11 Confidence Intervals, Conditional Expectation, LLSE

Computer Science Mentors 70

Apr 19 - Apr 23
Spring 2021

Amogh Gupta, Sylvia Jin, Aekus Bhathal, Abinav Routhu, Debayan Bandyopadhyay

Roast us here: `https://tinyurl.com/csm70-feedback20`

1 Confidence Intervals

1. Define i.i.d. variables $A_k \sim \text{Bern}(p)$ where $k \in \{1, \cdots, n\}$. Assume we can declare that $\Pr\left[\left|\frac{1}{n}\sum_{k=1}^{n} A_k - p\right| \geq 0.25\right] = 0.01$.

(a) Please give a 99% confidence interval for p given A_k.

Solution: Notice that the mean is $\frac{1}{n}\sum_{k=1}^{n} A_k$, and our confidence interval should cover the parts that are within 0.25 of the mean, leading to an interval of the form:

$$\left(\frac{1}{n}\sum_{k=1}^{n} A_k - 0.25, \frac{1}{n}\sum_{k=1}^{n} A_k + 0.25\right)$$

(b) We know that the variables $X_1, \ldots, X_n$, are i.i.d. random variables and have variance σ^2. We also have the observation that $A_n = \frac{X_1 + \ldots + X_n}{n}$. We want to estimate the mean, μ, of each X_i.

Prove that we have 95% confidence that μ lies in the interval $\left[A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}\right]$

That is, $\Pr\left[A_n - 4.5\frac{\sigma}{\sqrt{n}} \leq \mu \leq A_n + 4.5\frac{\sigma}{\sqrt{n}}\right] \geq 95\%$

Solution: To do this, we use Chebyshev's. Because $\mathbb{E}[A_n] = \mu$ (since A_n is the average of the X_i's), we bound the probability that $|A_n - \mu|$ is more than the interval size at 5%:

$$\Pr\left[|A_n - \mu| \geq 4.5\frac{\sigma}{\sqrt{n}}\right] \leq \frac{\text{Var}(A_n)}{(4.5\frac{\sigma}{\sqrt{n}})^2} \approx \frac{\frac{\sigma^2}{n}}{\frac{20\sigma^2}{n}} = \frac{1}{20} = 5\%$$

.

Thus, the probability that μ is *in* the interval is 95%.

(c) Give the 99% confidence interval for μ.

Solution: Solution is similar to that of the 95% confidence interval. Our 99% confidence interval is $\left[A_n - 10\frac{\sigma}{\sqrt{n}}, A_n + 10\frac{\sigma}{\sqrt{n}}\right]$ since

$$\Pr\left[|A_n - \mu| \geq 10\frac{\sigma}{\sqrt{n}}\right] \leq \frac{\text{Var}(A_n)}{(10\frac{\sigma}{\sqrt{n}})^2} \approx \frac{\frac{\sigma^2}{n}}{\frac{100\sigma^2}{n}} = \frac{1}{100} = 1\%$$

.

2. We have a die with 6 faces labeled $1, 2, 3, 4, 5, 6$.

(a) Develop a 99% confidence interval for the sum of n samples.

Solution: Consider the 99% confidence interval for the average of n rolls of a standard die, $X = \frac{1}{n}\sum_{i=1}^{n} D_i$.

D_i has expectation $\mu = 3.5$ and variance $\sigma^2 = \frac{105}{36}$. Accordingly, X has expectation μ and variance $\frac{\sigma^2}{n}$. Knowing this, we can bound using Chebyshev's inequality:

$$\Pr\left[|X-\mu| \geq a\frac{\sigma}{\sqrt{n}}\right] \leq \frac{\text{Var}(X)}{(a\frac{\sigma}{a\sqrt{n}})^2} = \frac{\frac{\sigma^2}{n}}{\frac{a^2\sigma^2}{n}} = \frac{1}{a^2}$$

.

We would like:

$$\Pr\left[|X-\mu| \geq a\frac{\sigma}{\sqrt{n}}\right] \leq \frac{1}{100} = 1\%$$

Thus:

$$\frac{1}{a^2} \leq \frac{1}{100} \longrightarrow a \geq 10$$

$$X \in \left(\mu - 10\frac{\sigma^2}{\sqrt{n}},\ \mu + 10\frac{\sigma^2}{\sqrt{n}}\right) \quad \text{with prob.} \geq 99\%$$

(b) Now, suppose the die's face values are just 6 consecutive integers $k+1, k+2, \ldots, k+6$, but we do not know k. For example, if $k = 6$, the die faces would take on the values $7, 8, 9, 10, 11, 12$. If we observe that the average of the n samples is 15.5, develop a 99% confidence interval for the value of k.

