Updated 10/31/2025

Part 3: Applications
L15: Maximum Entropy Principle
[Deriving the Maximum entropy principle]

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa25)

https://northeastern-datalab.github.io/cs7840/fa25/

10/30/2025

Pre-class conversations

- Last class recapitulation
- Projects: I added comments for all projects. Please talk to me often (just today I cannot meet after class).
- Idea for class contributions: how to find errors in my slides...
- Slide decks are being re-organized (new ideas and connections from our in-class discussions). Page numbers will likely change. Also new examples in the Python workbooks ☺
  - please submit yours too with scribes

- Today:
  - Why Max entropy? Involves just combinatorics, and limits, no "uncertainty"
  - Why Occam's razor?

A quick primer on Combinatorics & the Multinomial Distribution

(in preparation for our discussion of Wallis' argument for Max Entropy)

Permutations, Combinations, Binomial coefficient

Permutations

Given $n = 4$ objects $\{A, B, C, D\}$. There are how many permutations:
$ABCD, ABDC, ACBD, ACBD, \ldots, DCBA$

?

Permutations, Combinations, Binomial coefficient

Permutations

Given $n = 4$ objects $\{A, B, C, D\}$. There are $n! = 24$ different permutations:
$ABCD, ABDC, ACBD, ACBD, \dots, DCBA$

k-permutations (partial permutations)

There are how may different permutations of size $k = 2$:
$AB, AC, AD, BA, \dots DC$

?

Permutations, Combinations, Binomial coefficient

Permutations

Given $n = 4$ objects $\{A, B, C, D\}$. There are $n! = 24$ different permutations:
$ABCD, ABDC, ACBD, ACBD, \ldots, DCBA$

k-permutations (partial permutations)

There are $P(n,k) = \frac{n!}{(n-k)!} = n^{\underline{k}} = 12$ different permutations of size $k = 2$:
$AB, AC, AD, BA, \ldots DC$

INTUITION 1: We have n choices for the 1st, $n-1$ for the 2nd, ..., $(n-k+1)$ for the kth. Thus $n^{\underline{k}}$.
($n^{\underline{k}}$ is called the "falling factorial")

INTUITION 2: We don't distinguish between permutations of the items not shown:
$AB(CD) = AB(DC)$. Thus we divide by the number of such permutations $(n-k)! = 2$

k-combinations

There are how many different combinations (subsets) of size $k = 2$:
$\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}$

?

Permutations, Combinations, Binomial coefficient

Permutations

Given $n = 4$ objects $\{A, B, C, D\}$. There are $n! = 24$ different permutations:
$ABCD, ABDC, ACBD, ACBD, \ldots, DCBA$

k-permutations (partial permutations)

There are $P(n,k) = \frac{n!}{(n-k)!} = n^{\underline{k}} = 12$ different permutations of size $k = 2$:
$AB, AC, AD, BA, \ldots DC$

INTUITION 1: We have n choices for the 1st, $n-1$ for the 2nd, ..., $(n-k+1)$ for the k^{th}. Thus $n^{\underline{k}}$. ($n^{\underline{k}}$ is called the "falling factorial")

INTUITION 2: We don't distinguish between permutations of the items not shown: $AB(CD) = AB(DC)$. Thus we divide by the number of such permutations $(n-k)! = 2$

k-combinations

There are $C(n,k) = \frac{P(n,k)}{P(k,k)} = \frac{n^{\underline{k}}}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k} =$ 6 different combinations (subsets) of size $k = 2$:
$\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}$

INTUITION: We don't distinguish between permutations of the items shown: $AB = BA$. Thus we divide by the number of such permutations $k!$ This leads to the binomial coefficient.

multinomial outcomes

There are how many ways to partition the set into disjoint subsets of sizes $k_1 = 2$, $k_2 = 1$, $k_3 = 1$ with $\sum_i k_i = n$. ?
$\{AB|C|D\}, \{AB|D|C\}$, $\{AC|B|C\}$, ... $\{CD|B|A\}$

Permutations, Combinations, Binomial coefficient

Permutations

Given $n = 4$ objects $\{A, B, C, D\}$. There are $n! = 24$ different permutations: $ABCD, ABDC, ACBD, ACBD, \dots, DCBA$

k-permutations (partial permutations)

There are $P(n, k) = \frac{n!}{(n-k)!} = n^{\underline{k}} = 12$ different permutations of size $k = 2$: $AB, AC, AD, BA, \dots DC$

INTUITION 1: We have n choices for the 1st, $n-1$ for the 2nd, ..., $(n-k+1)$ for the k^{th}. Thus $n^{\underline{k}}$. ($n^{\underline{k}}$ is called the "falling factorial")

INTUITION 2: We don't distinguish between permutations of the items not shown: $AB(CD) = AB(DC)$. Thus we divide by the number of such permutations $(n-k)! = 2$

k-combinations

There are $C(n, k) = \frac{P(n,k)}{P(k,k)} = \frac{n^{\underline{k}}}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k} =$ 6 different combinations (subsets) of size $k = 2$: $\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}$

INTUITION: We don't distinguish between permutations of the items shown: $AB = BA$. Thus we divide by the number of such permutations $k!$ This leads to the binomial coefficient.

multinomial outcomes

There are $\binom{n}{k_1, k_2, k_3} = \frac{n!}{k_1! k_2! k_3!} = 12$ different ways to partition the set into disjoint subsets of sizes $k_1 = 2$, $k_2 = 1$, $k_3 = 1$ with $\sum_i k_i = n$.
$\{AB|C|D\}, \{AB|D|C\}$, $\{AC|B|C\}$, ... $\{CD|B|A\}$

INTUITION: We don't distinguish between permutations within each group. Thus we divide by the size of the equivalence class, i.e. $k_i!$ permutations for each group. That leads to the multinomial coefficient.

Permutations, Combinations, Binomial coefficient

Binomial Coefficient:
The number of distinct subsets of size k that can be chosen from a set of n elements.

Special case of multinomial coefficient: We partition the set into 2 groups, those that are in, and those that are not in the selection.

k-combinations

IN OUT

There are $C(n,k) = \frac{P(n,k)}{P(k,k)} = \frac{n^{\underline{k}}}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k} =$
6 different combinations (subsets) of size $k = 2$:
$\{A,B\},\{A,C\},\{A,D\},\{B,C\},\{B,D\},\{C,D\}$

Intuition: We don't distinguish between permutations of the items shown: $AB = BA$. Thus we divide by the number of such permutations k! This leads to the binomial coefficient.

Multinomial Coefficient:
A generalization of the binomial coefficient that calculates the number of ways to divide a set of n distinct elements into m distinct groups, where each group i contains a specified number of objects k_i, s.t. $\sum_i k_i = n$.

multinomial outcomes

There are $\binom{n}{k_1,k_2,k_3} = \frac{n!}{k_1!k_2!k_3!} = 12$ different ways to partition the set into disjoint subsets of sizes $k_1 = 2$, $k_2 = 1$, $k_3 = 1$ with $\sum_i k_i = n$.
$\{AB|C|D\},\{AB|D|C\}, \{AC|B|C\}, \ldots \{CD|B|A\}$

Intuition: We don't distinguish between permutations within each group. Thus we divide by the size of the equivalence class, i.e. $k_i!$ permutations for each group. That leads to the multinomial coefficient.

Binomial & Multinomial distribution

Binomial theorem (or Binomial expansion)

?

Binomial & Multinomial distribution

BACKUP

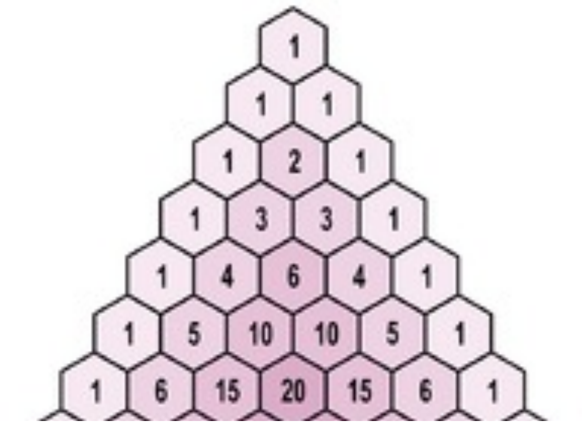
Binomial theorem (or Binomial expansion)

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot a^{n-k} b^k$$

Binomial coefficient $\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} = \frac{n^{\underline{k}}}{k!}$

Number of ways in which you can select k items from a total of n different items

$$(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$



Multinomial theorem (here, for $m = 3$)

?

