Chris Piech
CS 109

Section #6
May 15-17, 2019

Section 6 Solution

With questions from Will Monroe and Julia Daniel

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, $\text{Var}(\bar{X})$?

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

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:

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

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

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

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

3. **Food for thought** *CLT*

   Karel the dog eats an unpredictable amount of food. Every day, the dog is equally likely to eat between a continuous amount in the range 100 to 300 ml. How much Karel eats is independent of all other days. You only have 6.5kg of food for the next 30 days. What is the probability that 6.5kg will be enough for the next 30 days?

The distribution of the sum is given by the central limit theorem. Let $X_i \sim \text{Uni}(100, 300)$ where $E[X_i] = 200$ and $Var(X_i) = \frac{1}{12}(200)^2 \approx 3333$.

$$Y = \sum_i X_i$$

Let's approximate Y with a normal R.V.

$$\sim \mathcal{N}(6000, 316.212^2)$$

$$P(Y < 6500)$$

$$P\left(\frac{Y - 6000}{316.212} < \frac{6500 - 6000}{316.212}\right)$$

Let $\frac{Y-6000}{316.212} = Z \sim \mathcal{N}(0, 1)$

$$P\left(Z < \frac{6500 - 6000}{316.212}\right)$$

$$P\left(Z < 1.58\right)$$

$$\Phi\left(1.58\right)$$

4. **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, i.e. 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 the difference in sample mean? 95th percentile?

```
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."