Lec 3

Data Sampling and Probability

How to sample effectively, and how to quantify the samples we collect.

HW 1 will be posted today on Canvas.

Announcement: welcome new TA!

Name: Yucheng Hou

Junior, ECE

Email: hyc0716@sjtu.edu.cn

Office Hour: Wed. 13:00-15:00 (via Feishu)

Current Research Direction:

Reinforcement Learning in automated stock trading

Hobby:

Gaming, Badminton

Feel free to contact me via Feishu, Email and WeChat!

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

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

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.

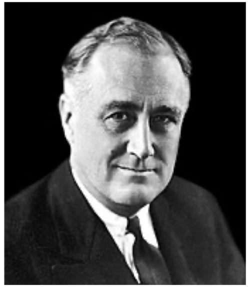
Then, the **multinomial probability** of drawing $\mathbf{k_1}$ individuals from Category 1, $\mathbf{k_2}$ individuals from Category 2, and $\mathbf{k_3}$ individuals from Category 3 (where $k_1 + k_2 + k_3 = n$) is

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

Revisit the "Literary Digest"

1936 U.S. Election:

- The *Literary Digest*'s sampling scheme was biased and did not represent the population. Their prediction was way off.
- But can we **quantify** this takeaway? What is the likelihood that the *Digest*'s differences arose simply due to **chance error** in their sample?



Roosevelt (D)

Landon (R)

We know the actual population distribution (i.e., election results).

- Assume the *Digest* did random sampling with replacement from the population.
- Simulate many different samples and generate many different predictions
- Draw a conclusion.

You have seen this process before in **Hypothesis Testing**.

	% Roosevelt	**# surveyed**
Actual election	61%	**All voters** (~45,000,000)
The Literary Digest poll	43%	10,000,000

Mark-Recapture Method

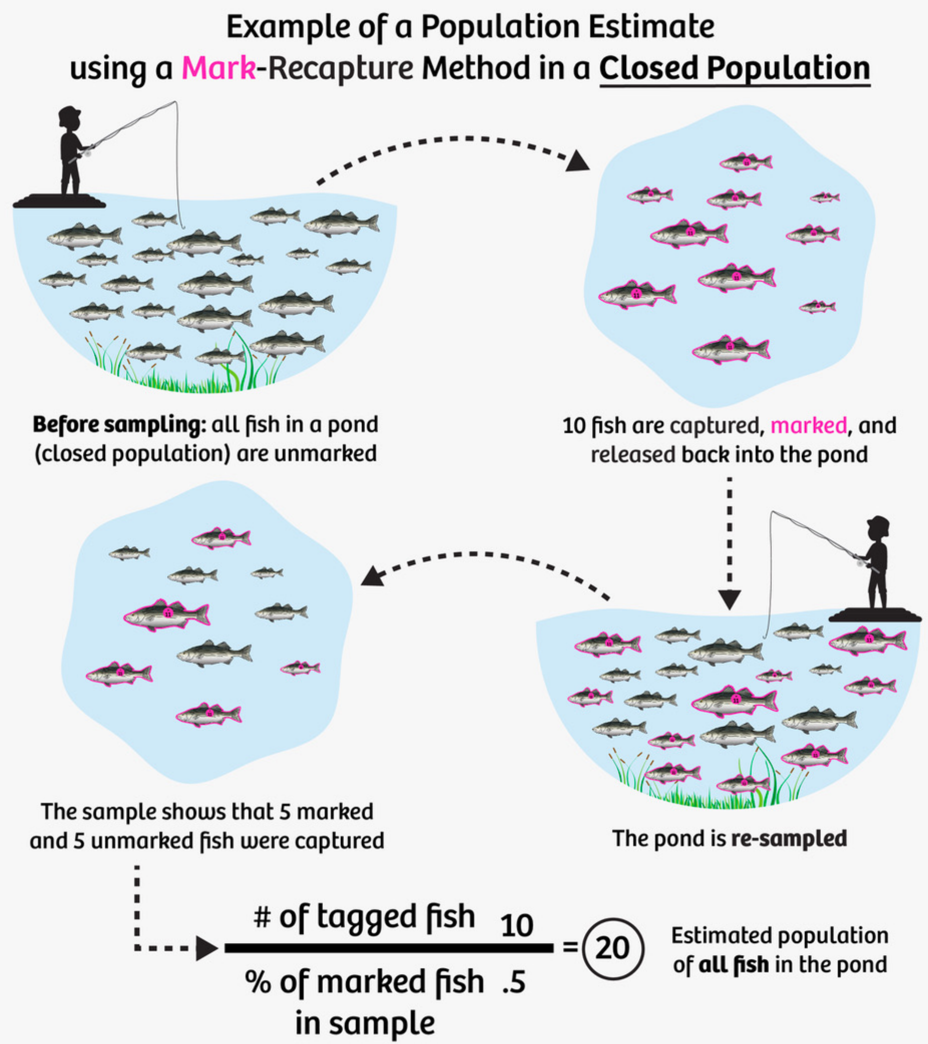
In the simplest case, a one-stage mark-recapture study produces the following data

