Announcements

• Pset grades release schedule
  • Pset 1 grades released earlier today
  • Pset 2 grades released by Monday’s class
  • PSet 3 grades released by Thursday’s class

• PSet 5 released
  • After today’s lecture, you should have everything you need to complete the entire pset!

• PSet 6 released on Monday
  • Can work on questions as we learn the material
  • No long extension for this pset!
Adding Random Variables
Imagine a game where each player independently scores 0 - 100 points:

• Let $X$ be the amount of points you score.
• Let $Y$ be the amount of points your opponent scores.
• Let’s say you know $P(X = x)$ and $P(Y = y)$.

What is the probability of a tie?

$$P(\text{tie}) = \sum_{i=0}^{100} P(X = i, Y = i) = \sum_{i=0}^{100} P(X = i)P(Y = i)$$
### Insight to Convolution

The problem is to evaluate $P(X + Y = n)$.

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$i$</th>
<th>$P(X=i, Y=n-i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$n$</td>
<td>0</td>
<td>$P(X = 0, Y = n)$</td>
</tr>
<tr>
<td>1</td>
<td>$n-1$</td>
<td>1</td>
<td>$P(X = 1, Y = n - 1)$</td>
</tr>
<tr>
<td>2</td>
<td>$n-2$</td>
<td>2</td>
<td>$P(X = 2, Y = n - 2)$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$n$</td>
<td>0</td>
<td>$n$</td>
<td>$P(X = n, Y = 0)$</td>
</tr>
</tbody>
</table>
**Insight to Convolution**

\[ P(X + Y = n) = \sum_{i=0}^{n} P(X = i, Y = n - i) \]

<table>
<thead>
<tr>
<th>( X )</th>
<th>( Y )</th>
<th>( i )</th>
<th>( P(X = 0, Y = n) )</th>
<th>( P(X = 1, Y = n - 1) )</th>
<th>( P(X = 2, Y = n - 2) )</th>
<th>( P(X = n, Y = 0) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( n )</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>( n - 1 )</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( n - 2 )</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n )</td>
<td>0</td>
<td>( n )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Insight to Convolution

For any discrete random variables $X$ and $Y$:

$$P(X + Y = n) = \sum_{i=0}^{n} P(X = i, Y = n - i)$$

In particular, for independent discrete random variables $X$ and $Y$:

$$P(X + Y = n) = \sum_{i=0}^{n} P(X = i)P(Y = n - i)$$
The distribution of a sum of 2 dice rolls is a convolution of 2 PMFs.

Example:

\[ P(X + Y = 4) = P(X = 1)P(Y = 3) + P(X = 2)P(Y = 2) + P(X = 3)P(Y = 1) \]
Convolution
**Sum of Independent Binomials**

\[ X \sim \text{Bin}(n_1, p) \]
\[ Y \sim \text{Bin}(n_2, p) \]
\[ X, Y \text{ independent} \]

\[ X + Y \sim \text{Bin}(n_1 + n_2, p) \]

**Holds in general case:**

\[ X_i \sim \text{Bin}(n_i, p) \]
\[ X_i \text{ independent for } i = 1, \ldots, n \]

\[ \sum_{i=1}^{n} X_i \sim \text{Bin} \left( \sum_{i=1}^{n} n_i, p \right) \]
Sum of Independent Poissons

\[ X \sim \text{Poi}(\lambda_1), Y \sim \text{Poi}(\lambda_2) \]
\[ X, Y \text{ independent} \]

\[ X + Y \sim \text{Poi}(\lambda_1 + \lambda_2) \]

Proof (just for reference):

\[
P(X + Y = n) = \sum_{k=0}^{n} P(X = k)P(Y = n - k)
\]

\[
= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} = e^{-(\lambda_1 + \lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k! (n-k)!}
\]

\[
= e^{-(\lambda_1 + \lambda_2)} \frac{n!}{n!} \sum_{k=0}^{n} \frac{n!}{k! (n-k)!} \lambda_1^k \lambda_2^{n-k} = e^{-(\lambda_1 + \lambda_2)} \frac{(\lambda_1 + \lambda_2)^n}{n!} = \text{Poi}(\lambda_1 + \lambda_2)
\]
Sum of Independent Poissons

\[ X \sim \text{Poi}(\lambda_1), \; Y \sim \text{Poi}(\lambda_2) \]
\[ X, \; Y \text{ independent} \]
\[ X + Y \sim \text{Poi}(\lambda_1 + \lambda_2) \]

Holds in general case:

\[ X_i \sim \text{Poi}(\lambda_i) \]
\[ X_i \text{ independent for } i = 1, \ldots, n \]
\[ \sum_{i=1}^{n} X_i \sim \text{Poi}(\sum_{i=1}^{n} \lambda_i) \]
Sum of Independent Gaussians

\[ X \sim \mathcal{N}(\mu_1, \sigma_1^2), \ Y \sim \mathcal{N}(\mu_2, \sigma_2^2) \]

\[ X, Y \text{ independent} \]

\[ X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \]

Holds in general case:

\[ X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \]

\[ X_i \text{ independent for } i = 1, \ldots, n \]

\[ \left( \sum_{i=1}^{n} X_i \right) \sim N \left( \sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2 \right) \]
Virus Infections

Say you are working to plan a response to a virus. There are two exposed groups

• P1: 50 people, each independently infected with $p = 0.1$
• P2: 100 people, each independently infected with $p = 0.4$

What is the probability of more than 40 infections?
Let’s say that $X \sim \mathcal{N} (\mu, \sigma^2)$.
We want to calculate the distribution of $Y$, where $Y = X + X = 2X$.

