Chris Piech
CS 109

Section #5
Oct 30, 2025

Section 5

1 Beta Sum Warmup

What is the distribution of the sum of 100 IID Betas? Let X be the sum:

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

Note the expectation and variance of a Beta:

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

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:

$$\begin{aligned} E(X_i) &= \frac{a}{a+b} && \text{Expectation of a Beta} \\ &= \frac{3}{7} \approx 0.43 \\ \text{Var}(X_i) &= \frac{ab}{(a+b)^2(a+b+1)} && \text{Variance of a Beta} \\ &= \frac{3 \cdot 4}{(3+4)^2(3+4+1)} \\ &= \frac{12}{49 \cdot 8} \approx 0.03 \end{aligned}$$

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

2 Microcontroller Precision Error

A low-powered microcontroller can only represent numbers to 3 decimal places.

Let X_i be an i.i.d. sample from Uniform(0, 1) with infinite precision. Let Y_i be the representation of X_i on the microcontroller. Y_i is equal to X_i truncated to 3 decimal places. Here are three examples:

X_i	Y_i
$0.12340809\ldots$	0.123
$0.28374110\ldots$	0.283
$0.55555555\ldots$	0.555

a (5 points) What is the distribution of $(X_i - Y_i)$? In other words, what is the distribution for the error when a single uniformly sampled value is represented in the microcontroller?

Solution: For each $X_i \in [0, 1]$, the rounding error is uniformly distributed between 0 and 0.001, so:

$$(X_i - Y_i) \sim \text{Unif}(0, 0.001)$$

b (10 points) We sample 1000 values, where X_i is the precise value and Y_i is the truncated value of X_i:

$$X = \sum_{i=1}^{1000} X_i \qquad Y = \sum_{i=1}^{1000} Y_i$$

What is the probability that the difference between X and Y is more than 0.51?

Solution: Apply Central Limit Theorem (CLT) since we are observing the distribution of the sum of 1000 IID samples. First, calculate the mean error and variance per sample (from the uniform distribution found in part a):

$$E[X_i - Y_i] = \frac{1}{2}(0.001) = 0.0005$$

$$Var(X_i - Y_i) = \frac{1}{12}(0.001)^2$$

We cannot apply a linear transformation between individual distributions for X and Y since Y_i is discrete and we dont have have a distribution for it. Instead, let $X - Y = \sum_{i=1}^{1000}(X_i - Y_i)$ and apply CLT to approximate $(X - Y)$ as a normal distribution:

$$(X - Y) \sim \mathcal{N}(\mu = 1000 * 0.0005, \sigma^2 = 1000 * \frac{1}{12}(0.001)^2)$$

Finally, solve for the final probability with the normal CDF:

$$P(X - Y \geq 0.51) = 1 - P(X - Y \leq 0.51) = 1 - \Phi\left(\frac{0.51 - 0.5}{\sqrt{\frac{1000}{12}(0.001)^2}}\right) = 1 - \Phi\left(\frac{0.01}{\sqrt{\frac{1000}{12}}(0.001)}\right)$$

Note: We accept both a one-tailed or two-tailed solution. However, since the event of interest is not symmetric around the mean, we cannot simply multiply the above probability by 2 for the two-tailed solution. Instead, we need to separately calculate $P(X - Y \leq -0.51)$, as well. The two tailed solution is:

$$P(|X - Y| \geq 0.51) = (1 - P(X - Y < 0.51)) + P(X - Y < -0.51)$$

3 Song of the Quarter

This quarter in CS109 there were 167 songs that were voted on. For each song, we have a list of votes where each vote is an integer in the set $\{1, 2, 3, 4, 5\}$. We assume all votes for a song are IID samples from the "true" distribution of CS109 opinion on the song.

For each song i we have m_i votes stored in a list `votes[i]` = $[x_1, x_2, \ldots, x_m]$. We have already calculated:

$$\mu_i = \frac{1}{m_i}\sum_{j=1}^{m_i} x_j \qquad \text{using } \texttt{np.mean(votes[i])}$$

$$\text{var}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} (x_j - \mu_i)^2 \qquad \text{using } \texttt{np.var(votes[i])}$$

$$\text{svar}_i = \frac{1}{m_i - 1}\sum_{j=1}^{m_i} (x_j - \mu_i)^2 \qquad \text{using } \texttt{np.var(votes[i], ddof=1)}$$

a (7 points) Song 1 has $m_1 = 45$ votes. We have calculated:

$$\mu_1 = 3.82 \qquad \text{var}_1 = 1.4 \qquad \text{svar}_1 = 1.5$$

Estimate the probability that the true average rating for song 1 is less than 3.

We can model the average rating of a song using the Central Limit Theorem, since each vote is IID. Let $\bar{X}_1$ be the average rating for song 1 from our collected votes.

$$\bar{X}_1 \sim \mathcal{N}(\mu_1, \frac{S^2}{n})$$

In this case, our best estimate of the mean is $\mu_1 = 3.82$, our sample variance is $S^2 = \text{svar}_1 = 1.5$, and $n = m_1 = 45$. Then, using the CDF of the Normal, the probability we want is:

$$P(\bar{X}_1 < 3) = \Phi\left(\frac{3 - 3.82}{\sqrt{\frac{1.5}{45}}}\right)$$

b (8 points) Song 1 has $m_1 = 45$ votes. Song 2 has $m_2 = 36$ votes. We have calculated:

Song 1:	$\mu_1 = 3.82$	$\text{var}_1 = 1.4$	$\text{svar}_1 = 1.5$
Song 2:	$\mu_2 = 3.79$	$\text{var}_2 = 1.7$	$\text{svar}_2 = 1.8$

What is the probability that the true average of Song 1 is greater than the true average for Song 2?

We again use the CLT to determine distributions of sample means:

$$\bar{X}_1 \sim \mathcal{N}(\mu_1, \frac{\text{svar}_1}{m_1}) \qquad \bar{X}_2 \sim \mathcal{N}(\mu_2, \frac{\text{svar}_2}{m_2})$$

Then, to determine if the average for song 1 is greater than the average for song 2, we can recall that $a > b$ is equivalent to $a - b > 0$ and create a distribution for the difference in averages between the two songs.

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\text{svar}_1}{m_1} + \frac{\text{svar}_2}{m_2})$$

$$P(\bar{X}_1 - \bar{X}_2 > 0) = 1 - P(\bar{X}_1 - \bar{X}_2 < 0)$$

$$= 1 - \Phi\left(\frac{0 - (3.82 - 3.79)}{\sqrt{\frac{1.5}{45} + \frac{1.8}{36}}}\right)$$