Noah Arthurs
CS 109

Section #5
August 5, 2019

Section 5 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:

   $$\begin{aligned} E(X_i) &= \frac{a}{a+b} && \text{Expectation of a Beta} \\ &= \frac{3}{7} \approx 0.43 \end{aligned}$$

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

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

3. **Variance of 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."

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

   Suppose that you bike to work with a constant speed that is normally distributed with mean 10 and std dev 2 mph. 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?

Let T be the commute time and S be your speed:

$$P(T < 10) = P(S > 12) = 1 - \Phi\left(\frac{12-10}{2}\right) \approx 0.16$$

   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 the Central Limit Theorem to approximate.

Let T_i be the amount of time we spend waiting at the ith traffic light. To use the CLT, we need to know the mean and Variance of T_i. To do this, let W_i be the event that

we have to wait at the ith traffic light. Note that $T_i|W_i \sim Uni(0,4)$ and $T_i|W_i^C = 0$. Using the law of total expectation, we have:

$$E[T_i] = E[T_i|W_i]P(W_i) + E[T_i|W_i^C]P(W_i^C) = 2 * 0.5 + 0 * 0.5 = 1$$

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

$$\begin{aligned} E[T_i^2] &= E[T_i^2|W_i]P(W_i) + E[T_i^2|W_i^C]P(W_i^C) \\ &= \frac{1}{2}\int_{t=0}^{4} t^2 f_T(t)dt + 0 * 0.5 \\ &= \frac{1}{2}\int_{t=0}^{4} t^2 \frac{1}{4}dt = \frac{8}{3} \end{aligned}$$

We then have $Var(T_i) = E[T_i^2] - E[T_i]^2 = \frac{8}{3} - 1 = \frac{5}{3}$. According to the CLT:

$$\sum_{i=1}^{10} T_i \approx N(10, 50/3) \implies P\left(\sum_{i=1}^{10} T_i < 8\right) \approx \Phi\left(\frac{8-10}{\sqrt{50/3}}\right) \approx 0.31$$

c. What is the probability of a commute duration under 20 minutes if the route has 10 traffic lights? As in the last part, feel free to use reasonable approximations.

Let T = total time, T_W = wait time, and S = speed:

$$\begin{aligned} P(T < 20|T_W = w) &= P\left(S > \frac{2}{20-w}\right), \text{ so:} \\ P(T < 20) &= \int_{w=-\infty}^{\infty} f_W(w)P(T < 20|T_W = w)dw \\ &= \int_{w=0}^{40} f_W(w)\left(S > \frac{2}{20-w}\right)dw \\ &= \int_{w=0}^{40} f_W(w)\left(1 - F_S\left(\frac{2}{20-w}\right)\right)dw \approx 0.3054 \end{aligned}$$

The last integral was approximated using a Riemann sum due to its unintegratability.