**Linear Transform**

$Y \sim \mathcal{N} (a\mu, a^2\sigma^2)$

$Y \sim \mathcal{N} (2\mu, 4\sigma^2)$

**Sum of Independent Normals**

$Y \sim \mathcal{N} (\mu + \mu, \sigma^2 + \sigma^2)$

$Y \sim \mathcal{N} (2\mu, 2\sigma^2)$
ELO ratings

Each team has an ELO score $S$, calculated based on its past performance.

- Each game, a team has ability $A \sim \mathcal{N}(S, 200^2)$.
- The team with the higher sampled ability wins.

What is the probability that Warriors win this game?

Want:

$P(Warriors \; win) = P(A_W > A_O)$
Probability of Winning

\[ A_W \sim \mathcal{N}(S = 1657, 200^2) \]
\[ A_O \sim \mathcal{N}(S = 1470, 200^2) \]

\[ P(\text{Warriors win}) = P(A_W > A_O) \]

\[ P(\text{Warriors win}) = P(A_W - A_O > 0) \]
Discrete vs. Continuous

Discrete:

\[ P(X + Y = a) = \sum_{y=-\infty}^{\infty} P(X = a - y, Y = y) \]

Continuous:

\[ P(X + Y = a) = \int_{y=-\infty}^{\infty} f(X = a - y, Y = y) \, dy \]
Sum of Independent Uniforms

\[ X \sim \text{Uni}(0,1), \ Y \sim \text{Uni}(0,1) \]

\[ X, Y \text{ independent} \]

\[ f(X + Y = a) = \int_{y=-\infty}^{\infty} f(X = a - y)f(Y = y) \, dy \]

\[ f(X + Y = a) = \begin{cases} 
  a & 0 < a < 1 \\
  2 - a & 1 < a < 2 \\
  0 & \text{otherwise}
\end{cases} \]
Central Limit Theorem

Consider \( n \) independent and identically distributed (i.i.d.) variables \( X_1, X_2, \ldots, X_n \) with \( E[X_i] = \mu \) and \( \text{Var}(X_i) = \sigma^2 \).

\[
\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)
\]

As \( n \to \infty \)

The sum of \( n \) i.i.d. random variables is normally distributed with mean \( n\mu \) and variance \( n\sigma^2 \).
Roll $n$ independent dice. Let $X_i$ be the outcome of roll $i$. $X_i$ are i.i.d.

- Sum of 1 die roll: $\sum_{i=1}^{1} X_i$
- Sum of 2 dice rolls: $\sum_{i=1}^{2} X_i$
- Sum of 3 dice rolls: $\sum_{i=1}^{3} X_i$
CLT explains a lot

As $n \to \infty$

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

The sum of $n$ i.i.d. random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$. 

$n = 5$
CLT explains a lot

The sum of $n$ i.i.d. random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.

\[ \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \]

Proof:

Let $X_i \sim \text{Ber}(p)$ for $i = 1, \ldots, n$, where $X_i$ are i.i.d. $E[X_i] = p$, $\text{Var}(X_i) = p(1-p)$

\[ X = \sum_{i=1}^{n} X_i \]

$X \sim \mathcal{N}(n\mu, n\sigma^2)$ (CLT, as $n \to \infty$)

$X \sim \mathcal{N}(np, np(1-p))$ (substitute mean, variance of Bernoulli)
CLT explains a lot

\[ \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \]

As \( n \to \infty \)

The sum of \( n \) i.i.d. random variables is normally distributed with mean \( n\mu \) and variance \( n\sigma^2 \).

Sample of size 15, sum values

Distribution of \( X_i \)

Distribution of \( \sum_{i=1}^{15} X_i \)
CLT explains a lot

\[
\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)
\]

As \( n \to \infty \)

The sum of \( n \) i.i.d. random variables is normally distributed with mean \( n\mu \) and variance \( n\sigma^2 \).

Sample of size 15, average values

Distribution of \( X_i \) (sample mean)

Distribution of \( \frac{1}{15} \sum_{i=1}^{15} X_i \)
Traffic Light

You hit 10 traffic lights on your way to work. You don't know the full distribution of the wait time, but for each you observe the average wait time is 45 seconds and the standard deviation is 5 seconds. You will be on time if your total wait time is less than 8 mins. What is the probability that you are on time? Assume the wait times are IID.
Proof of CLT

\[ \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \]

As \( n \to \infty \)

The sum of \( n \) i.i.d. random variables is normally distributed with mean \( n\mu \) and variance \( n\sigma^2 \).

