Lisa Yan and Jerry Cain
CS109

Section #6
October 27-28, 2020

Section #6: Samples Solution

1 Warmups

1.1 Sums of Random Variables

For each X and Y below, let X and Y be independent.

1. Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. What is μ and σ for $X + Y \sim \mathcal{N}(\mu, \sigma)$?
2. Let $X \sim \text{Uni}(0, 1)$ and $Y \sim \text{Uni}(0, 1)$. What is the PDF for $X + Y$?
3. In general, two random variables X and Y, what is the PDF f of $X + Y$?

Solution

1. $\mu = \mu_1 + \mu_2$ and $\sigma = \sigma_1^2 + \sigma_2^2$. How convenient!
2. $f_{X+Y}(a) = \begin{cases} a & 0 \le a \le 1 \\ 2 - a & 1 \le a \le 2 \\ 0 & \text{otherwise} \end{cases}$
3. $f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a - y) f_y(y) dy$

   It is good to remember these equations, but perhaps another message from lecture that it is difficult to sum random variables. The derivation for Uniform distributions is difficult. And solving for the general random variables is even worse. But we can pick distributions, like the Normal distribution, that are easy to use!

1.2 Sample and Population Mean

Computing the sample mean is similar to the population mean: sum all available points and divide by the number of points. However, sample variance is slightly different from population variance.

1. Consider the equation for population variance, and an analogous equation for sample variance.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{1}$$

$$S^2_{biased} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{2}$$

   S^2_{biased} is a random variable which estimates the constant σ^2. Is $E[S^2_{biased}]$ greater or less than σ^2?

2. Write the equation for $S^2_{unbiased}$ (known simply as S^2 in the slides). This is known as *Bessel's correction*.

Solution

1. $E[S^2_{biased}] < \sigma^2$. The intuition is that the spread of a sample of points is generally smaller than the spread of all the points considered together.
2. $S^2_{unbiased} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

1.3 Sum of I.I.D Random Variables

What is the distribution (with name and parameter(s)) of the average of n i.i.d. random variables, X_1, ..., X_n, each with mean μ and variance σ^2?

Solution

According to the central limit theorem, this can be modeled as $N(\mu, \sigma^2/n)$.

2 Problems

2.1 Variance of Hemoglobin Levels

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 $(g/dL)^2$ **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

How would this calculation be different if you were interested in looking at the statistical significance of the difference in sample mean? Or the 95th percentile?

Solution

```
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."

2.2 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.

Solution 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:

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