CS109 February 22, 2024

# Continuous Joint Distributions, Central Limit Theorem

Before you leave lab, make sure you click here so that you're marked as having attended this week's section. The CA leading your discussion section can enter the password needed once you've submitted.

## 1 Warmups

### *1.1 Sums of Random Variables*

For each of the problems below, assume that $X$ and $Y$ are independent.

1. Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. What is $\mu$ and $\sigma^2$ for $X + Y \sim \mathcal{N}(\mu, \sigma^2)$?
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 $X + Y$?

1. $\mu = \mu_1 + \mu_2$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2$. 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 Food for Thought*

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 $\frac{Y-6000}{316.212} = Z \sim \mathcal{N}(0, 1)$

$$P\left(Z < \frac{6500 - 6000}{316.212}\right)$$

$$P\left(Z < 1.58\right)$$

$$\Phi\left(1.58\right)$$

# 2 Problems

## *2.1 Grading Exams*

Jacob and Kathleen are planning to grade Problem 1 on your Week 7 exam, and they'll each grade their half independently of the other. Jacob takes $X \sim Exp(\frac{1}{3})$ hours to finish his half while Kathleen takes $Y \sim Exp(\frac{1}{4})$ hours to finish his half.

a. Find the CDF of $X/Y$, which is the ratio of their grading completion times.

The random variable of interest is the ratio $X/Y$, so the CDF, $F(r)$, in this case would be $P(\frac{X}{Y} < r)$, where $r$ stands for **r**atio and ranges from 0 to $\infty$. Rearranging, we are interested

in computing $P(X < rY)$, which can be computed in terms of the PDFs for $X$ and $Y$:

$$\begin{aligned}
F(r) = P(\frac{X}{Y} < r) &= P(X < rY) \\
&= \int_0^\infty \int_0^{ry} \frac{1}{12} e^{-\frac{1}{3}x} e^{-\frac{1}{4}y} dx dy \\
&= \frac{1}{12} \int_0^\infty e^{-\frac{1}{4}y} \int_0^{ry} e^{-\frac{1}{3}x} dx dy \\
&= -\frac{1}{4} \int_0^\infty e^{-\frac{1}{4}y} \left( e^{-\frac{1}{3}x} \right) \Big|_0^{ry} dy \\
&= -\frac{1}{4} \int_0^\infty e^{-\frac{1}{4}y} \left( e^{-\frac{1}{3}ry} - 1 \right) dy \\
&= \frac{1}{4} \int_0^\infty e^{-\frac{1}{4}y} dy - \frac{1}{4} \int_0^\infty e^{-(\frac{1}{3}r + \frac{1}{4})y} dy \\
&= 1 + \frac{\frac{1}{4}}{\frac{1}{3}r + \frac{1}{4}} e^{-(\frac{1}{3}r + \frac{1}{4})y} \Big|_0^\infty \\
&= 1 - \frac{\frac{1}{4}}{\frac{1}{3}r + \frac{1}{4}} = \frac{\frac{1}{3}r + \frac{1}{4}}{\frac{1}{3}r + \frac{1}{4}} - \frac{\frac{1}{4}}{\frac{1}{3}r + \frac{1}{4}} \\
&= \frac{\frac{1}{3}r}{\frac{1}{3}r + \frac{1}{4}}
\end{aligned}$$

For those question why that first of two integrals vanished to 1, note that the integrand is just the PDF of Expo($\lambda = \frac{1}{4}$)!

Incidentally, we can compute the probability density function from the CDF by differentiating with respect to $r$:

$$\begin{aligned}
f(r) &= \frac{dF(r)}{dr} \\
&= \frac{d}{dr} \frac{\frac{1}{3}r}{\frac{1}{3}r + \frac{1}{4}} \\
&= \frac{1}{12(\frac{1}{3}r + \frac{1}{4})^2}
\end{aligned}$$

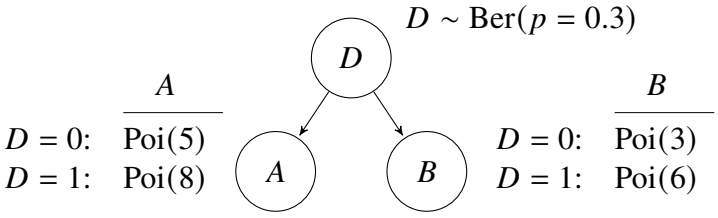
b. What is the probability that Kathleen finishes before Jacob does?

