STAT 113: TESTING THE FIT OF CATEGORICAL DATA TO A KNOWN SET OF PROPORTIONS

Mannan and Meslow (1984) studied bird foraging behavior in a forest in Oregon. In a managed forest, 54% of the canopy volume was Douglas fir, 40% was ponderosa pine, 5% was grand fir, and 1% was western larch. They made 156 observations of foraging by red-breasted nuthatches. The null hypothesis is that the birds choose trees to forage in at random, without regard to species.

1. If the null hypothesis is true, what proportion of the time would you expect the nuthatches to forage in each type of tree in the long run?

2. Each observation in this data is an instance of foraging. What "population" can we think of these observations as being drawn from?

3. Translate the null hypothesis to a statement about a set of proportions that pertain to this population.

Date: November 25, 2019.

4. Propose a method for creating *one* randomization sample of size $n = 156$. Remember that randomization samples are simulated based on a world in which the null hypothesis is true. There may be many sensible choices!

5. What do you expect the "average" randomization sample to look like?

6. Of the 156 foraging observations made, 79 (51% of the total) occurred in Douglas firs, 70 (45%) in ponderosa pines, 3 (2%) in grand firs, and 4 (3%) in western larches. Are these proportions different from what you expect if the null hypothesis is true? If they are, give two possible reasons why that might be.

7. Can you conclude that the null hypothesis is false? Why or why not?

8. In order to do a hypothesis test, **we need a test statistic which is a single number**. When our null parameter is a single number, and our sample statistic is a single number, our test statistic has been based on the difference between the two.

   Now, though, our null hypothesis is a statement about *multiple* proportions, and our sample statistic is a collection of *multiple* proportions. **Propose at least one way you could calculate the "distance" between the observed set of proportions and the null set of proportions.** The key here is that your measures of distance need to provide a *single number* that takes into account how "far away" the five counts or proportions in the sample are from the average counts or proportions across all randomization samples. There is no single right answer!

9. For your measure(s), do larger values look more favorable for the alternative hypothesis? Smaller values? Both? (That is, if we created a randomization distribution for your measure, which values would "count" toward the P-value?)

10. One way we could generate one randomization sample is the following:
    (a) Take 100 pieces of paper, write "Douglas fir" on 54 (that is, 54%) of them, "Ponderosa pine" on 40 (40%), "Grand fir" on 5 (5%) and "Western larch" on 1 (1%). These proportions are chosen based on what the null hypothesis predicts the "long run" foraging proportions should be.
    (b) Put the slips of paper in a hat.
    (c) Pick one slip out at random. Write down what it says, and put it back. Repeat this until we have 156 observations.
    (d) Count up how often each type of tree appears in the randomization sample.
    (e) Calculate the distance between the proportions of each tree species in this sample and the proportions according to the null hypothesis. (This difference is our test statistic).

    If this procedure were repeated thousands of times to obtain a randomization distribution, how would you obtain a P-value?

A statistic with good mathematical properties is *Pearson's Chi-squared (χ^2) statistic* (chi, or χ is a Greek letter; annoyingly, this goes against the convention of using Greek letters for *parameters*... — appropriately though, the letter's name rhymes with the word "sigh", but with a 'k' sound in front)

It is used in a situation like this, with a non-binary categorical variable, to measure the "distance" between category counts in a sample and expected counts according to some H_0. Here's how it's calculated:

(i) Based on the sample size and the null hypothesis set of proportions, find the "expected" number of observations we should see in each category in the long run if H_0 were true. This is just the null proportion for each category times the sample size.

(ii) Calculate the differences between each observed count and the corresponding expected count (if H_0 were true)

(iii) Divide this difference by the square root of the expected count (this functions like a sort of standard error).

(iv) We now have one standardized "distance" for each category. To combine these distances into a single number, we use the same principle we used to compute the discrepancy between a dataset and a regression model: sum the squared distances.

All together, this yields the following formula for χ^2

$$\chi^2_{observed} = \sum_c \frac{(\mathsf{Observed\ count_c} - \mathsf{Expected\ count_c})^2}{\mathsf{Expected\ count_c}}$$

where we have one term in the sum for each category (c indexes categories.) Note that the square roots disappeared because we squared each term in the sum.

The reason we divide each difference in counts by the square root of the expected count is that the bigger the expected count, the bigger the differences we will tend to get by chance: so we want to "downplay" differences when the expected count is large. For example, if we expect a count of 10 and get 15 this is a bigger deal than if we expect 1,000,000 and get 1,000,005, even though both differ by 5 cases.

11. **For the foraging data, we had observed counts of 79, 70, 3, and 4**, for Douglas fir, ponderosa pine, grand fir, and western larches. **The null "long run" proportions are 0.54, 0.40, 0.05, and 0.05. Find the "expected" (average) counts according to H_0. Then use these, together with the expected counts, to find the χ^2 test statistic for this data.**

12. StatKey has an applet for constructing a randomization distribution as described above.

    (i) Go to StatKey and select "χ^2 Goodness-of-fit" under "More Advanced Randomization Tests".

    (ii) Click, "Edit Data", and enter the observed counts as below:

    ```
    Tree, Count
    Douglas Fir, 79
    Ponderosa Pine, 70
    Grand Fir, 3
    Western Larch, 4
    ```

    (iii) Click the button marked "Null hypothesis" and enter the H_0 proportions. ***Note: you must edit the data first then enter the null proportions afterwards; if you need to edit the data, you'll have to re-enter the null proportions. StatKey has a bug that causes the null proportions to be reset whenever new observed counts are entered.***

    (iv) **generate a few thousand randomization samples.** StatKey will compute the χ^2 statistic for each one and build the plot.

    (v) Write down the *observed* value of the χ^2 statistic from the upper right above the observed counts. Click "Show Details" to see the contribution of the individual categories.

13. Describe the shape of the randomization distribution.

14. Consider the formula for the χ^2 statistic. What values do you think constitute big discrepancies from the null? Large values? Small values? Large or small values on either extreme? (Hint: What will χ^2 be if the observed and expected counts are exactly equal?)

15. Select the tail(s) consisting of the values that are *least* consistent with H_0, set the threshold (the "cutoff") to the observed χ^2 value, and report the P-value. What is your conclusion in the context of foraging behavior, using a significance level of $\alpha = 0.01$?

16. Supposing the proportions were the same but the sample size were doubled. How do you expect the distance measure and the P-value to change?

17. Click "Edit data" and enter new counts that are double the old ones. You will need to enter the null proportions again. What is the new χ^2 value?

18. Repeat the randomization procedure. How did the distribution change (if at all)?

19. Compute the new P-value. What would your new conclusion be in context?

20. Why do you think the P-value got smaller even though the sample proportions and null proportions were the same?