Chris Piech
CS109

Section #6
Nov 10, 2017

Section #6: Samples Solution

1. **Warmup**:

   - Population variance, σ^2: The true variance of a population (or random variable).
   - Sample variance, S^2: the unbiased estimate of the true variance based on an independent subsample.
   - Variance of sample mean, $\text{Var}(\bar{X})$: How much spread there is in the estimation of the true mean.

2. **Binary Tree**:

   Let X_1 and X_2 be number of nodes the left and right calls to `randomTree`.
   $E[X_1] = E[X_2] = E[X]$.

$$\begin{aligned}
E[X] &= p \cdot E[X \mid \texttt{if}] + (1-p)E[X \mid \texttt{else}] \\
&= p \cdot E[1 + X_1 + X_2] + (1-p) \cdot 0 \\
&= p \cdot (1 + E[X] + E[X]) \\
&= p + 2pE[X] \\
(1-2p)E[X] &= p \\
E[X] &= \frac{p}{1-2p}
\end{aligned}$$

3. **Beta Sum**:

   By the Central Limit Theorem, the sum of equally weighted IID random variables will be Normally distributed. First, we calculate the expectation and variance of X_i using the beta formulas:

$$\begin{aligned}
E(X_i) &= \frac{a}{a+b} && \text{Expectation of a Beta} \\
&= \frac{3}{7} \approx 0.43 \\
\text{Var}(X_i) &= \frac{ab}{(a+b)^2(a+b+1)} && \text{Variance of a Beta} \\
&= \frac{3 \cdot 4}{(3+4)^2(3+4+1)} \\
&= \frac{12}{49 \cdot 8} \approx 0.03
\end{aligned}$$

$$\begin{aligned}
X &\sim N(\mu = n \cdot E[X_i], \sigma^2 = n \cdot \text{Var}(X_i)) \\
&\sim N(\mu = 100 \cdot 0.43, \sigma^2 = 100 \cdot 0.03) \\
&\sim N(\mu = 43, \sigma^2 = 3)
\end{aligned}$$

4. Variance of Height among Island Corgis:

```
def bootstrap(pop1, pop2):
    # make the universal population
    totalPop = copy.deepcopy(pop1)
    totalPop.extend(pop2)

    # Run a bootstrap experiment
    countDiffGreaterThanObserved = 0
    print 'starting bootstrap'
    for i in range(50000):
        # resample and recalculate the statistic
        sample1 = resample(totalPop, len(pop1))
        sample2 = resample(totalPop, len(pop2))
        sampleStat1 = calcSampleVariance(sample1)
        sampleStat2 = calcSampleVariance(sample2)
        diff = abs(sampleStat2 - sampleStat1)

        # count how many times the statistic is more extreme
        if diff >= 3:
            countDiffGreaterThanObserved += 1

    # compute the p-value
    p = float(countDiffGreaterThanObserved) / 50000
    print 'p-value:', p
```

For this data, the two-tailed (eg using absolute value) test returns a null hypothesis probability **p = 0.12**. There is a pretty decent chance that the observed difference in sample variance was random chance – and it doesn't fall under what scientists often call "statistically significant." Here is a histogram of all the diff values from the bootstrap experiment: