CS 109

Section #6
August 2nd, 2023

Section 6

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 300g. 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?

```
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."

5. **Timing Attack**:

   In this problem we are going to show you how to crack a password in linear time, by measuring how long the password check takes to execute (see code below). Assume that our server takes T ms to execute any line in the code where $T \sim N(\mu = 5, \sigma^2 = 0.5)$ seconds. The amount of time taken to execute a line is always independent of other values of T.

```
# An insecure string comparison
def string_equals(guess, password):
    n_guess = len(guess)
    n_password = len(password)
    if n_guess != n_password:
        return False                # 4 lines executed to get here
    for i in range(n_guess):
        if guess[i] != password[i]:
            return False            # 6 + 2i lines executed to get here
    return True                     # 5 + 2n lines executed to get here
```

   On our site all passwords are length 5 through 10 (inclusive) and are composed of lower case letters only. A hacker is trying to crack the root password which is "gobayes" by carefully measuring how long we take to tell them that her guesses are incorrect.

   a. What is the distribution of time that it takes our server to execute k lines of code? Recall that each line independently takes $T \sim N(\mu = 5, \sigma^2 = 0.5)$ ms.

> Let Y be the amount of time to execute k lines. $Y = \sum_{i=1}^{k} X_i$ where X_i is the amount of time to execute line i. $X_i \sim N(\mu = 5, \sigma^2 = 0.5)$.
>
> Since Y is the sum of independent normals:
>
> $$Y \sim N(\mu = \sum_{i=1}^{k} 5, \sigma^2 = \sum_{i=1}^{k} 0.5)$$
> $$\sim N(\mu = 5k, \sigma^2 = 0.5k)$$

   b. First the hacker needs to find out the length of the password. What is the probability that the time taken to test a guess of correct length (server executes 6 lines) is longer than the time taken to test a guess of an incorrect length (server executes 4 lines)? Assume that the first letter of the guess does not match the first letter of the password. Hint: $P(A > B)$ is the same as $P(A - B > 0)$.

> From last problem:
> Time to run 6 lines of code $A \sim N(\mu = 30, \sigma^2 = 3)$

Time to run 4 lines of code $B \sim N(\mu = 20, \sigma^2 = 2)$

$$\begin{aligned}
-B &\sim N(\mu = -20, \sigma^2 = 2) \\
A - B &\sim N(\mu = 10, \sigma^2 = 5) \\
P(A > B) &= P(A - B > 0) \\
&= 1 - F_{A-B}(0) \\
&= 1 - \Phi\Big(\frac{0 - 10}{\sqrt{5}}\Big) \\
&\approx 1.0
\end{aligned}$$

c. Now that our hacker knows the length of the password, to get the actual string she is going to try and figure out each letter one at a time, starting with the first letter. The hacker tries the string "aaaaaaa" and it takes 27s. Based on this timing, how much more probable is it that first character did not match (server executes 6 lines) than the first character did match (server executes 8 lines)? Assume that all letters in the alphabet are equally likely to be the first letter.

Let M be the event that the first letter matched.

$$\begin{aligned}
\frac{P(M^C|T = 27)}{P(M|T = 27)} &= \frac{f(T = 27|M^C)P(M^C)}{f(T = 27|M)P(M)} \\
&= \frac{f(T = 27|M^C)\frac{25}{26}}{f(T = 27|M)\frac{1}{26}} \\
&= 25 \cdot \frac{f(T = 27|M^C)}{f(T = 27|M)} \\
&= 25 \cdot \frac{\frac{1}{\sqrt{6\pi}} e^{-\frac{(27-30)^2}{6}}}{\frac{1}{\sqrt{8\pi}} e^{-\frac{(27-40)^2}{8}}} \\
&= 25 \cdot \frac{\sqrt{8}}{\sqrt{6}} \cdot \frac{e^{-\frac{9}{6}}}{e^{-\frac{169}{8}}} \\
&\approx 9.6 \text{ million}
\end{aligned}$$

d. If it takes the hacker 6 guesses to find the length of the password, and 26 guesses per letter to crack the password string, how many attempts does she need to crack our password, "gobayes"? Yikes!

$7 \cdot 26 + 6 = 188$