In comparison, that is delightfully straightforward, because we get to plug $r = 1$ into our result from part a. $P(X < Y) = \frac{1}{3} \cdot \frac{12}{7} = \frac{4}{7}$. That, however, is the probability that Jacob finishing before Kathleen, and we want to opposite. Therefore, the probability of interest

is really $\frac{3}{7}$. Given the expected completion times of 3 and 4 hours for Jacob and Kathleen, respectively, this seems right.

## 2.2 *Fish Sticks [courtesy of Lisa Yan]*

Fish Sticks, the online platform designed to meet all of your fish stick needs, wants to model their hourly homepage traffic from Stanford. The company decides to model two different behaviors for homepage visits according to the Bayesian Network on the right:




$A$ and $B$ are the numbers of Stanford students and faculty, respectively, who visit the Fish Sticks homepage in an hour. Since Fish Sticks does not know when Stanford people eat, the company models demand as a "hidden" Bernoulli random variable $D$, which determines the distribution of $A$ and $B$. Recall that in a Bayesian Network, random variables are conditionally independent given their parents. For example, given $D = 0$, $A \sim \text{Poi}(5)$ and $B \sim \text{Poi}(3)$, two independent random variables.

a. Given that 6 users from group $A$ visit the homepage in the next hour, what is the probability that $D = 0$?

Note that given $D = 0$, $A \sim \text{Poi}(\lambda = 5)$, and given $D = 1$, $A \sim \text{Poi}(\lambda = 8)$. By Bayes' Theorem,

$$\begin{aligned} P(D = 0|A = 6) &= \frac{P(A = 6|D = 0)P(D = 0)}{P(A = 6|D = 0)P(D = 0) + P(A = 6|D = 1)P(D = 1)} \\ &= \frac{\frac{5^6 e^{-5}}{6!}(1 - 0.3)}{\frac{5^6 e^{-5}}{6!}(1 - 0.3) + \frac{8^6 e^{-8}}{6!}(0.3)} \\ &= \frac{5^6 e^{-5}(1 - 0.3)}{5^6 e^{-5}(1 - 0.3) + 8^6 e^{-8}(0.3)} \approx 0.7364 \end{aligned}$$

b. What is the probability that in the next hour, the *total* number of users who visit the homepage from groups $A$ and $B$ is equal to 12, i.e., what is $P(A + B = 12)$?

By Law of Total Probability,

$$P(A + B = 12) = P(A + B = 12|D = 0)P(D = 0) + P(A + B = 12|D = 1)P(D = 1).$$

$A$ and $B$ are conditionally independent Poisson random variables given $D$, and therefore $A + B|D = 0 \sim \text{Poi}(\lambda = 8)$ and $A + B|D = 1 \sim \text{Poi}(\lambda = 14)$. Using the Poisson PMF,

$$P(A+B=12) = \frac{8^{12}e^{-8}}{12!} \cdot (1-0.3) + \frac{14^{12}e^{-14}}{12!} \cdot (0.3) \approx 0.0632.$$

c. Now simulate $P(A+B=\texttt{total})$, where $\texttt{total} = 12$, by implementing the `infer_prob_total(total, ntrials)` function below using rejection sampling.

- `total` is the total number of users from groups $A$ and $B$ in the event $A+B=\texttt{total}$.
- `ntrials` is the number of observations to generate for rejection sampling.
- `prob` is the return value to the function, where $\texttt{prob} \approx P(A+B=\texttt{total})$.
- The function call is implemented for you at the bottom of the code block.

You can call the following functions from the `scipy` package:

- `stats.bernoulli.rvs(`$p$`)`, which randomly generates a 1 with probability $p$, and generates a 0 otherwise.
- `stats.poisson.rvs(`$\lambda$`)`, which randomly generates a value according to a Poisson distribution with parameter $\lambda$

You are not required to use lists or `numpy` arrays in this question (but you can if you want). **Pseudo-code is fine** as long as your code accurately conveys your approach.

```
import numpy as np
from scipy import stats

def infer_prob_total(total, ntrials):
    # here's where your implementation belongs
    return prob

total = 12
ntrials = 50000
print('Simulated% P(A + B)=', infer_prob_total(total, ntrials))
```

This is the full implementation right here:

```
import numpy as np
from scipy import stats

def infer_prob_total(total, ntrials):
    n_samples_event = 0
    for i in range(ntrials):
       d = stats.bernoulli.rvs(0.3)
       if d == 0:
          user_sum = stats.poisson.rvs(5) + stats.poisson.rvs(3)
       else:
```

```
            user_sum = stats.poisson.rvs(8) + stats.poisson.rvs(6)
        if user_sum == total: n_samples_event += 1
    prob = n_samples_event/ntrials
    return prob

ntrials = 50000
total = 12
print("Simulated P(A + B)=", infer_prob_total(total, ntrials))
```

## 3 ChatGPT, Watermarking, and Bayesian Inference

ChatGPT is a generative AI technology that can be coarsely summarized to be a chat bot with a seemingly boundless ability to discuss any topic—history, computer science, art, nuclear physics, probability, and even the ethics of using ChatGPT—in any one of several written languages, including English, French, Spanish, C++, JavaScript, Python, and some 100 others.

