Lisa Yan
CS 109

Section #6
November 6-8, 2019

Section 6 Solution

Adapted for Fall 2019 by Oishi Banerjee. Contains 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

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

   Note the variance of a Beta:

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

   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. **Variance of Hemoglobin Levels**: *sampling and bootstrapping*

   A medical researcher treats patients with dangerously low hemoglobin levels. She has formulated two slightly different drugs and is now testing them on patients. First, she administered drug A to one group of 50 patients and drug B to a separate group of 50 patients. Then, she measured all the patients' hemoglobin levels post-treatment.

   For simplicity, assume that all variation in the patient outcomes is due to their different reactions to treatment.

   The researcher notes that the sample mean is similar between the two groups: both have mean hemoglobin levels around 10g/dL. However, drug B's group has a **sample variance** that is 3 $(\text{g/dL})^2$ **greater** than drug A's group. The researcher thinks that patients respond to drugs A and B differently. Specifically, she wants to make the scientific claim that drug A's patients will end up with a significantly different spread of hemoglobin levels compared to drug B's.

   You are skeptical. It is possible that the two drugs have practically identical effects and that the observed different 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 level of an independently sampled patient:

   Hemoglobin Levels of Drug A's Group ($S^2 = 6.0$):
   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

   Hemoglobin Levels of Drug B's Group ($S^2 = 9.1$):
   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(sample1, sample2):
   \# make the universal population
   totalSample = copy.deepcopy(sample1)
   totalSample.extend(sample2)

   \# Run a bootstrap experiment
   countDiffGreaterThanObserved = 0
   print 'starting bootstrap'
   for i in range(50000):
      \# resample and recalculate the statistic
      resample1 = resample(totalSample, len(sample1))
      resample2 = resample(totalSample, len(sample2))
      resampleStat1 = calcSampleVariance(resample1)
```

```
            resampleStat2 = calcSampleVariance(resample2)
            diff = abs(resampleStat2 - resampleStat1)
            \# 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 (e.g. 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 from random chance – and it doesn't fall under what scientists often call "statistically significant."

4. **Medicine Doses**:

   Megha has a health condition that requires unpredictable amounts of medication. Every day, there is a 20% chance that she feels perfectly fine and requires no medicine. Otherwise, she needs to take a dose of medication. The necessary dose is equally likely to be any value in the continuous range 1 to 5 ounces. How much medicine she needs on any given day is independent of all other days.

   Megha's insurance will fully cover 90 ounces of medicine for each 30-day period. What is the probability that 90 ounces will be enough for the next 30 days? Make your life easier by using Central Limit Theorem.

Let M be the amount of medicine Megha will need in the next thirty days. Let M_i be the amount of medicine Megha needs on the ith day. M is a sum of M_1 through M_{30} and can be modeled with the CLT.

To use the CLT, we need to first know the mean and variance of M_i. To do this, let D_i be the event that she needs to take a dose on the ith day. Note that $M_i|D_i \sim Uni(1,5)$ and $M_i|D_i^C = 0$. Using the law of total expectation, we have:

$$E[M_i] = E[M_i|D_i]P(D_i) + E[M_i|D_i^C]P(D_i^C) = 3 * 0.8 + 0 * 0.2 = 2.4$$

To find the variance of M_i, we need to know $E[M_i^2]$. We can use a similar approach as the previous problem along with the law of the unconscious statistician:

$$\begin{aligned} E[M_i^2] &= E[M_i^2|D_i]P(D_i) + E[M_i^2|D_i^C]P(D_i^C) \\ &= \frac{4}{5}\int_{m=1}^{5} m^2 f_M(m)dm + 0 * .2 \\ &= \frac{4}{5}\int_{m=1}^{5} m^2 \frac{1}{4}dt \approx 8.267 \end{aligned}$$

We then have $Var(M_i) = E[M_i^2] - E[M_i]^2 = 8.267 - 2.4^2 = 2.507$. According to the CLT:

$$\sum_{i=1}^{30} M_i \approx N(30*2.4, 30*2.507) \implies M \sim N(72, 75.21) P(M < 90) \approx \Phi\left(\frac{90-72}{\sqrt{75.21}}\right) \approx 0.98$$