LIN 313 Language and Computers
9/8/2025
Probaility Distributions

Our corpus: "The cat gave the dog the fig"

1. How many tokens does the corpus have?

2. How many types?

3. Draw a histogram of the tokens with types on the x axis and token counts on the y axis.

4. What are the predicted probabilities of each word? (remember probabilities add up to 100, you may have to do some rounding to get this to work out)
    a. P("the") =
    b. P("cat") =
    c. P("dog") =
    d. P("fig") =
    e. P("gave") =

5. **Uniform Distribution:** Open the weighted dice roller. Our weighted die has a face for every word type in the corpus, and we can assign probability weights to make the die loaded. The default settings record a uniform distribution. Let's run some trials.
    a. Run one trial with 10 dice rolls. What is the **observed probability distribution?**

    b. Run another single trial with 10 dice rolls. What is the **observed probability distribution**?

    c. Run a few more trials like this (don't record the results). What do you notice about the observed distribution?

d. When you run multiple trials, the dice roller reports the average frequency counts across all trials. Play around with the experiment settings. Perform 100 trials of 100 rolls. What is the observed probability distribution?

6. **Uniform Language Model:** The 'Roll the dice' button **samples** from the probabilty distribution (the actual values, not the observed ones), to pick a word. We can roll over and over again to build a new string based on our distribution.
    a. Use the 'Roll and add to string' button to generate a string of 20 words based on the uniform probability distribution. Record it here.

    b. Generate a second string and record it here.

7. **Unigram Distribution**: Now change the weights of the die to reflect the observed probabilities of each word in our corpus (from above).
    a. Run one trial with 10 runs. What is the observed probability distribution?

    b. Run a few more trials like this (don't record the results). What do you notice about the observed distribution?

    c. When you run multiple trials, the dice roller reports the average frequency counts across all trials. Perform 100 trials of 10 runs. What is the observed probability distribution

8. Unigram Language Model. Let's use the observed unigram distribution as a generative model for building new strings, just like before.
    a. Use the 'Roll and add to string' button to generate a string of 20 words. Record it.

    b. Generate a second string and record it here.

9. Look at the strings you generated with the uniform model and the unigram model. Which model is better? Which one produces strings that are more like English? What about the string makes it more like English?