Unsurprisingly, we will soon be left to wonder whether a poem, a Tweet, a C++ function, or a college thesis is written by ChatGPT or a human being. Questions about authorship, accuracy, and attribution have prompted OpenAI, the company behind ChatGPT, to address these concerns by implementing ChatGPT to employ what's termed **watermarking** and insert certain words more or less often than is customary in even the best of human-authored writing.

To illustrate, let's assume that most humans use the word **the** an average of 4.8 times per 100 words, whereas ChatGPT might generate prose where **the** appears on average 6.5 times per 100 words. Similarly, humans use the word **of** on average about 3.9 times per 100 words, whereas ChatGPT might leverage **of** about 6.2 times per 100 words. Conversely, ChatGPT might generate the word **by** only 1.6 times per 100 words, whereas humans use the word **by** about 2.7 times per 100 words.

a. You elect to model word frequency of all words using either a Poisson for paragraphs of 200 or so words—but as a Gaussian for larger documents—say 10000 words or more. Explain why the Gaussian might be the better choice for larger documents than the Poisson, whereas Poisson is more easily defended for smaller documents.

> The Poisson distribution provides an accurate estimate for documents with low word counts since we expect mostly low word counts with a long right tail for higher word counts.
>
> As our documents get larger, by the CLT we know that the Gaussian is an accurate, computationally efficient approximate of the word frequencies.
>
> Other answers we accepted were: discussions of variance, discussion of the requirements for binomial approximation, discussion of CLT, discussion of computation efficiency.

b. A deeper statistical analysis of many human-written documents strongly suggests that $H_{\text{the}} \sim \mathcal{N}(5, 1)$ and $H_{\text{by}} \sim \mathcal{N}(4, 1)$, whereas a separate but equally deep analysis strongly suggests that $C_{\text{the}} \sim \mathcal{N}(3, 1)$ and $C_{\text{by}} \sim \mathcal{N}(2, 1)$. (The parameters are rounded values for simplicity and assumed to be per-100-words.)

Assuming a prior belief that a very long document was written by a human is 0.99, what is your posterior belief that the document was written by a human if the document contains an average of 5 **the**'s every 100 words but only 1 **by** every 100 words. You may assume all Gaussian distributions of interest are independent.

We want to find our posterior belief:

$$f(H|5_{\text{the}}, 1_{\text{by}})$$

By Bayes:

$$= \frac{f(5_{\text{the}}, 1_{\text{by}}|H)P(H)}{f(5_{\text{the}}, 1_{\text{by}})}$$

By LOTP:

$$= \frac{f(5_{\text{the}}, 1_{\text{by}}|H)P(H)}{f(5_{\text{the}}, 1_{\text{by}}|H)P(H) + f(5_{\text{the}}, 1_{\text{by}}|C)P(C)}$$

By our independence, we can expand the joint PDF with a simple product

$$= \frac{0.99 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(0)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(-3)^2}}{0.99 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(0)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(-3)^2} + 0.01 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(2)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(-1)^2}}$$

$$= \frac{0.99 \cdot e^{-\frac{1}{2}(-3)^2}}{0.99 \cdot e^{-\frac{1}{2}(-3)^2} + 0.01 \cdot e^{-\frac{1}{2}(2)^2} \cdot e^{-\frac{1}{2}(-1)^2}}$$

$$= \frac{0.99 \cdot e^{-\frac{9}{2}}}{0.99 \cdot e^{-\frac{9}{2}} + 0.01 \cdot e^{-\frac{5}{2}}}$$

which simplifies to about 0.9305.

In reality, the Gaussians here are not independent, since the presence of one word implies the absence of another. Assuming a correlation value $\rho$ that is slightly negative, would you expect the observations stated in part b) to result in a larger posterior probability or a smaller one? Briefly explain why.

We anticipate a larger posterior. Negative correlation implies that an increased usage of "the" is correlated with a decreased usage of "by". This is aligned with what we observed, with "by" being used less than average. Thus, we believe that $f(5_{\text{the}}, 1_{\text{by}}) > f(5_{\text{the}})f(1_{\text{by}})$.