Chris Piech
CS 109

Section #6
Nov 7, 2018

Section 6: Samples

With questions from Will Monroe and Julia Daniel

Recent topics in CS109: Beta distribution, Central Limit Theorem, sampling and unbiased estimates of population parameters, bootstrapping, randomized algorithms

1. **Warmup**: *populations vs. samples*

   What is the difference between the population variance, σ^2, and sample variance, S^2? What is the difference between sample variance, S^2, and variance of the sample mean, $\text{Var}(\bar{X})$?

2. **Beta Sum**: *beta distribution and sum of RVs*

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

   $$X = \sum_{i=0}^{100} X_i \qquad \text{Where each } X_i \sim \text{Beta}(a = 3, b = 4)$$

   Either simulate the summation 10,000 times or use theory. Note the variance of a Beta:

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

3. **Variance of Height among Island Corgis**: *sampling and bootstrapping*

   A colleague has collected samples of heights of corgis that live on two different islands. The colleague collects 50 samples from both islands.

   

   The colleague notes that the sample mean is the same between the two groups: both are around 10 inches. However, island B has a sample **variance** that is 3 in^2 greater than island A. The colleague wants to make a scientific claim that corgis on island A have a significantly higher spread of heights than corgis on island B. You are skeptical. It is possible that heights are identically distributed across both islands and that the observed difference in variance was a result of chance and a small sample size (the null hypothesis).

Calculate the probability of the null hypothesis using bootstrapping. Here is the data. Each number is the height, in inches, of an independently sampled corgi:

Island A Corgi Heights (S^2 = 6.0):
13, 12, 7, 16, 9, 11, 7, 10, 9, 8, 9, 7, 16, 7, 9, 8, 13, 10, 11, 9, 13, 13, 10, 10, 9, 7, 7, 6, 7, 8, 12, 13, 9, 6, 9, 11, 10, 8, 12, 10, 9, 10, 8, 14, 13, 13, 10, 11, 12, 9

Island B Corgi Heights (S^2 = 9.1):
8, 8, 16, 16, 9, 13, 14, 13, 10, 12, 10, 6, 14, 8, 13, 14, 7, 13, 7, 8, 4, 11, 7, 12, 8, 9, 12, 8, 11, 10, 12, 6, 10, 15, 11, 12, 3, 8, 11, 10, 10, 8, 12, 8, 11, 6, 7, 10, 8, 5

Discuss: How would this calculation be different if you were interested in looking at the statistical significance of: sample mean? 95th percentile?

4. **Traffic Lights**: *Stretch problem with multiple types of continuous RVs and convolution*

   On the midterm, we reasoned a bit about traffic lights, but let's take it further. Suppose that a rider bikes to work with an average speed that is normally distributed with mean 10 and std dev 2 mph. (There is no need to account for variations in speed over the course of the ride, since this figure represents average speed - assume the only variations in ride duration come from time spent at traffic lights.) The route from home to work is two miles.

   For all lights on your commute: when you arrive at the light, there is a 50% chance that the light is green and a 50% chance that the light is red (we treat yellow as green). If the light is green, your wait time is 0. If the light is red, your wait time is equally likely to be any value in the continuous range 0 to 4 mins.

   a. What is the probability of a commute duration under 10 minutes if the route has 0 traffic lights?
   b. What is the probability of a total wait time under 8 minutes if the route has 10 traffic lights? Make your life easier by using what we know about the sum of independent random variables.
   c. What is the probability of a commute duration under 20 minutes if the route has 10 traffic lights? Hint: you may want to approximate using a Riemann sum, which may require coding.