Figure source: https://study.com/academy/lesson/binomial-coefficient-formula-examples.html

Binomial & Multinomial distribution

BACKUP

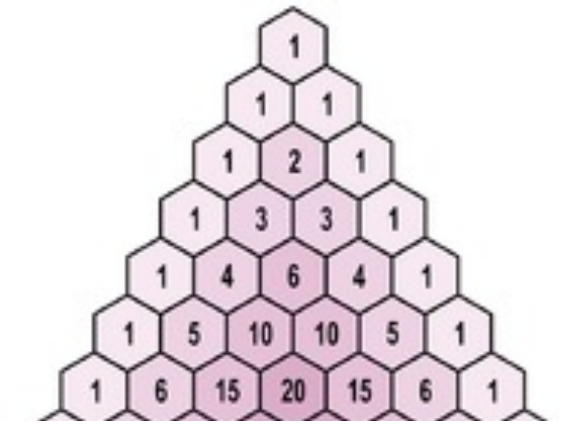
Binomial theorem (or Binomial expansion)

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot a^{n-k} b^k$$

Binomial coefficient $\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} = \frac{n^{\underline{k}}}{k!}$

Number of ways in which you can select k items from a total of n different items

$$(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$



Multinomial theorem (here, for $m = 3$)

$$(a+b+c)^n = \sum_{k_1+k_2+k_3=n} \binom{n}{k_1, k_2, k_3} a^{k_1} b^{k_2} c^{k_3}$$

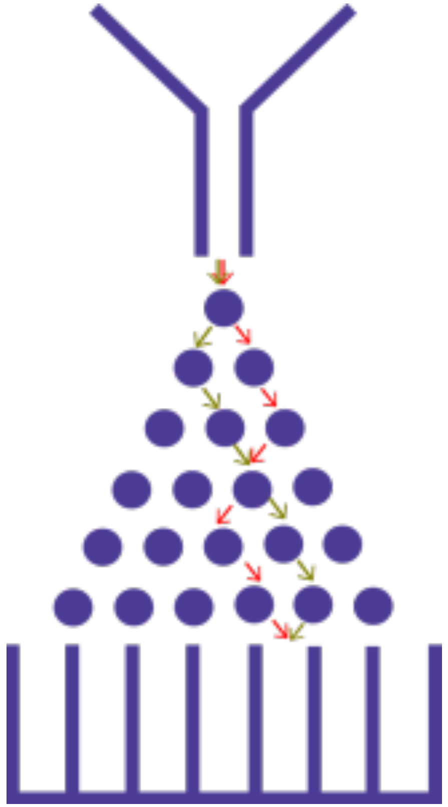
Multinomial coefficient $\binom{n}{k_1, k_2, k_3} = \frac{n!}{k_1! k_2! k_3!}$

Number of ways in which to partition an n-element set into disjoint subsets of sizes k_1, k_2, k_3 w/ $\sum_i k_i = n$.

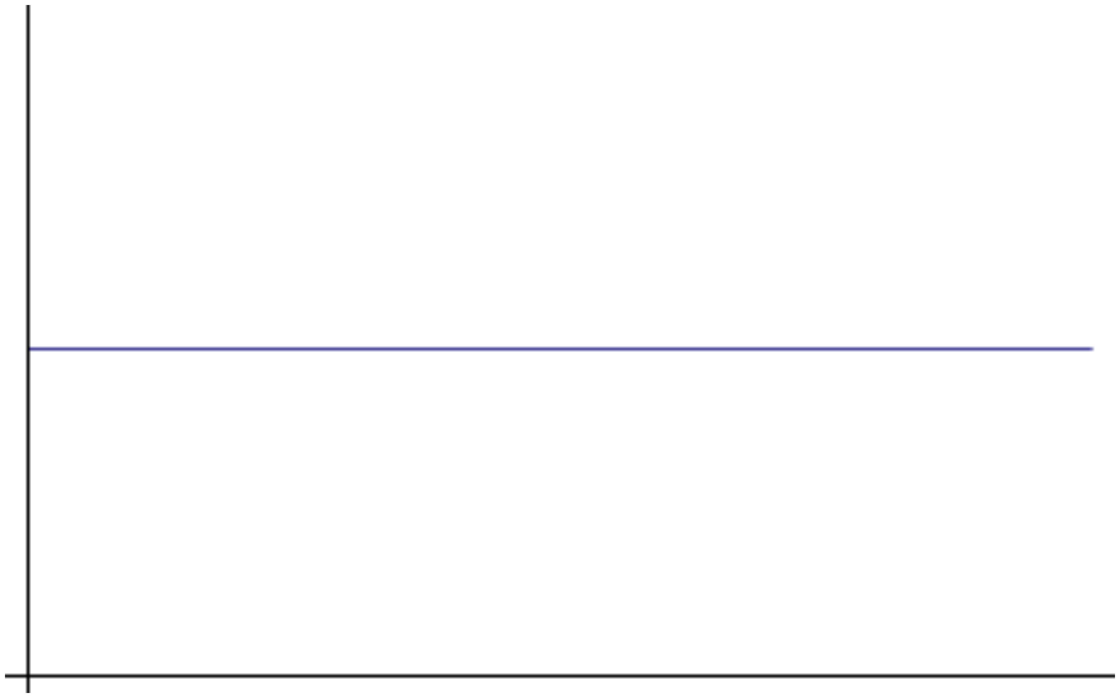
$$\begin{aligned}(a+b+c)^4 = {} & a^4 + b^4 + c^4 \\ & +4a^3b + 4a^3c + 4b^3a + 4b^3c + 4c^3a + 4c^3b \\ & +6a^2b^2 + 6a^2c^2 + 6b^2c^2 \\ & +12a^2bc + 12ab^2c + 12abc^2\end{aligned}$$

Figure source: https://study.com/academy/lesson/binomial-coefficient-formula-examples.html

Binomial distribution towards Normal distribution



"Two possible paths leading to the same bin within the bean machine."



"This animation captures the way a binomial distribution with increasing n will begin to look like a normal distribution."

Likely for $p \approx 0.5$, yet cut-off on the right.

Figure source: https://www.thephysicsmill.com/2014/04/13/probability-part-2-distributions/

Binomial distribution towards Normal distribution

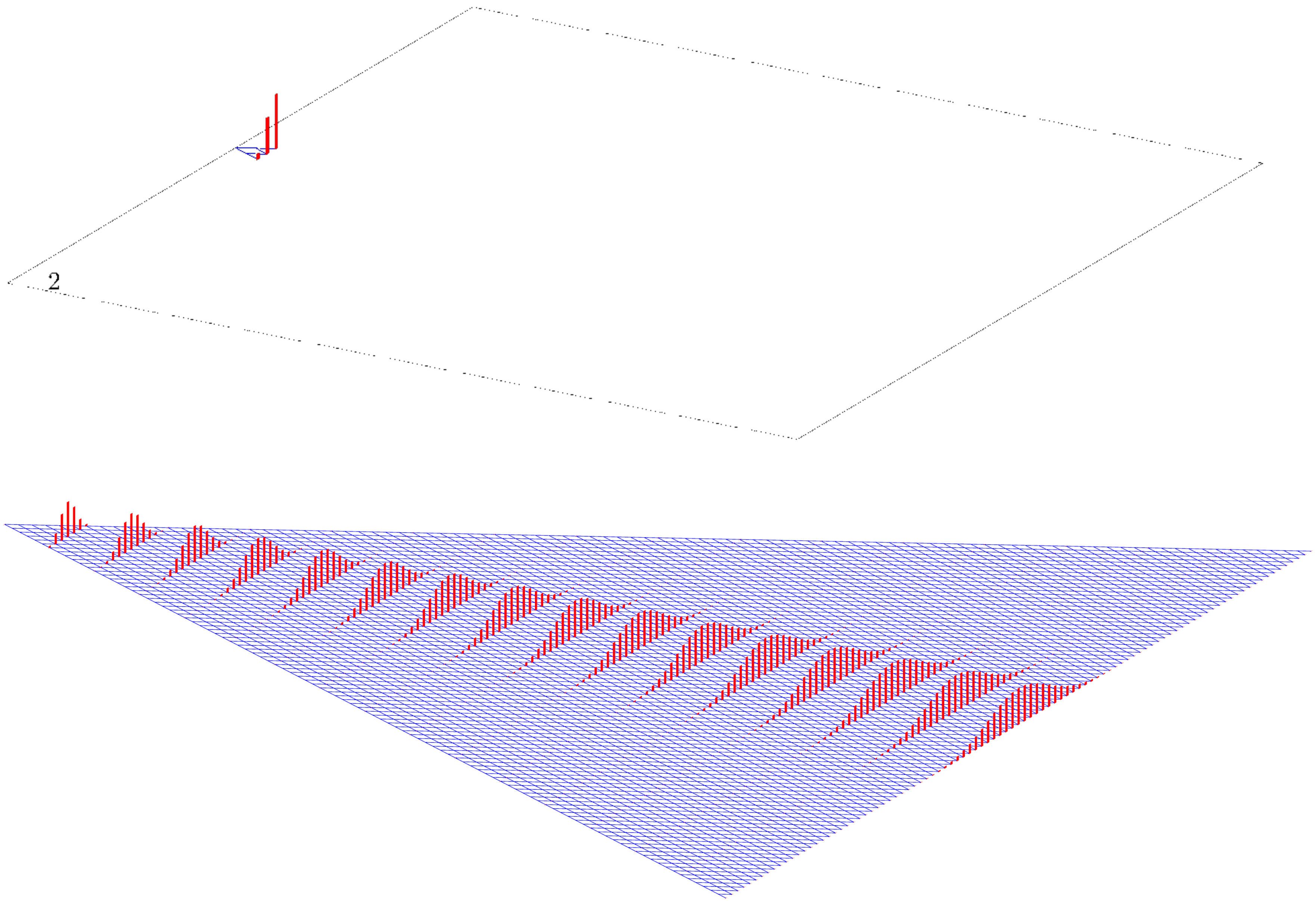


Figure Source: https://tex.stackexchange.com/questions/471912/binomial-tree-converging-to-a-normal-distribution-3d

