Lisa Yan
CS 109

Section #7

The Honor Code

We have decided that automated tools should be used to identify if two submissions are suspiciously similar. (N.B. these tools are never used as a basis for community standards cases — those always require professional human opinion.) However, automated tools are never perfect. As active CS109 students we would like to explore the chance of a false positive in automated tools. For this task, we will consider Breakout, a CS106A assignment where students implement Breakout.

1. **Beta Sum**

   What is the distribution of the sum of 100 IID Betas? Let X be the sum

   $$X = \sum_{i=1}^{100} X_i \qquad \text{where each } X_i \sim \text{Beta}(a = 3, b = 4)$$

   Note the variance of a Beta:

   $$\text{Var}(X_i) = \frac{ab}{(a + b)^2(a + b + 1)} \qquad \text{where } X_i \sim \text{Beta}(a, b)$$

2. **Single Match**

   Say there are 1000 decision points when writing Breakout. Assume: Each decision point is binary. Each student makes all 1000 decisions. For each decision there is a probability p that a student takes the more popular choice. What is the probability distribution for the number of matching decisions (we are going to refer to this as the "score")? If possible, could you approximate this probability?

3. **Maximum Match**

   When we look at two assignments, the probability of a false match is exceedingly small. What would the max similarity score look like when we compare one student to 5000 historical breakout submissions? Let X_i be the similarity score between a student who worked on their own and student i. Let Y be the highest match score between the student and all other submissions:

   $$Y = \max_i X_i$$

   The Central Limit Theorem tells us about the distribution of the sum of IID random variables. A more obscure theorem, the Fisher-Tippett-Gnedenko theorem, tells us about the *max* of

IID random variables. It says that the max of IID exponential or normal random variables will be a "Gumbel" random variable.

$$Y \sim \text{Gumbel}(\mu, \beta) \qquad \qquad \text{The max of IID vars}$$

$$f(Y = k) = \frac{1}{\beta} e^{-(z + e^{-z})} \text{ where } z = \frac{k - \mu}{\beta} \qquad \text{The Gumbel PDF}$$

You are given data of 1300 students' max scores from three quarters (we believe they all worked independently): $y^{(1)} \ldots y^{(1300)}$. Set up (but do not solve) simultaneous equations we could solve to find the values of μ and β.