice mean minus different-voice mean. Report this value.

    d. Combine this result with your classmates' to create a dotplot that shows the distribution of several possible values of the difference in sample means that could have happened due to

pure chance if the voice condition has no impact on surprise score. Sketch the dotplot, being sure to label and scale the horizontal axis.

e. At about what value is the dotplot centered? Explain why this makes sense. (*Hint*: What are we assuming to be true when we conduct the simulation?)

f. Where is the observed difference in means from the original study (as reported in #8) on the dotplot? Did this value happen often, somewhat rarely, or very rarely? How are you deciding?

10. As before with simulation-based analyses, you would now like to conduct many, many more repetitions to determine what is typical and what is not typical for the difference in group means, assuming that voice condition has no impact on surprise score. We think you would prefer to use a computer applet to do this rather than continue to shuffle cards for a very long time, calculating the difference of group means by hand. Go back to the **Multiple Means** applet, check the **Show Shuffle Options** box, select the **Plot** display, and press **Shuffle Responses**.

    a. Describe what the applet is doing and how this relates to your null hypothesis from #4.

    b. Record the shuffled difference in sample means for the rerandomized groups, as given in the applet output. Is this difference more extreme than the observed difference from the study (as reported in #8)? How are you deciding?

    c. Click on **Shuffle Responses** again and record the simulated difference in sample means for the rerandomized groups. Did it change from #10b?

    d. Click on **Re-Randomize** again and record the simulated difference in sample means for the rerandomized groups. Did it change from #10b and #10c?

    Now to see many more possible values of the difference in sample means, assuming voice condition has no impact on surprise score, do the following in the **Multiple Means** applet:

    - Change **Number of Shuffles** from 1 to 997.
    - Press **Shuffle Responses** to produce a total of 1,000 shuffles and rerandomized statistics.

    e. Consider the histogram of the 1,000 could-have-been values of difference in sample means, assuming that voice condition has no effect on surprise scores.

        i. What does one dot on the dotplot represent? (*Hint*: Think about what you would have to do to put another dot on the graph.)

        ii. Describe the overall shape of the null distribution displayed in this dotplot.

        iii. Where does the observed difference in sample means (as reported in #8) fall in this dotplot: near the middle or out in a tail? Are there a lot of dots that are even more extreme than the observed difference, assuming voice condition has no impact on surprise? How are you deciding?

f. To estimate a p-value, continue with the **Multiple Means** applet. Type in the observed difference in group means (as reported in #8), choose the appropriate alternative hypothesis in the **Count Samples** box, and press **Count**. What is your approximate p-value?

g. Complete the following sentence to provide the interpretation of the p-value.

The p-value of _____ is the probability of observing _____assuming _____.

3. Strength of evidence

11. Based on the p-value, evaluate the strength of evidence provided by the data against the null hypothesis that the voice condition has no effect on surprise score: not much evidence, moderate evidence, strong evidence, or very strong evidence?

12. Use the 2SD method to approximate a 95% confidence interval for the difference in long-run mean surprise score for cats who hear the same voice minus the long-run mean surprise score for cats who hear the different voice. (*Hints*: Remember the observed value of the difference in group means and obtain an estimate of the SD of the difference in group means from the applet's simulation results. The interval should be *observed difference in means* ± *2SD*.)

STEP 5: Formulate conclusions.

13. ***Significance:*** Summarize your conclusion with regard to strength of evidence in the context of this study.

14. ***Estimation:*** Fill in the following interpretation of what this confidence interval reveals, paying particular attention to whether the interval is entirely positive, entirely negative, or contains zero. (*Hint*: Include the appropriate numbers and then choose the appropriate "direction" (higher or lower) in your interpretation.)

I'm 95% confident that the long-run mean surprise score with the _____ treatment is _____ (higher/lower) to _____(higher/lower) than the long-run mean surprise score with the _____ treatment.

15. ***Generalization:*** Were the cats in this study randomly selected from a larger population? Describe the population to which you would feel comfortable generalizing the results of this study.

16. ***Causation:*** Were the participants in the study randomly assigned to a voice condition? How does this affect the scope of conclusion that you can draw?

STEP 6: Look back and ahead.

17. ***Looking back:*** Did anything about the design and conclusions of this study concern you? In particular, are there things that could have been done to give a better chance finding strong evidence of a true difference between the two groups? Issues you may want to critique include:

- Any mismatch between the research question and the study design
- How the experimental units were selected

- How the treatments were assigned to the experimental units
- How the measurements were recorded
- The number of experimental units in the study
- Whether what we observed is of practical value

18. ***Looking ahead:*** What should the researchers' next steps be to fix the limitations or build on this knowledge?

Exploring Further: Another statistic

We could have chosen a statistic other than the difference in group means to summarize how different the two groups' surprise scores were. For example we could have used the difference in group *medians*. Why might we do this? For one reason, the median is less affected by outliers than the mean. To analyze the difference in group medians, we carry out the 3S strategy as before, except we use medians instead of means.

19. Return to the **Multiple Means** applet and use the **Statistic** pull-down menu to select **Difference in medians**.

    a. From the **Summary Statistics** for the original data, record the median surprise score for each group. Also record the difference between the medians (same-voice median minus diff-voice median).

    b. Enter 1000 for **Number of Shuffles** and press **Shuffle Responses**. Describe the resulting null distribution of difference in group medians. Does this null distribution appear to be centered near zero? Does it seem to have a bell-shaped distribution?

    c. To calculate a p-value based on the difference in medians, enter the observed value in the **Count samples** box. Then press **Count**. Report both the value that you enter into the applet and the resulting p-value.

    d. Does this p-value indicate strong evidence of a voice effect on surprise score? Explain how you are deciding.

    e. In this study, for which statistic (difference in *means* or difference in *medians*) do the data provide *stronger* evidence that there is a voice effect on surprise score? Explain how you are deciding.

    f. In the next section, we will use a theory-based test to determine whether there is strong evidence of a true difference between the means of two groups. Based on your null distribution for the differences in medians you saw in part (b), explain why using medians instead of means might be problematic in a theory-based test.

Reference

Takagi S, Chijiiwa H, Arahori M, Saito A, Fujita K, Kuroshima H (2021) Socio-spatial cognition in cats: Mentally mapping owner's location from voice. PLoS ONE 16(11): e0257611. https://doi.org/10.1371/journal.pone.0257611