- M : number of animals marked in first capture
- C : number animals in second capture
- R : number of marked animals in second capture.
- We are interested in N : number of animals in the population

$$\widehat{N} = \frac{MC}{R}$$

This population estimate would arise from a probabilistic model in which the number of recaptured animals is distributed binomially

- R ~ Binomial(C, p), where p = M/N
  (prerequisite: N is large, M/N > 0.1)



Example of a Population Estimate
using a Mark-Recapture Method in a Closed Population

Before sampling: all fish in a pond (closed population) are unmarked

10 fish are captured, marked, and released back into the pond

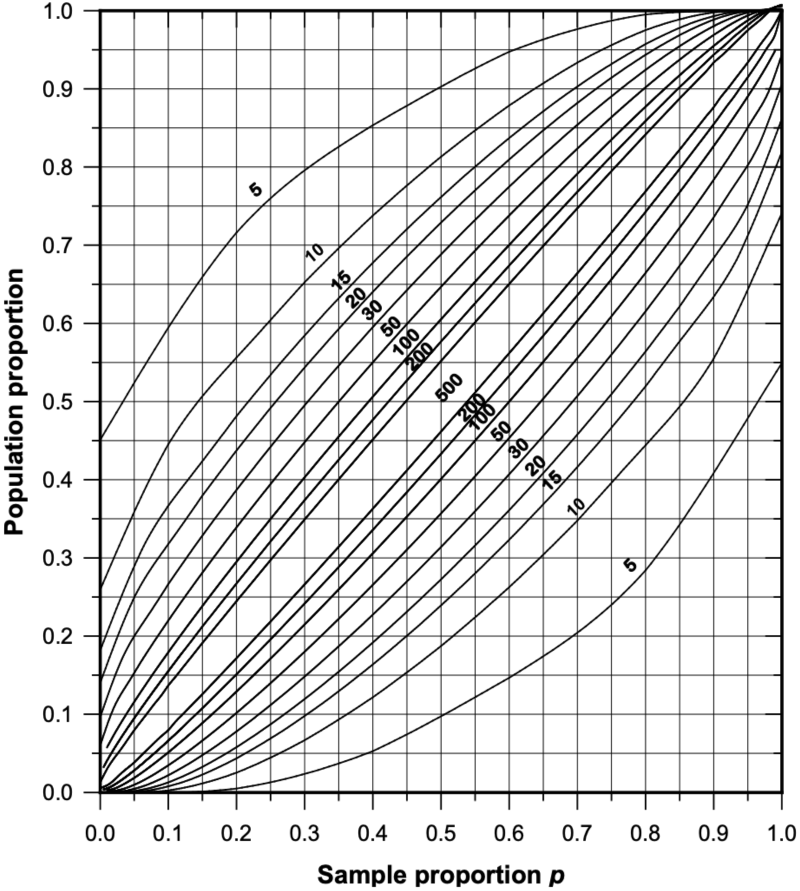
The pond is **re-sampled**

The sample shows that 5 marked and 5 unmarked fish were captured

$$\frac{\text{\# of tagged fish} \quad 10}{\text{\% of marked fish in sample} \quad .5} = 20$$

Estimated population of **all fish** in the pond

Binomial 95% Confidence Limits



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