Binomial distribution towards Normal distribution

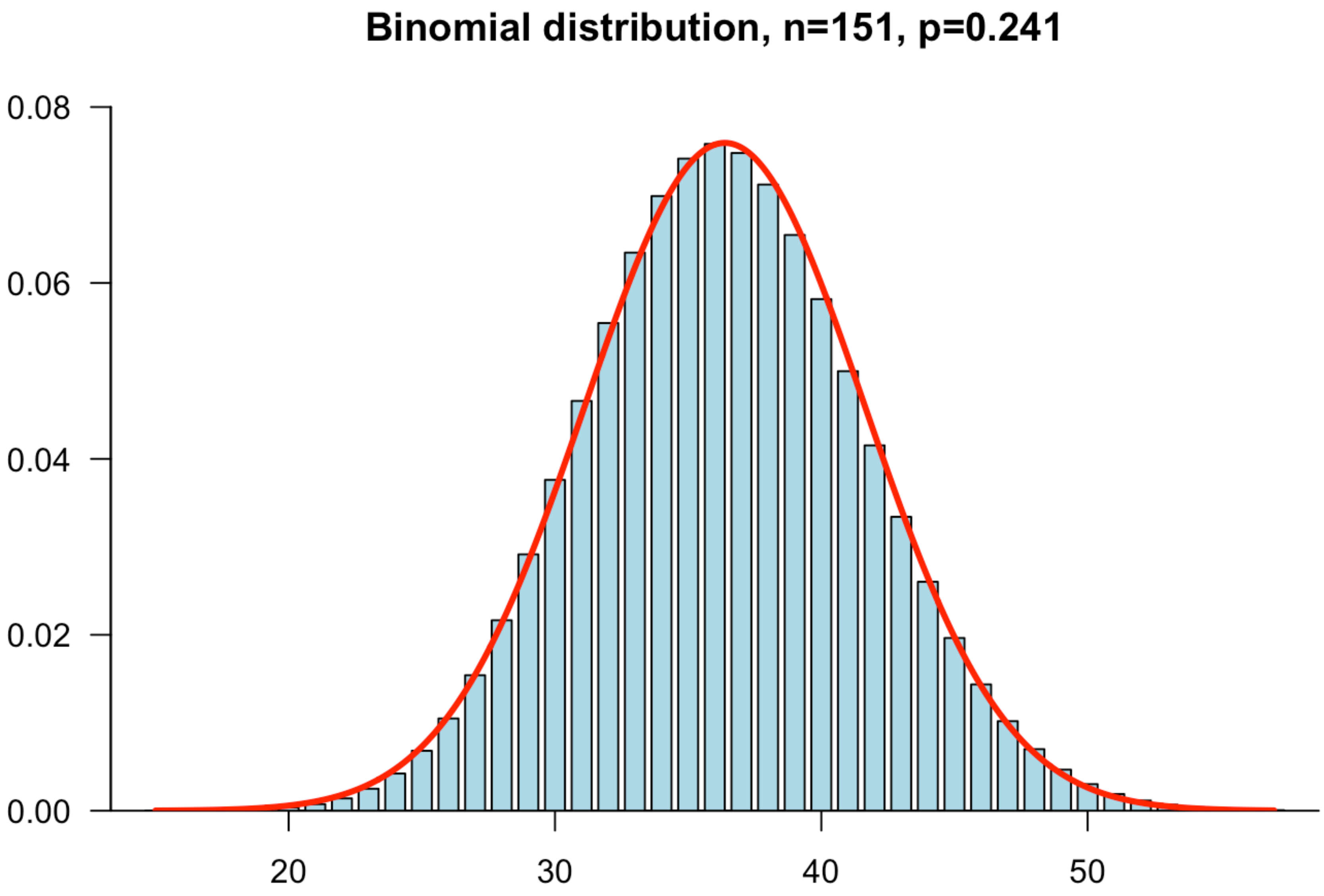
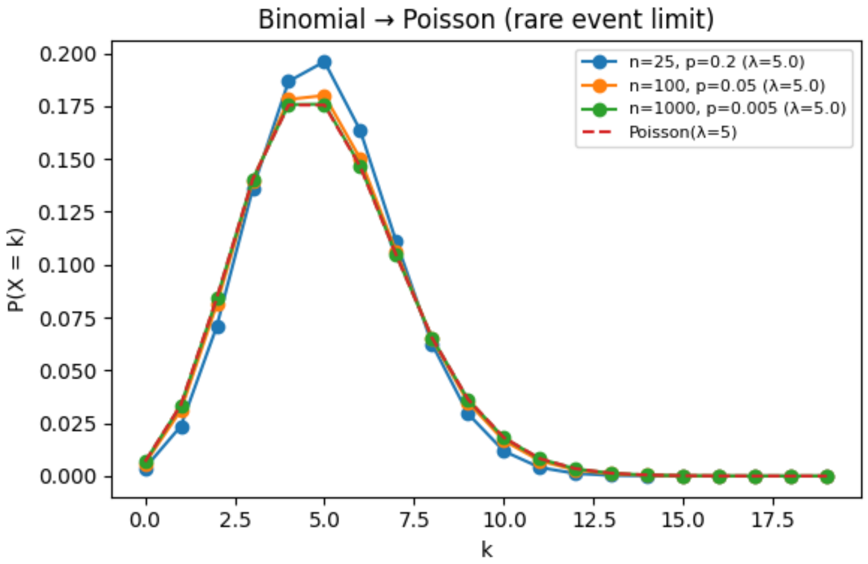
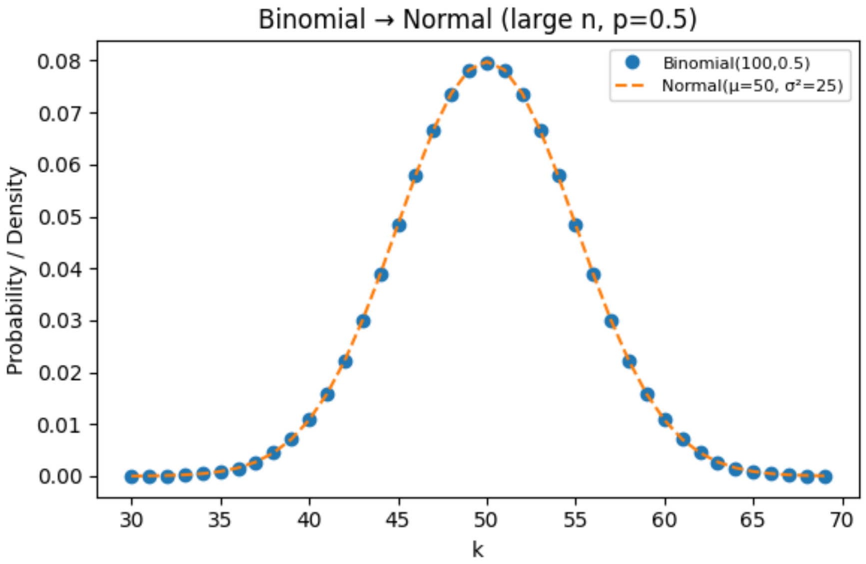
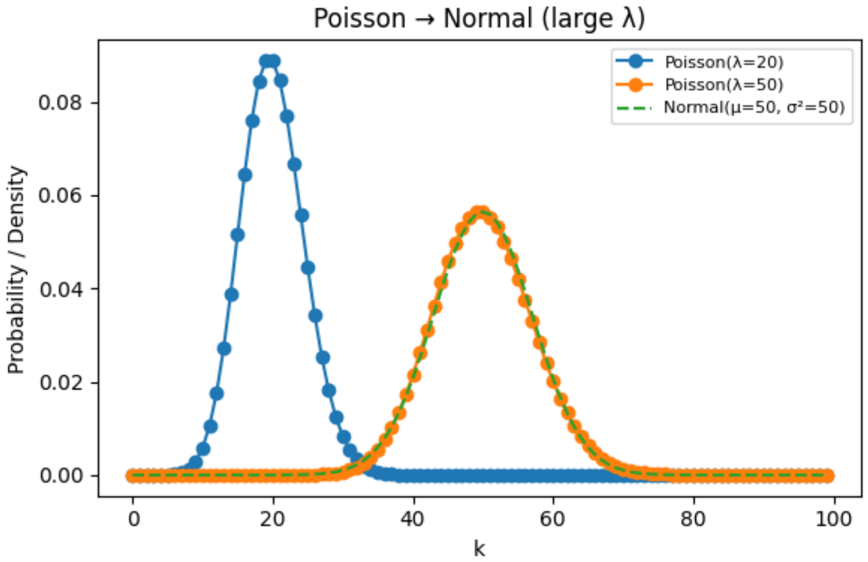


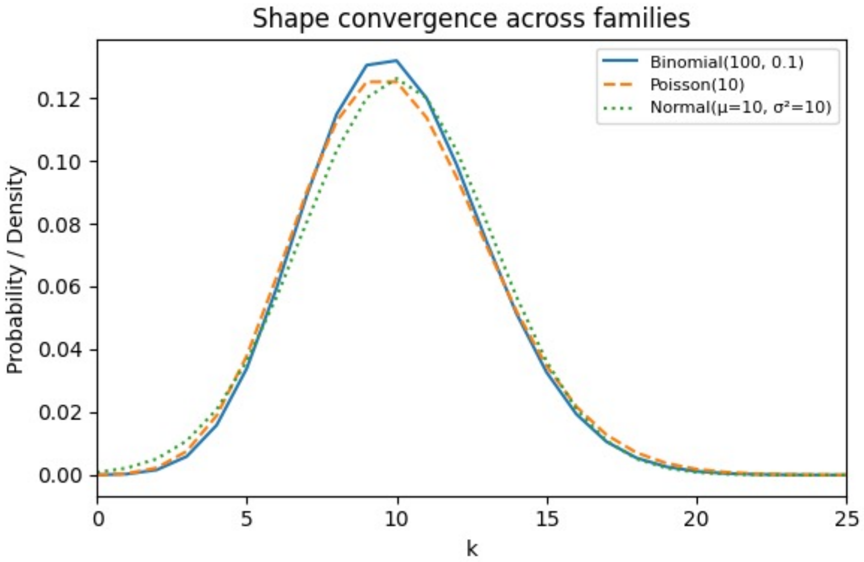
Figure Source: https://stackoverflow.com/questions/60546225/plotting-the-normal-and-binomial-distribution-in-same-plot