Proof:

- The Fourier Transform of a PDF is called a **characteristic function**.
- Take the characteristic function of the probability mass of the sample distance from the mean, divided by standard deviation
- Show that this approaches an exponential function in the limit as \( n \to \infty \):
  \[ f(x) = e^{-\frac{x^2}{2}} \]
- This function is in turn the characteristic function of the Standard Normal, \( Z \sim \mathcal{N}(0,1) \).

(this proof is beyond the scope of CS109)
Sum of \( n \) independent Uniform RVs

Let \( X = \sum_{i=1}^{n} X_i \) be sum of i.i.d. RVs, where \( X_i \sim \text{Uni}(0,1) \).

\[ \mu = E[X_i] = 1/2 \]
\[ \sigma^2 = \text{Var}(X_i) = 1/12 \]

For different \( n \), how close is the CLT approximation of \( P(X \leq n/3) \)?

\( n = 2 \):

Exact \( \quad P(X \leq 2/3) \approx 0.2222 \)

CLT approximation

\[ X \approx Y \sim \mathcal{N}(n\mu, n\sigma^2) \quad \Rightarrow Y \sim \mathcal{N}(1, 1/6) \]

\[ P(X \leq 2/3) \approx P(Y \leq 2/3) \]

\[ = \Phi \left( \frac{2/3 - 1}{\sqrt{1/6}} \right) \approx 0.2071 \]
Sum of \( n \) independent Uniform RVs

Let \( X = \sum_{i=1}^{n} X_i \) be sum of i.i.d. RVs, where \( X_i \sim \text{Uni}(0,1) \).

\[
\mu = E[X_i] = \frac{1}{2} \\
\sigma^2 = \text{Var}(X_i) = \frac{1}{12}
\]

For different \( n \), how close is the CLT approximation of \( P(X \leq n/3) \)?

\( n = 5 \):

**Exact** \( P(X \leq 5/3) \approx 0.1017 \)

**CLT approximation**

\[
X \approx Y \sim \mathcal{N}(n\mu, n\sigma^2) \quad \Rightarrow \quad Y \sim \mathcal{N}(5/2, 5/12)
\]

\[
P(X \leq 5/3) \approx P(Y \leq 5/3)
\]

\[
= \Phi \left( \frac{5/3 - 5/2}{\sqrt{5/12}} \right) \approx 0.0984
\]
Sum of \( n \) independent Uniform RVs

Let \( X = \sum_{i=1}^{n} X_i \) be sum of i.i.d. RVs, where \( X_i \sim \text{Uni}(0,1) \).

\[ \mu = E[X_i] = \frac{1}{2} \]

\[ \sigma^2 = \text{Var}(X_i) = \frac{1}{12} \]

For different \( n \), how close is the CLT approximation of \( P(X \leq n/3) \)?

\( n = 10 \):

Exact \( P(X \leq 10/3) \approx 0.0337 \)

CLT approximation

\( X \approx Y \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow Y \sim \mathcal{N}(5, 5/6) \)

\[ P(X \leq 10/3) \approx P(Y \leq 10/3) \]

\[ = \Phi\left(\frac{10/3 - 5}{\sqrt{5/6}}\right) \approx 0.0339 \]
Sum of $n$ independent Uniform RVs

Let $X = \sum_{i=1}^{n} X_i$ be sum of i.i.d. RVs, where $X_i \sim \text{Uni}(0,1)$. 

For different $n$, how close is the CLT approximation of $P(X \leq n/3)$?

$n = 2$: 

$n = 5$: 

$n = 10$: 

Most books will tell you that CLT holds if $n \geq 30$, but it can hold for smaller $n$ depending on the distribution of your i.i.d. $X_i$’s.
What about other functions?

Let $X_1, X_2, ..., X_n$ be i.i.d., where $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$. As $n \to \infty$:

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

Sum of i.i.d. RVs

Average of i.i.d. RVs (sample mean)

Max of i.i.d. RVs
Distribution of sample mean

Let $X_1, X_2, \ldots, X_n$ be i.i.d., where $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$. As $n \to \infty$:

Define:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{(sample mean)} \quad Y = \sum_{i=1}^{n} X_i \quad \text{(sum)}$$

The average of i.i.d. random variables (i.e., sample mean) is normally distributed with mean $\mu$ and variance $\sigma^2 / n$.

What about other functions?

Let $X_1, X_2, \ldots, X_n$ be i.i.d., where $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$. As $n \to \infty$:

\[
\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)
\]

Sum of i.i.d. RVs

\[
\frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})
\]

Average of i.i.d. RVs (sample mean)

Gumbel

Max of i.i.d. RVs

(see Fisher-Tippett Gnedenko Theorem)
Dice game

You will roll 10 6-sided dice \((X_1, X_2, \ldots, X_{10})\).
- Let \(X = X_1 + X_2 + \cdots + X_{10}\), the total value of all 10 rolls.
- You win if \(X \leq 25\) or \(X \geq 45\).

What is the probability of winning?

\[
E[X_i] = 3.5, \\
\text{Var}(X_i) = 35/12
\]