Solution: If we define $\tilde{D}_i = D_i + k$, then we find $\tilde{D}_i$ has expectation $\tilde{\mu} = \mu + k$ and variance $\tilde{\sigma}^2 = \sigma^2$.
Then, $E[\tilde{X}] = E[X] + k$ and $Var[\tilde{X}] = Var[X]$. This is great because this means there is no need to redo our work.

Earlier, we used our knowledge of μ to estimate X,

$$X \in \left(\mu - 10\frac{\sigma^2}{\sqrt{n}},\ \mu + 10\frac{\sigma^2}{\sqrt{n}}\right) \quad \text{with prob.} \geq 99\%$$

If instead, we were to estimate μ based on knowledge of X, we'd have

$$\mu \in \left(X - 10\frac{\sigma^2}{\sqrt{n}},\ X + 10\frac{\sigma^2}{\sqrt{n}}\right) \quad \text{with prob.} \geq 99\%$$

Since the variances of $\tilde{X}$ and X are the same:

$$\tilde{\mu} \in \left(\tilde{X} - 10\frac{\sigma^2}{\sqrt{n}},\ \tilde{X} + 10\frac{\sigma^2}{\sqrt{n}}\right) \quad \text{with prob.} \geq 99\%$$

$$\mu + k \in \left(\tilde{X} - 10\frac{\sigma^2}{\sqrt{n}},\ \tilde{X} + 10\frac{\sigma^2}{\sqrt{n}}\right) \quad \text{with prob.} \geq 99\%$$

$$k \in \left((\tilde{X}-\mu) - 10\frac{\sigma^2}{\sqrt{n}},\ (\tilde{X}-\mu) + 10\frac{\sigma^2}{\sqrt{n}}\right) \quad \text{with prob.} \geq 99\%$$

$$k \in \left(12 - 10\frac{\sigma^2}{\sqrt{n}},\ 12 + 10\frac{\sigma^2}{\sqrt{n}}\right) \quad \text{with prob.} \geq 99\%$$

3. **Looping Ropes**

Frobenius has n ropes in his backyard, which he likes a lot. But when he goes to work everyday, they grow a mind of their own and begin behaving weirdly. At every timestep t, two ends of a rope(s) are uniformly chosen at random and knotted together. If the two ends are from the same rope, they form a loop. If the two ends are from different ropes, they join together to form a new rope. By the time Frobenius comes home, this process has completed (meaning no more loose ends are left). How many loops can Frobenius expect to see? **Bonus:** Does this converge as $n \longrightarrow \infty$?

Solution: let X_n be the number of loops when there are n ropes left. Then, at the next timestep, we can either pick two ends from the same rope or two ends from different ropes.

If we pick two ends from the same rope, we form one more loop and there are $n-1$ ropes left. If we pick two ends from different ropes, we dont change the number of loops, and still the number of ropes decreases by 1 (since we combine two ropes into one). Let I be the event that we pick two ends of the same rope. Now, we can use conditional expectation to represent the number of loops we expect to see.

$$\begin{aligned}\mathbb{E}[X_n] &= \mathbb{E}[X_n|I]\mathbb{P}[I] + \mathbb{E}[X_n|\bar{I}]\mathbb{P}[\bar{I}] && (1)\\ &= (\mathbb{E}[X_{n-1}]+1)\mathbb{P}[I] + \mathbb{E}[X_{n-1}]\mathbb{P}[\bar{I}] && (2)\\ &= E[X_{n-1}] + \mathbb{P}[I] && (3)\end{aligned}$$

If there are n ropes, there are $2n$ ends and thus $\frac{1}{2n-1}$ probability of pick two ends on the same rope (after we pick the first, there are $2n-1$ ends left). Thus $\mathbb{P}[I] = \frac{1}{2n-1}$. This gives us the recursive relation,

$$\mathbb{E}[X_n] = \mathbb{E}[X_{n-1}] + \frac{1}{2n-1} \tag{4}$$

with the base case $\mathbb{E}[X_1] = 1$ (because if there is one rope, both ends must on the same rope and thus form a loop). Fortunately, this is just the recursive relation for an arithmetic sequence, hence:

$$\mathbb{E}[X_n] = \sum_{i=1}^{n} \frac{1}{2i-1} = 1 + \frac{1}{3} + \frac{1}{5} + \ldots \tag{5}$$