Binomial / Normal / Poisson distribution

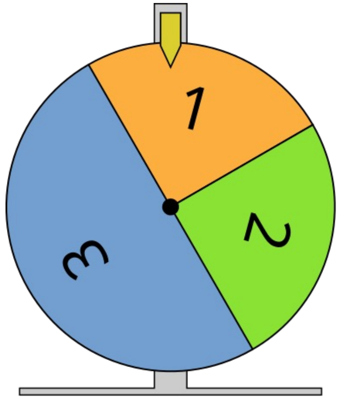








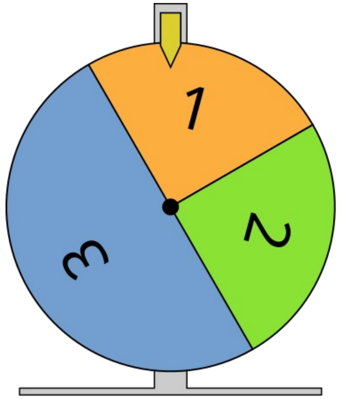
Conditioned Multinomial Distribution



EXAMPLE:
Suppose a lucky wheel has three numbers 1, 2, and 3 with areas covering 25%, 25%, and 50% of the wheel, respectively. If we spin the wheel 6 times independently, what is the probability of getting exactly one "1", two "2"s, and three "3"s?

?

Conditioned Multinomial Distribution

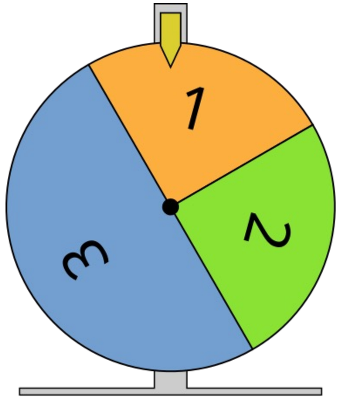


EXAMPLE:
Suppose a lucky wheel has three numbers 1, 2, and 3 with areas covering 25%, 25%, and 50% of the wheel, respectively. If we spin the wheel 6 times independently, what is the probability of getting exactly one "1", two "2"s, and three "3"s?

This is exactly the multinomial distribution:

$$\mathbb{P}(\mathrm{A}=1, \mathrm{B}=2, \mathrm{C}=3) = \frac{6!}{1!\,2!\,3!} \cdot (0.25)^1 (0.25)^2 (0.5)^3 \approx 0.1172$$

Conditioned Multinomial Distribution



EXAMPLE:

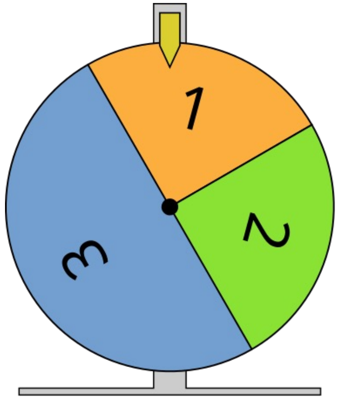
Suppose a lucky wheel has three numbers 1, 2, and 3 with areas covering 25%, 25%, and 50% of the wheel, respectively. If we spin the wheel 6 times independently, what is the probability of getting exactly x “1”s, y “2”s, and z “3”s?

?

This is exactly the multinomial distribution:

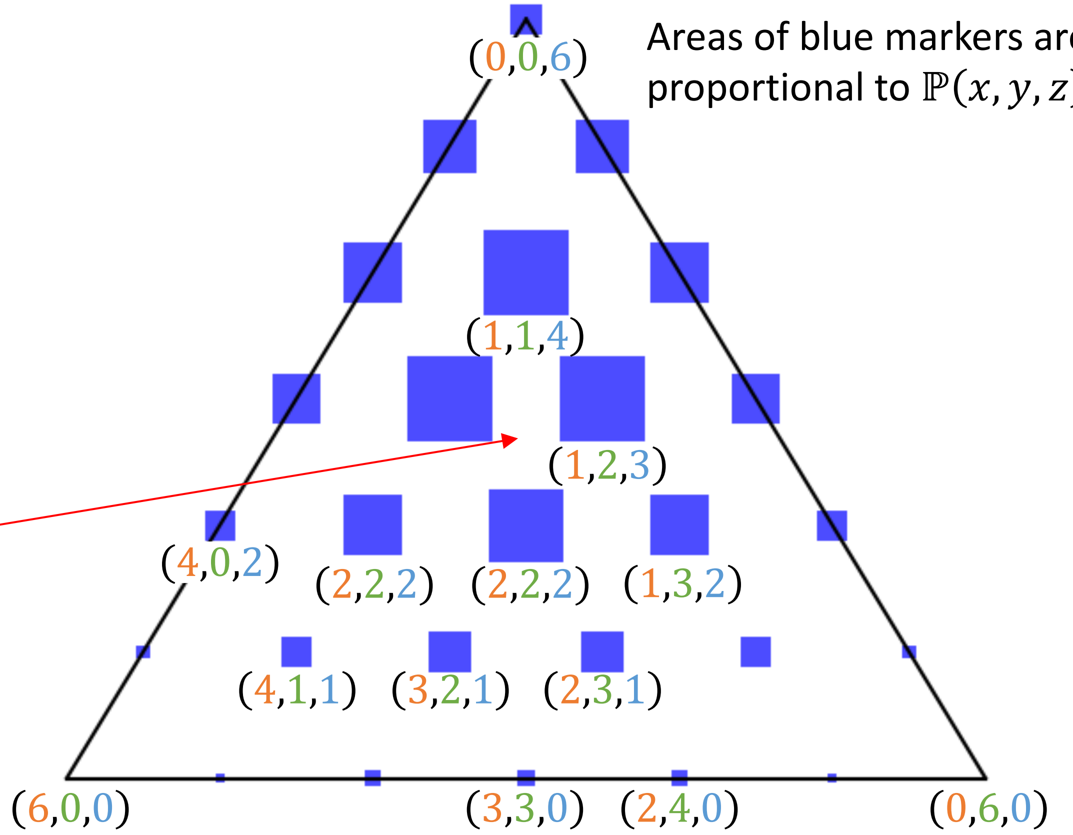
$$\mathbb{P}(\mathrm{A}=1, \mathrm{B}=2, \mathrm{C}=3) = \frac{6!}{1!\,2!\,3!} \cdot (0.25)^1 (0.25)^2 (0.5)^3 \approx 0.1172$$

Conditioned Multinomial Distribution



EXAMPLE:
Suppose a lucky wheel has three numbers 1, 2, and 3 with areas covering 25%, 25%, and 50% of the wheel, respectively. If we spin the wheel 6 times independently, what is the probability of getting exactly x "1"s, y "2"s, and z "3"s?

Areas of blue markers are proportional to $\mathbb{P}(x, y, z)$

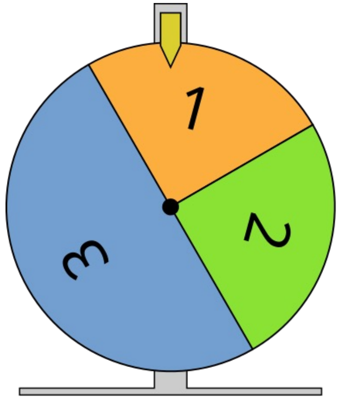


This is exactly the multinomial distribution:

$$\mathbb{P}(\mathrm{A} = x, \mathrm{B} = y, \mathrm{C} = z) = \frac{6!}{x!\, y!\, z!} \cdot (0.25)^x (0.25)^y (0.5)^z$$

Notice that the highest ones are close to (25%, 25%, 50%). But (1.5, 1.5, 3) is not an integral solution.

