**DISCOVERY:**

**Overall:**

Made up of a few components

Discovery, experiment, data, questions, probability, statistical inference, hypothesis testing, and resampling

**Hypothesis Testing:**

Made up of a few steps:

1. From vague null hyp. Identify test statistic t that is sensitive to truth of null hypothesis
2. Work the hypothesis into specific null hypothesis H0 that is a model for the scientific experiment and can simulate test statistics T
3. Repeat 1 and 2 for a good null hypothesis
4. Compute the p-value (prob. Of observing test statistic T as or more extreme that the observed test statistic t0 when H0 is true)

Vague hypothesis is not sufficient to mimic the experiment

Vague: the sample (x1 … xn) is iid normally distributed

Specific: (x1 … xn) N(0, 3); (the mean and variance are specified)

With specific null hypothesis H0, we can simulate T

Can simulate with python and modeling skills

Important statistic is Z-test

Z Value = (sample mean – hypothesis mean) / (Standard Dev. / sqrt(number of samplings))

= (avg(x) – µ0) / (stdDev/sqrt(n)) (normal distribution)

Where H0: x1 … xn ~ N(µ, stdDev2)

T Test = (sample mean – hypothesis mean) / (sample standard dev. / sqrt(number of samplings))

Nt – nn, nn is number of reads to normal cells, nt is number of reads mapped to tumor cells

Just review Binomial, Poisson, Bernoulli

**P-Values:**

A p value is the probability of obtaining a test statistic T as extreme or more extreme than the observed test statistic t0 when H0 is true

For a z test, the p-value is the sum of the tail areas (the areas outside the neg. and pos. z values

**Handout Answers Q3:**

1. Values closer to 1 suggest that X is bound by the protein (since, if the value is closer to 1, the most likely letter is chosen at each site)
2. The sequence X is not a binding site.

To simulate: X1 … Xn ~ Uniform({A, C, T, G}) -> P(X = A) = ¼

Better Sim: X1 … Xn ~ Multinoulli(Pa, Pt, Pc, Pg) -> P(Xi = N) = Pn

Used for HW,