Bonus

——————-

This can be represented as a partial sum of the difference of two harmonic series:
(It's the harmonic series with every other term cancelled)

$$\sum_{i=1}^{n} \frac{1}{2i-1} = \sum_{i=1}^{2n} \frac{1}{i} - \frac{1}{2}\sum_{i=1}^{n} \frac{1}{i} = H_{2n} - \frac{1}{2}H_n \tag{6}$$

Since $H_{2n} > H_n$, we have that $H_{2n} - \frac{1}{2}H_n > \frac{1}{2}H_n$. Therefore,

$$\lim_{n\to\infty} \mathbb{E}[X_n] \geq \lim_{n\to\infty} \frac{1}{2}H_n \tag{7}$$

Since the harmonic series doesn't converge neither does the number of loops.

4. **Rolling Chopsticks**

The content mentors were trying to eat noodles in a new way. Rather than eating noodles by chopsticks directly, they tried eating noodles by rolling one noodle on the chopstick and eat it. This is seemingly a hard way to eat noodles so the probability they successfully eat a noodle on each attempt is p.

(a) Suppose they attempt to eat a noodle, and eat the noodle on the attempt X. What is the distribution of X? What is the distribution of unsuccessful attempts to eat that noodle, X', in terms of X?

Solution: We notice that X is modeling how long it takes until the first successful attempt hence $X \sim \text{Geometric}(p)$. Now if $X = k$ we know the last one is the successful attempt so there are $k-1$ unsuccessful attempts. Then $X' = X - 1$. So the distribution of unsuccessful attempts to eat that noodle $X' \sim X - 1$. That is $\mathbb{P}[X' = k] = \mathbb{P}[X = k+1]$

(b) Let Y be the number of unsuccessful attempts in trying to eat 2 noodles. What is the distribution of Y?

Solution: We can model Y as $Y = X_1' + X_2'$, where X_1' and X_2' are identical and independent random variables over the same distribution as X'.

For the distribution of Y, if $Y = k$ then there were k unsuccessful attempts between trying to eat both noodles. We observe that if we made w unsuccessful attempts before 1^{st} successful attempt, then we made $k - w$ unsuccessful attempts before the 2^{nd} successful event. There are k+1 different possibilities from $w = 0$ to $w = k$.

Also we know the distribution of X' only in terms of X, so we'll need to substitute $X' = X - 1 \implies X = X' + 1$.

$$\begin{aligned}
\mathbb{P}[Y = k] &= \mathbb{P}[X_1' + X_2' = k] \\
&= \sum_{w=0}^{k} \mathbb{P}[X_1' = w \cap X_2' = k - w] \\
&= \sum_{w=0}^{k} \mathbb{P}[X_1 = w + 1 \cap X_2 = k - w + 1] \\
&= \sum_{w=0}^{k} (1-p)^w p \cdot (1-p)^{k-w} p \\
&= (1-p)^k p^2 \cdot \sum_{w=0}^{k} 1 \\
&= (k+1)(1-p)^k p^2
\end{aligned}$$

(c) Not content with their distribution Y and eating 2 noodles, the content mentors want to find the distribution Z for the total unsuccessful attempts of eating the whole bowl of R noodles. They were planning to proceed as part b) but then Aekus, a random variable distribution enthusiast, suggested to use $P(Z = k) = \binom{r+k-1}{k}(1-p)^k p^r$ where $r = R$.

The distribution Z is defined by 2 parameters: 1) r - the number of successful attempts and 2) p - the probability of a successful attempt so we will write Z as $Z(r, p)$.