Conditioned Multinomial Distribution



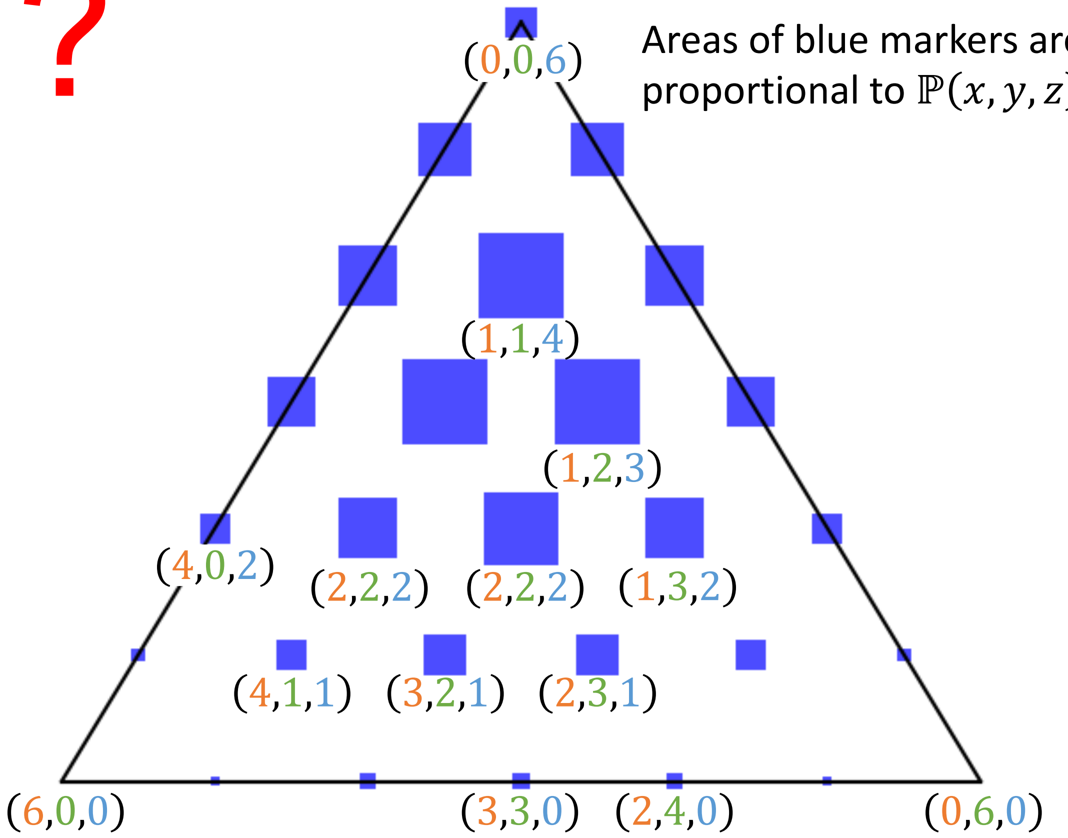
EXAMPLE:

Suppose a lucky wheel has three numbers 1, 2, and 3 with areas covering 25%, 25%, and 50% of the wheel, respectively. If we spin the wheel 6 times independently, what is the probability of getting exactly x "1"s, y "2"s, and z "3"s if we know additionally that condition COND holds, namely the sum of the draws is equal to 10?

?

This is exactly the multinomial distribution:

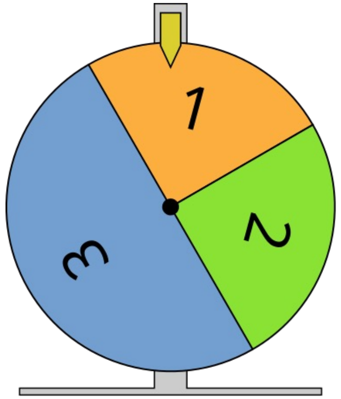
$$\mathbb{P}(\mathrm{A} = x, \mathrm{B} = y, \mathrm{C} = z) = \frac{6!}{x!\,y!\,z!} \cdot (0.25)^x (0.25)^y (0.5)^z$$



Areas of blue markers are proportional to $\mathbb{P}(x, y, z)$

Python file 206: https://github.com/northeastern-datalab/cs7840-activities/tree/main/notebooks/206_maxEntropy.ipynb

Conditioned Multinomial Distribution

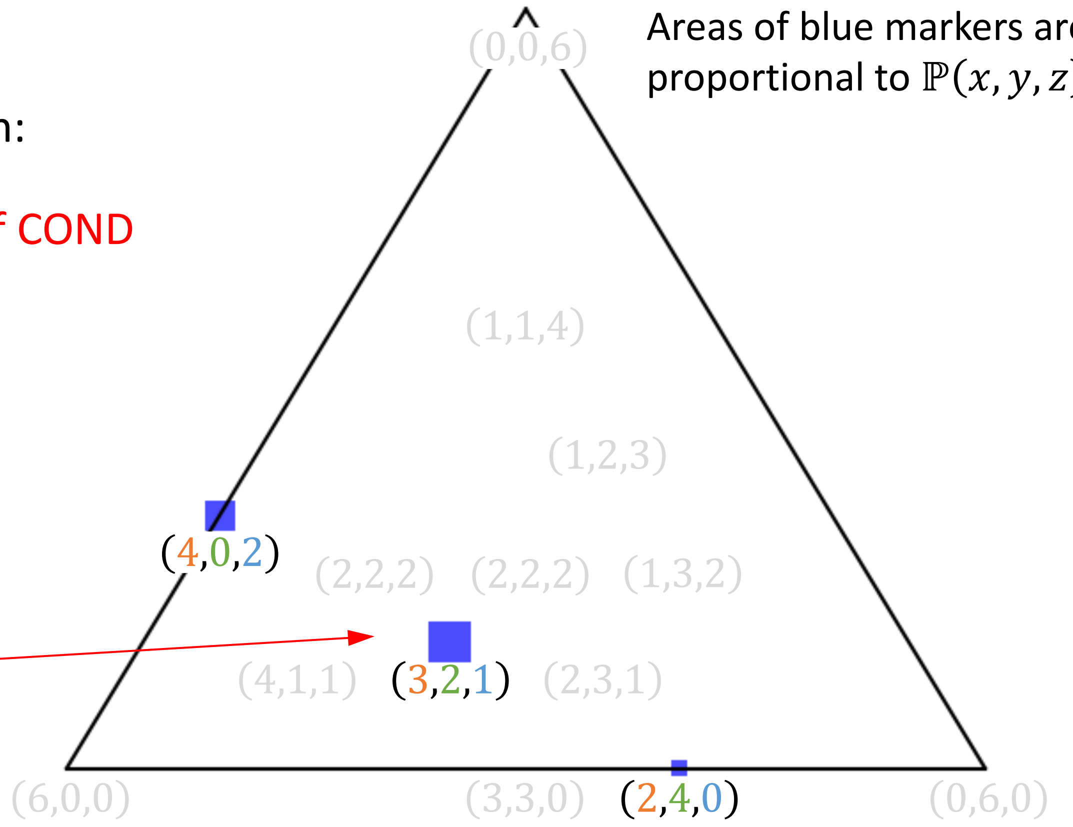


EXAMPLE:
Suppose a lucky wheel has three numbers 1, 2, and 3 with areas covering 25%, 25%, and 50% of the wheel, respectively. If we spin the wheel 6 times independently, what is the probability of getting exactly x "1"s, y "2"s, and z "3"s if we know additionally that condition COND holds, namely the sum of the draws is equal to 10?

This is exactly the conditional (or truncated) multinomial distribution:

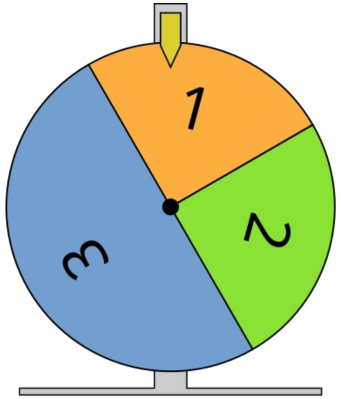
$$\mathbb{P}(\mathrm{A}=x,\mathrm{B}=y,\mathrm{C}=z)\begin{cases}\propto \dfrac{6!}{x!\,y!\,z!}\cdot(0.25)^x(0.25)^y(0.5)^z & \text{if COND}\\[2ex] = 0 & \text{else}\end{cases}$$

need to be scaled appropriately



Areas of blue markers are proportional to $\mathbb{P}(x,y,z)$

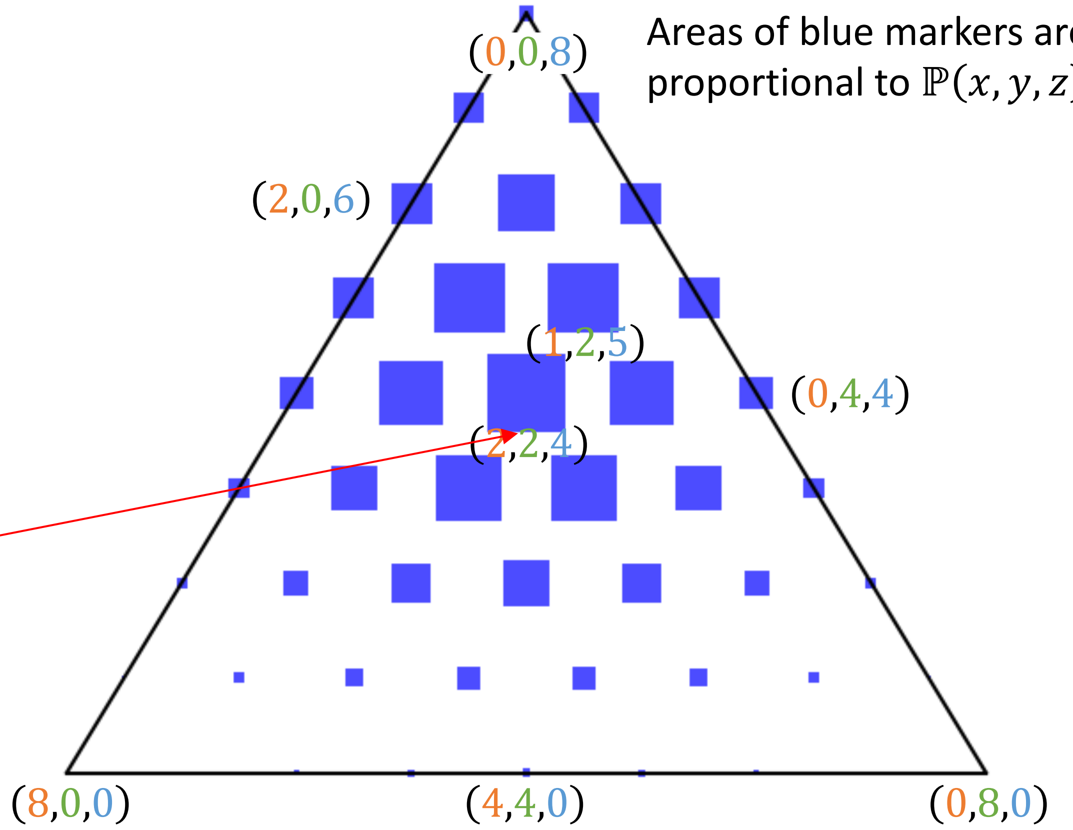
Conditioned Multinomial Distribution



EXAMPLE:
Suppose a lucky wheel has three numbers 1, 2, and 3 with areas covering 25%, 25%, and 50% of the wheel, respectively. If we spin the wheel 8 times independently, what is the probability of getting exactly x "1"s, y "2"s, and z "3"s?

