# STAT 113: TESTING THE FIT OF CATEGORICAL DATA TO A KNOWN DISTRIBUTION

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 biological null hypothesis is that the birds forage randomly, without regard to what species of tree they're in.

(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) Translate the null hypothesis to a statement about a set of population proportions.

*Date*: November 28, 2017.

(3) 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!

(4) What do you expect the "average" randomization sample to look like?

(5) 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. Can you conclude that the null is false? Why or why not?

(6) Propose at least one way you could calculate the "distance" between the real sample and the average randomization sample. 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!

(7) For your measure(s), are larger values stronger evidence against the null hypothesis? Smaller values? Both?

(8) 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, "Grand fir" on 5 and "Western larch" on 1. Put the slips of paper in a hat

   (b) Pick one slip out at random. Write down what it says, and put it back. Repeat this until we have 156 observations.

   (c) Count up how often each type of tree appears in the randomization sample. Calculate the distance between this sample and the "average" randomization sample.

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

(9) A statistic with good mathematical properties is *Pearson's Chi-square ($\chi^2$) statistic* (chi, or $\chi$, is a Greek letter). 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:

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

   (b) Square the differences

   (c) "Standardize" the squared differences by dividing by the expected count

   (d) Sum these standardized differences across the categories.

   The formula for $\chi^2$ is

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

   where $c$ indexes categories, and the Expected count for each category is the null proportion for that category multiplied by the total sample size, $n$.

   The rationale for the "standardization" step is that the larger the expected count, the larger 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 more surprising than if we expect 1,000,000 and get 1,000,005, even though the differences are the same.

(10) StatKey has an applet for constructing a randomization distribution as described above. Go to StatKey and select "$\chi^2$ Goodness-of-fit" under "More Advanced Randomization Tests". Click, "Edit Data", uncheck "Raw Data", and enter the observed counts as below:

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

Now 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. StatKey has a bug that causes the null proportions to be reset whenever new observed counts are entered.**

Now generate a few thousand randomization samples. StatKey will compute the $\chi^2$ statistic for each one and build the plot. Write down the observed value of the $\chi^2$ statistic from the upper right above the observed counts. Click "Show Details" to see the contribution of the individual categories.

Describe the shape of the randomization distribution. Consider the formula for the $\chi^2$ 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 $\chi^2$ be if the observed and expected counts are exactly equal?) Select the relevant tail(s), and set the threshold to the observed $\chi^2$ value, and report the $P$-value.

(11) 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?

(12) 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 $\chi^2$ value? Repeat the randomization procedure. How did the distribution change? Compute the new $P$-value. What would you conclude? Explain why the $P$-value is smaller even though the sample proportions and null proportions were the same.

(13) The theoretical distribution for the $\chi^2$ statistic is called a $\chi^2$ distribution. Like the $t$, it is defined by a "degrees of freedom" parameter; but instead of coming from the sample size, the degrees of freedom is one less than the number of *categories*. The rationale as usual is that with only one category there is no notion of distance from expected proportions; but with each additional category, we have a new proportion that provides information about the discrepancy from our expectations.

For the theoretical distribution to be a good approximation, the conventional rule is that sample size should be large enough that the smallest *expected* count should be at least five or six. Is that condition satisfied for the datasets (original and doubled) that we have used?

(14) For now, ignore the fact that our sample size is too small, and find the theoretical tail proportions in StatKey (use the Theoretical Distributions: $\chi^2$ applet) by plugging in the two different $\chi^2$ values from the two datasets above. Are the $P$-values roughly the same as in the randomization tests?