Show by induction on r with base case $r = 1$ that Aekus's suggestion is correction; Z is the sum of independent random variables drawing from the distribution of X'. (*Hint*: Remember the "Hockey stick" identity $\sum_{i=0}^{k-1} \binom{n+i}{i} = \binom{n+k}{k-1}$)

Solution: For intuition let's try to put small values for r.

$$\begin{aligned}
\mathbb{P}[Z(1, p) = k] &= \binom{1 + k - 1}{k}(1-p)^k p^1 \\
&= (1-p)^k p^1 \\
&= \mathbb{P}(X = k + 1) = \mathbb{P}(X' = k)
\end{aligned}$$

When $r = 1$ we get that $Z(1, p) \sim X'$ and when $r = 2$ we get $Z(2, p) \sim Y$. Now this helps us to see that we could use our old friend induction to prove this.
Base Case: The base case is when $r = 1$ and is true as we saw above.
Hypothesis: Assume that $Z(r, p)$ is the sum of r X_i', where $X_i' = X_i - 1$ for IID $X_i \sim$ Geometric(p).
Step : Now let $W = Z + X'$. We show that W is the Z distribution with $r_W = r + 1$.

$$\begin{aligned}
\mathbb{P}[W = w] &= \mathbb{P}[X' + Z = w] \\
&= \sum_{k=0}^{w} \mathbb{P}[(Z = k) \cap (X' = w - k)] \\
&= \sum_{k=0}^{w} \mathbb{P}[(Z = k) \cap (X = w - k + 1)] && \text{Using the substitution } X' = X - 1 \\
&= \sum_{k=0}^{w} \binom{r + k - 1}{k} p^r (1 - p)^k \cdot p(1 - p)^{w-k} \\
&= \sum_{k=0}^{w} \binom{r + k - 1}{k} p^{r+1} (1 - p)^w \\
&= p^{r+1}(1 - p)^w \sum_{k=0}^{w} \binom{r + k - 1}{k} \\
&= \binom{r + w}{w} p^{r+1} (1 - p)^w && \text{Using "Hockey Stick" with } n = r - 1, k = w + 1.
\end{aligned}$$

This shows that W is the Z distribution with $r_W = r + 1$ and probability of success as p, $Z(r + 1, p)$ and it is now evident that if $X_i' = X_i - 1$ where $X \sim$ Geometric(p) for $i = 1, 2, \ldots, n$ are IID, then $\sum_i^r X_i' = Z(r, p)$
The distribution Z has a name, it's called the Negative Binomial Distribution, just a fun fact.

(d) What is the expected value of total unsuccessful attempts of eating the whole bowl of R noodles, the random variable Z?.

Solution: From part c) we proved that Z is the sum of r independent X' random variables. Then

$$\begin{aligned}
Z &= rX' \\
E[Z] &= E[rX'] \\
&= rE[X'] && \text{Using Linearity of expectation} \\
&= rE[X - 1] && \text{Using the substitution } X' = X - 1 \\
&= r(E[X] - 1) \\
&= \frac{r}{p} - r && \text{As } X \sim \text{Geometric}(p)\text{, then } E[X] = 1/p \\
&= \frac{r(1 - p)}{p}
\end{aligned}$$

2 Linear Least-Squares Estimation

5. **Linear Least Squares Estimate: Derivation**
The LLSE of Y given X, denoted $L[Y|X]$, is the linear estimator $\hat{Y} = g(X) = a + bX$ that minimizes least-squares error:

$$C(g) = E(|Y - g(X)|^2) = E(|Y - a - bX|^2).$$

It turns out $L[Y|X] = E(Y) + \frac{cov(X,Y)}{var(X)}(X - E(X))$.
Let's try to derive this.

(a) Write $C(g)$ as linear function of $E(Y^2)$, $E(X^2)$, $E(Y)$, $E(X)$ and $E(XY)$

Solution:

$$\begin{aligned} C(g) &= E(|Y - a - bX|^2) = E((Y - a - bX)(Y - a - bX)) \\ &= E(Y^2 + a^2 + b^2X^2 - 2aY - 2bYX + 2abX) \\ &= E(Y^2) + a^2 + b^2E(X^2) - 2aE(Y) - 2bE(YX) + 2abE(X). \end{aligned}$$

(b) Using calculus, find the values of a and b that minimize the expression in part a. To simplify the calculation use

$$Cov(X,Y) = E(YX) - E(Y)E(X) \text{ and } Var(X) = E(X^2) - E(X)^2.$$

Solution: To find the values of a and b that minimize that expression, we set to zero the partial derivatives with respect to a and b. This gives the following two equations:

$$\begin{aligned} 0 &= \frac{d}{da}C(g) = 2a - 2E(Y) + 2bE(X) \\ 0 &= \frac{d}{da}C(g) = 2bE(X^2) - 2E(YX) + 2aE(X) \end{aligned}$$