This is exactly the multinomial distribution:

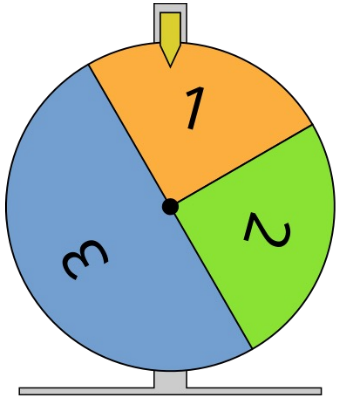
$$\mathbb{P}(\mathrm{A}=x, \mathrm{B}=y, \mathrm{C}=z) = \frac{8!}{x!\,y!\,z!} \cdot (0.25)^x (0.25)^y (0.5)^z$$



Areas of blue markers are proportional to $\mathbb{P}(x, y, z)$

Now with 8 total spins, the highest one with (2, 2, 4) fits perfectly what we would expect: (25%, 25%, 50%)

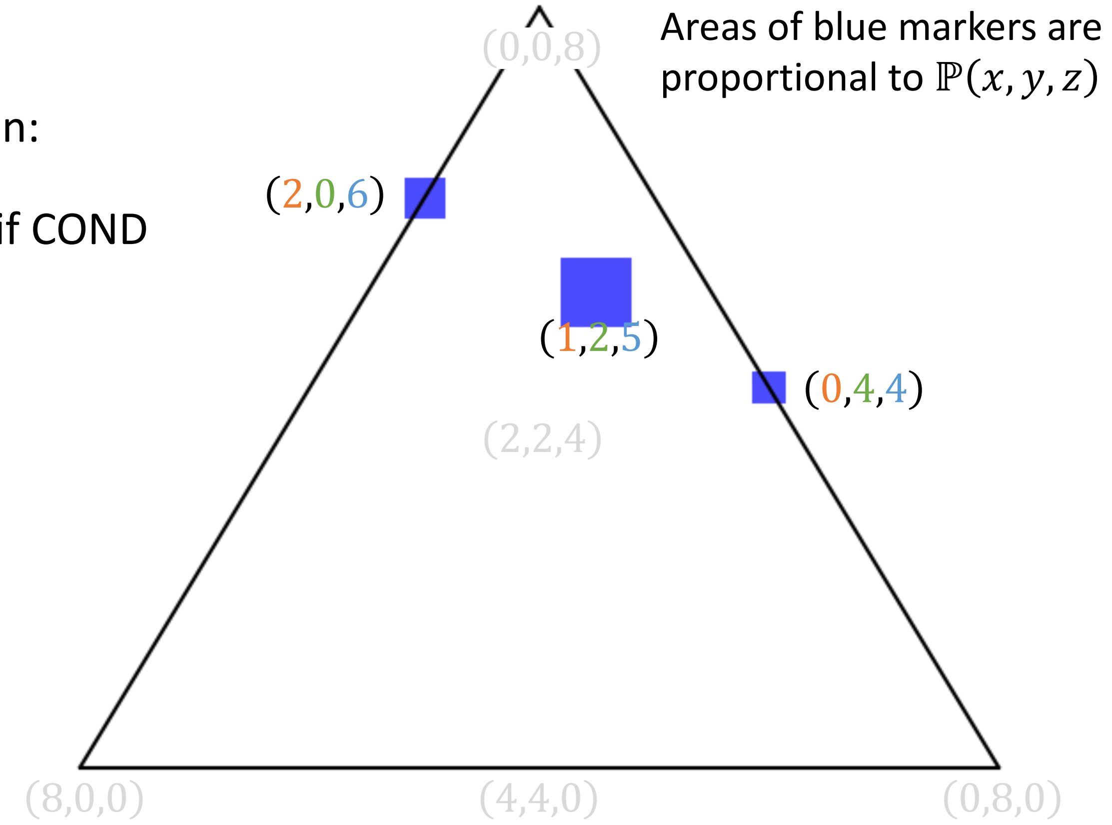
Conditioned Multinomial Distribution



EXAMPLE:
Suppose a lucky wheel has three numbers 1, 2, and 3 with areas covering 25%, 25%, and 50% of the wheel, respectively. If we spin the wheel 8 times independently, what is the probability of getting exactly x "1"s, y "2"s, and z "3"s if we know additionally that condition COND holds, namely the sum of the draws is equal to 20?

This is exactly the conditional (or truncated) multinomial distribution:

$$\mathbb{P}(\mathrm{A}=x,\mathrm{B}=y,\mathrm{C}=z)\begin{cases}\propto \dfrac{8!}{x!\,y!\,z!}\cdot(0.25)^x(0.25)^y(0.5)^z & \text{if COND}\\[2ex] =\ 0 & \text{else}\end{cases}$$



Wallis' argument for Max Entropy

The following argument is based on Wallis' argument given in [Jaynes'03] "Probability theory: the logic of science", Cambridge press, 2003, Section 11.4 (https://doi.org/10.1017/CBO9780511790423). The argument is also given on https://en.wikipedia.org/wiki/Principle_of_maximum_entropy#The_Wallis_derivation

Maximum Entropy Principle

Recall: Entropy as a measure of uncertainty

For discrete RV X with distribution $\mathbb{P}[X = x_i] = p_i$:

$$H(X) = -\sum_{i=1} p_i \cdot \lg(p_i) = \mathbb{E}_{X \sim p}\left[\lg\left(\frac{1}{p(X)}\right)\right]$$

For continuous RV X with PDF $p(x)$, the "differential entropy"

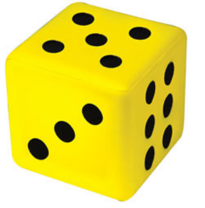
$$H(X) = -\int_{-\infty}^{\infty} p(x) \cdot \lg\big(p(x)\big) \cdot dx$$

MAXIMUM ENTROPY PRINCIPLE: The probability distribution with largest entropy is the one which best represents the current state of knowledge about a system.

But why ?

The Wallis derivation

Take an unbiased die with $m = 6$ sides and throw it $n = 7$ times.



Assume we are searching for an outcome probability distribution (e.g. the fraction of times that each of the $m = 6$ faces of an unbiased die come up if we throw it $n = 7$ times).

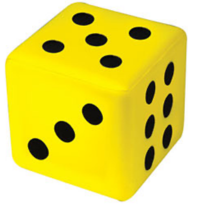
We have some other information I (some constraint) about the distribution (e.g. that the average roll was $\mu = 4$)

What is the most likely <u>outcome probability distribution</u> ?

The Wallis derivation

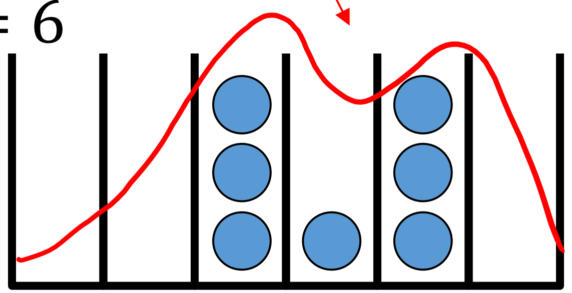
Assume we are searching for an outcome probability distribution (e.g. the fraction of times that each of the $m = 6$ faces of an unbiased die come up if we throw it $n = 7$ times).

Take an unbiased die with $m = 6$ sides and throw it $n = 7$ times.

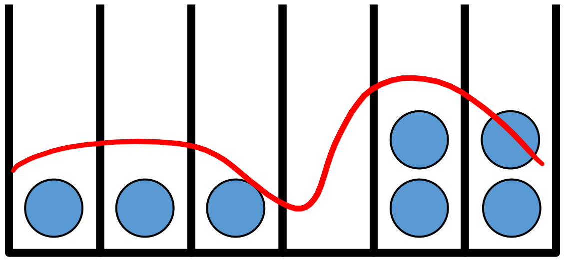


We have some other information I (some constraint) about the distribution (e.g. that the average roll was $\mu = 4$)

What is the most likely outcome distribution (given I)?



What is the most likely <u>outcome probability distribution</u>?



