Data Sampling and Probability

How to sample effectively, and how to quantify the samples we collect.
(Continued Discussion)

Recap: Generalization of binomial probabilities

If we are drawing at random with replacement **n** times, from a population in which a proportion **p** of the individuals are called "successes" (and the remaining **1 - p** are "failures"), then the probability of **k successes** (and hence, **n - k failures**) is

$$P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}$$

Multinomial probabilities

Suppose we again sample at random with replacement 7 times from a bag of marbles, but this time, 60% of marbles are **blue**, 30% are **green**, and 10% are **red**.

- What is P(**bgbbbgr**)?
  - Following the same steps as before:

$$P(\mathbf{bgbbbgr}) = 0.6 \times 0.3 \times 0.6 \times 0.6 \times 0.6 \times 0.3 \times 0.1 = (0.6)^4(0.3)^2(0.1)^1$$

- What is P(4 blue, 2 green, 1 red)?
  - As we saw before, we multiply the above probability by the total number of ways to draw 4 blue, 2 green, and 1 red marbles. This gives

$$P(4 \text{ blue}, 2 \text{ green}, 1 \text{ red}) = \frac{7!}{4!2!1!}(0.6)^4(0.3)^2(0.1)^1$$

Generalization of multinomial probabilities

If we are drawing at random with replacement **n** times, from a population broken into three separate categories (where $p_1 + p_2 + p_3 = 1$):

- Category 1, with proportion $\mathbf{p_1}$ of the individuals.
- Category 2, with proportion $\mathbf{p_2}$ of the individuals.
- Category 3, with proportion $\mathbf{p_3}$ of the individuals.

Then, the probability of drawing $\mathbf{k_1}$ individuals from Category 1, $\mathbf{k_2}$ individuals from Category 2, and $\mathbf{k_3}$ individuals from Category 3 (where $k_1 + k_2 + k_3 = n$) is

$$\frac{n!}{k_1!k_2!k_3!}p_1^{k_1}p_2^{k_2}p_3^{k_3}$$

At no point in this class will you be forced to memorize this! In practice, we use np.random.multinomial to compute these quantities.

Summary

- Formalized various ideas about sampling
  - Why we need to sample
  - What it means for the sample to biased
  - How to prevent these biases in the samples

- Compute probabilities from samples
  - Binomial and multinomial probabilities