From equation 1 we can find a

$$a = E(Y) - bE(X)$$

Substituting that in equation 2 gives

$$\begin{aligned} 0 &= 2bE(X^2) - 2E(YX) + 2aE(X) \\ &= 2bE(X^2) - 2E(YX) + 2[E(Y) - bE(X)]E(X) \\ &= bE(X^2) - E(YX) + E(Y)E(X) - bE(X)^2 \\ &= b[E(X^2) - E(X)^2] - [E(YX) - E(Y)E(X)] \\ &= bVar(X) - Cov(X,Y) \\ b &= \frac{Cov(X,Y)}{Var(X)} \end{aligned}$$

So finally we get values a and b as

$$\begin{aligned} a &= E(Y) - \frac{Cov(X,Y)}{Var(X)}E(X) \\ b &= \frac{Cov(X,Y)}{Var(X)} \end{aligned}$$

(c) Construct $L[Y|X]$ using the values you found for a and b.

Solution:

$$\begin{aligned} L[Y|X] &= a + bX \\ &= E(Y) - \frac{Cov(X,Y)}{Var(X)}E(X) + \frac{Cov(X,Y)}{Var(X)}X \\ &= E(Y) + \frac{cov(X,Y)}{var(X)}(X - E(X)) \end{aligned}$$

6. **Linear Least Squares Estimate: Projection**

It turns out there is an alternate derivation of the LLSE from a geometric perspective:
$L[Y|X]$ is the projection of Y onto the space of linear functions of X.

Without delving into the linear algebra, $L[Y|X]$ is the linear estimator $\hat{Y} = g(X) = a + bX$ that satisfies

$$E((Y - g(X)(c + dX)) = E((Y - a - bX)(c - dX)) = 0 \qquad \forall c, d \in \mathbb{R}$$

Use this property to show the $L[Y|X]$ minimizes $C(g)$. (Hint: Show $E(|Y - h(X)|^2) \geq E(|Y - L[Y|X]|^2)$ for any linear $h(X)$.) This picture may help.

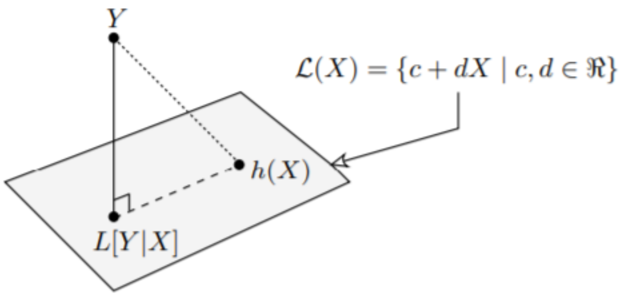

Figure 5: $L[Y|X]$ is the projection of Y onto $\mathscr{L}(X)$.

Solution: $E[|Y - h(X)|^2]$ corresponds to the squared length hypotenuse of the triangle in the diagram. We'll attempt to decompose the squared length of the hypotenuse into the sum of squared side lengths (Pythagorean Theorem).

To use the projection property, we need a $|Y - L[Y|X]|^2$ term, so let's get one.

$$\begin{aligned} E(|Y - h(X)|^2) &= E(|Y - (L[Y|X] - L[Y|X]) - h(X)|^2) \\ &= E(|(Y - L[Y|X]) + (L[Y|X] - h(X))|^2) \\ &= E(|Y - L[Y|X]|^2) + E(|L[Y|X] - h(X)|^2) + 2E((Y - L[Y|X])(L[Y|X] - h(X))) \end{aligned}$$

Let's look at the third term more closely. It looks suspiciously similar to our projection property, except the function multiplying $(Y - L[Y|X])$ is $L[Y|X] - h(X)$ not $c + dX$. But recall both $L[Y|X]$ and $h(X)$ are linear functions of the form $a + bX$!

Here's a neat fact: the difference of 2 linear functions are linear.

$$(a_1 + b_1X) - (a_2 + b_2X) = (a_1 - a_2) + (b_1 - b_2)X$$

So, by the projection property the third term is 0.

$$E(|Y - h(X)|^2) = E(|Y - L[Y|X]|^2) + E(|L[Y|X] - h(X)|^2)$$

The second term is a square so it can never be negative. Thus,

$$E(|Y - h(X)|^2) \geq E(|Y - L[Y|X]|^2)$$