Wallis' thought experiment:

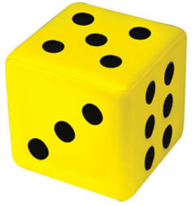
- We have $n \gg m$ balls and throw them randomly into m bins, each bin is treated the same (like an unbiased die with m sides)
- Repeat this until the resulting outcome probability distribution in the bins conforms to our information (constraint) I
- What is the most likely probability distribution to result from this game?

We will see this is the one that maximizes entropy ☺

The Wallis derivation

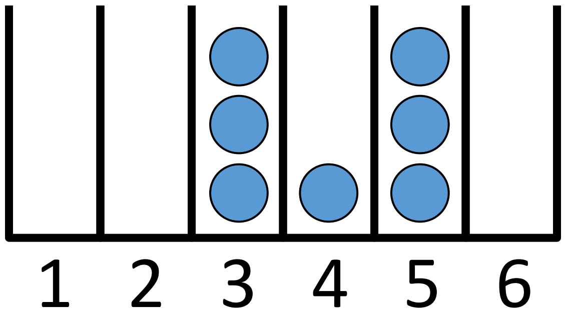
What is the PDF of the possible (unconstrained) outcomes ?

Take an unbiased die with $m = 6$ sides and throw it $n = 7$ times.

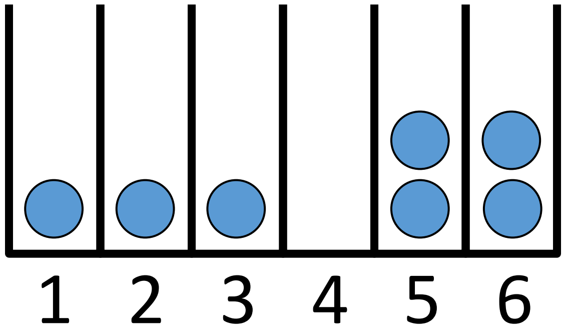


What is the most likely outcome distribution (given I)?

$m = 6$



$n = 7$



$n = 7$

The Wallis derivation

What is the PDF of the possible (unconstrained) outcomes?

Multinomial distribution

$$\text{pmf} = \underbrace{m^{-n}}_{} \cdot \underbrace{\frac{n!}{n_1! \cdot n_2! \cdots n_m!}}_{}$$

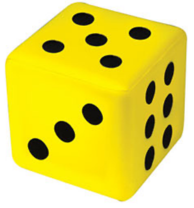
Number of balls in each bin

if all balls had a unique id

Multinomial coefficient $\binom{n}{n_1,\ldots,n_m} =: W$

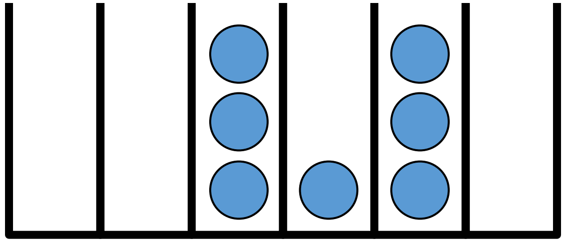
This is the multiplicity = the number of ways in which you can partition an n-element set into disjoint subsets of sizes $n_1, n_2, \ldots, n_m$ with $\sum_i n_i = n$

Take an unbiased die with $m = 6$ sides and throw it $n = 7$ times.



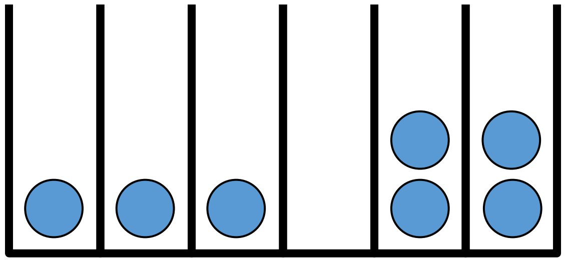
What is the most likely outcome distribution (given I)?

$m = 6$



1 2 3 4 5 6

$n = 7$ $W = 140$



1 2 3 4 5 6

$n = 7$ $W = 1260$

The Wallis derivation

New goal: Maximize the following expression
s.t. constraint I (not shown):

$$\max \left[C = \frac{n!}{n_1! \cdot n_2! \cdots n_m!} \right]$$

We will show that maximizing W can be achieved by maximizing the entropy

The Wallis derivation

New goal: Maximize the following expression
s.t. constraint I (not shown):

$$\max \left[C = \frac{n!}{n_1! \cdot n_2! \cdots n_m!} \right]$$

We will show that maximizing W can be achieved by maximizing the entropy

$$\frac{1}{n} \cdot \lg(C) = \frac{1}{n} \cdot \lg\left(\frac{n!}{n_1! \cdot n_2! \cdots n_m!}\right)$$

$$= \frac{1}{n} \cdot \lg\left(\frac{n!}{(np_1)! \cdot (np_2)! \cdots (np_m)!}\right)$$

$$= \frac{1}{n} \cdot \left(\lg(n!) - \sum_{i=1}^{m} \lg\big((np_i)!\big)\right)$$

Now we are stuck. What next ?

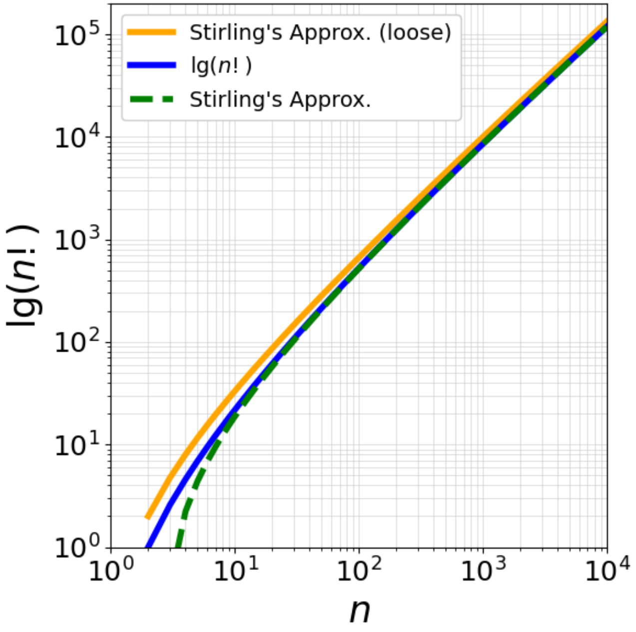
The Wallis derivation

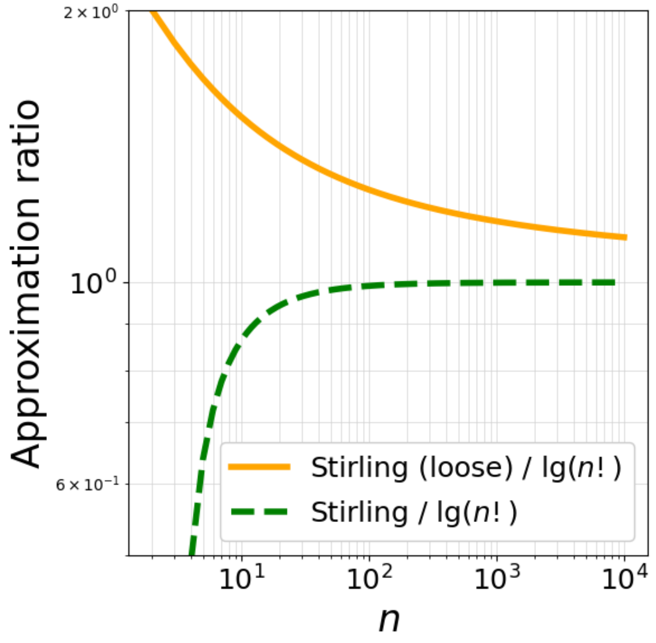
New goal: Maximize the following expression s.t. constraint I (not shown):

$$\max \left[C = \frac{n!}{n_1! \cdot n_2! \cdots n_m!} \right]$$

$$\begin{aligned}
\frac{1}{n} \cdot \lg(C) &= \frac{1}{n} \cdot \lg\left(\frac{n!}{n_1! \cdot n_2! \cdots n_m!}\right) \\
&= \frac{1}{n} \cdot \lg\left(\frac{n!}{(np_1)! \cdot (np_2)! \cdots (np_m)!}\right) \\
&= \frac{1}{n} \cdot \left(\lg(n!) - \textstyle\sum_{i=1}^{m} \lg\big((np_i)!\big)\right) \\
&\approx \cancel{\frac{1}{n}} \cdot \left(\cancel{n} \cdot \lg(n) - \textstyle\sum_{i=1}^{m} \cancel{n} p_i \cdot \lg(np_i)\right) \\
&= \lg(n) - \textstyle\sum_{i=1}^{m} p_i \cdot (\lg n + \lg p_i) \\
&= \cancel{\lg(n)} - \cancel{\lg(n) \cdot \textstyle\sum_{i=1}^{m} p_i} - \textstyle\sum_{i=1}^{m} p_i \cdot \lg(p_i)
\end{aligned}$$

$$\boxed{\lim_{n\to\infty} \frac{1}{n} \cdot \lg(C) = -\textstyle\sum_{i=1}^{m} p_i \cdot \lg(p_i)} = H(\boldsymbol{p})$$





Assume $n \to \infty$, then apply Stirling's formula:

$$\ln(n!) \approx n \cdot \ln(n)$$

$$\lg(n!) \approx n \cdot \left(\frac{\lg(n)}{\lg(e)}\right) = n \cdot \lg(n) - n \cdot \lg(e)$$

$$\approx n \cdot \lg(n)$$

All we need to do is to maximize entropy under the constraints of our testable information I. There is no need for any interpretation of H in terms of information theoretic notion like "amount of uncertainty"

Python file 224: https://github.com/northeastern-datalab/cs7840-activities/blob/main/notebooks/224_Sterling.ipynb , see also: https://en.wikipedia.org/wiki/Stirling%27s_approximation

Jaynes' die

The following argument is based on Wallis' argument given in [Jaynes'03] "Probability theory: the logic of science", Cambridge press, 2003, Section 11.4 (https://doi.org/10.1017/CBO9780511790423). The argument is also given on https://en.wikipedia.org/wiki/Principle_of_maximum_entropy#The_Wallis_derivation

Jaynes' die

I find this interpretation problematic. Rather use the interpretation we used in the Wallis derivation: what is the most likely distribution on the outcomes

Example 3: Jaynes' Dice

A die has been tossed a very large number N of times, and we are told that the average number of spots per toss was not 3.5, as we might expect from an honest die, but 4.5. Translate this information into a probability assignment $p_n, n = 1, 2, \ldots, 6$, for the n-th face to come up on the next toss.

This problem is similar to the above except for two changes: our support is $\{1, \ldots, 6\}$ and the expectation of the die roll is 4.5. We can formulate the problem in a similar way with the following Lagrangian with an added term for the expected value (B):

$$\mathcal{L}(p_1, \ldots, p_6, \lambda_0, \lambda_1) = -\sum_{k=1}^{6} p_k \log(p_k) - \lambda_0(\sum_{k=1}^{6} p_k - 1) - \lambda_1(\sum_{k=1}^{6} kp_k - B) \tag{11}$$

Taking the partial derivatives and setting them to zero, we get:

$$\begin{aligned} \log(p_k) = -1 - \lambda_0 - k\lambda_1 &= 0 \\ \log(p_k) &= -1 - \lambda_0 - k\lambda_1 \\ p_k &= e^{-1-\lambda_0-k\lambda_1} \end{aligned} \tag{12}$$

$$\sum_{k=1}^{6} p_k = 1 \tag{13}$$

$$\sum_{k=1}^{6} kp_k = B \tag{14}$$

Define a new quantity $Z(\lambda_1)$ by substituting Equation 12 into 13:

$$Z(\lambda_1) := e^{-1-\lambda_0} = \frac{1}{\sum_{k=1}^{6} e^{-k\lambda_1}} \tag{15}$$

Substituting Equation 12, and dividing Equation 14 by 13

$$\begin{aligned} \frac{\sum_{k=1}^{6} ke^{-1-\lambda_0-k\lambda_1}}{\sum_{k=1}^{6} e^{-1-\lambda_0-k\lambda_1}} &= B \\ \frac{\sum_{k=1}^{6} ke^{-k\lambda_1}}{\sum_{k=1}^{6} e^{-k\lambda_1}} &= B \end{aligned} \tag{16}$$

Going back to Equation 12 and defining it in terms of Z:

$$p_k = \frac{1}{Z(\lambda_1)} e^{-k\lambda_1} \tag{17}$$

Unfortunately, now we're at an impasse because there is no closed form solution. Interesting to note that the solution is just an exponential-like distribution with parameter λ_1 and $Z(\lambda_1)$ as a normalization constant to make sure the probabilities sum to 1. Equation 16 gives us the desired value of λ_1. We can easily find a solution using any root solver, such as the code below:

Source: https://bjlkeng.io/posts/maximum-entropy-distributions/

Jaynes' die

```
from numpy import exp
from scipy.optimize import newton

a, b, B = 1, 6, 4.5

# Equation 15
def z(lamb):
    return 1. / sum(exp(-k*lamb) for k in range(a, b + 1))

# Equation 16
def f(lamb, B=B):
    y = sum(k * exp(-k*lamb) for k in range(a, b + 1))
    return y * z(lamb) - B

# Equation 17
def p(k, lamb):
    return z(lamb) * exp(-k * lamb)

lamb = newton(f, x0=0.5)
print("Lambda = %.4f" % lamb)
for k in range(a, b + 1):
    print("p_%d = %.4f" % (k, p(k, lamb)))

# Output:
#   Lambda = -0.3710
#   p_1 = 0.0544
#   p_2 = 0.0788
#   p_3 = 0.1142
#   p_4 = 0.1654
#   p_5 = 0.2398
#   p_6 = 0.3475
```

Define a new quantity $Z(\lambda_1)$ by substituting Equation 12 into 13:

$$Z(\lambda_1) := e^{-1-\lambda_0} = \frac{1}{\sum_{k=1}^{6} e^{-k\lambda_1}} \tag{15}$$

Substituting Equation 12, and dividing Equation 14 by 13

$$\begin{aligned} \frac{\sum_{k=1}^{6} ke^{-1-\lambda_0-k\lambda_1}}{\sum_{k=1}^{6} e^{-1-\lambda_0-k\lambda_1}} &= B \\ \frac{\sum_{k=1}^{6} ke^{-k\lambda_1}}{\sum_{k=1}^{6} e^{-k\lambda_1}} &= B \end{aligned} \tag{16}$$

Going back to Equation 12 and defining it in terms of Z:

$$p_k = \frac{1}{Z(\lambda_1)} e^{-k\lambda_1} \tag{17}$$

Unfortunately, now we're at an impasse because there is no closed form solution. Interesting to note that the solution is just an exponential-like distribution with parameter λ_1 and $Z(\lambda_1)$ as a normalization constant to make sure the probabilities sum to 1. Equation 16 gives us the desired value of λ_1. We can easily find a solution using any root solver, such as the code below:

Source: https://bjlkeng.io/posts/maximum-entropy-distributions/

Using the Max Entropy principle to derive the Normal Distribution and outcomes of dice rolls

Maximum Entropy Distributions

EXAMPLE: Suppose a continuous random variable X has given mean (1st moment) μ and variance (2nd moment) σ^2. Which PDF $p(x)$ has the maximum entropy $H(x)$?

How would you formalize this problem ?

Maximum Entropy Distributions

EXAMPLE: Suppose a continuous random variable X has given mean (1st moment) μ and variance (2nd moment) σ^2. Which PDF $p(x)$ has the maximum entropy $H(x)$?

Differential Entropy

$$H(X) = -\int_{-\infty}^{\infty} p(x) \cdot \lg\big(p(x)\big) \cdot dx$$

PDF constraint

$$\int_{-\infty}^{\infty} p(x) \cdot dx = 1$$

Moment constraint(s)

$$\int_{-\infty}^{\infty} (x-\mu)^2 \cdot p(x) \cdot dx = \sigma^2$$

"Only one constraint is needed, because the definition of σ^2 already includes μ."

Maximum Entropy Distributions

EXAMPLE: Suppose a continuous random variable X has given mean (1st moment) μ and variance (2nd moment) σ^2. Which PDF $p(x)$ has the maximum entropy $H(x)$?

Differential Entropy

$$H(X) = -\int_{-\infty}^{\infty} p(x) \cdot \lg\big(p(x)\big) \cdot dx$$

Lagrangian

$$\mathcal{L} = \ ?$$

PDF constraint

$$\int_{-\infty}^{\infty} p(x) \cdot dx \ = 1$$

Moment constraint(s)

$$\int_{-\infty}^{\infty} (x-\mu)^2 \cdot p(x) \cdot dx \ = \sigma^2$$

Maximum Entropy Distributions

EXAMPLE: Suppose a continuous random variable X has given mean (1st moment) μ and variance (2nd moment) σ^2. Which PDF $p(x)$ has the maximum entropy $H(x)$?

Differential Entropy

$$H(X) = -\int_{-\infty}^{\infty} p(x) \cdot \lg\big(p(x)\big) \cdot dx$$

PDF constraint

$$\int_{-\infty}^{\infty} p(x) \cdot dx \ = 1$$

Moment constraint(s)

$$\int_{-\infty}^{\infty} (x-\mu)^2 \cdot p(x) \cdot dx \ = \sigma^2$$

Lagrangian

$$\mathcal{L} = -\int_{-\infty}^{\infty} p(x) \cdot \lg\big(p(x)\big) \cdot dx$$

$$+\lambda_0\left(\int_{-\infty}^{\infty} p(x) \cdot dx \ - 1\right)$$

$$+\lambda_1\left(\int_{-\infty}^{\infty} (x-\mu)^2 \cdot p(x) \cdot dx \ - \sigma^2\right)$$

Maximum Entropy Distributions

EXAMPLE: Suppose a continuous random variable X has given mean (1st moment) μ and variance (2nd moment) σ^2. Which PDF $p(x)$ has the maximum entropy $H(x)$?

Partial derivation (calculus of variation)

(functional) function of a function

$$\frac{\partial \mathcal{L}}{\partial p(x)} = -\frac{1}{\ln(2)}\left(1+\ln\left(p(x)\right)\right)$$

Calculus cheat sheet

$$\lg(x)' = \left(\frac{\ln(x)}{\ln(2)}\right)' = \frac{1}{x \cdot \ln(2)}$$

$$(x \cdot \ln(x))' = \cancel{x}\frac{1}{\cancel{x}} + \ln(x)$$

$$+\lambda_0$$

$$+\lambda_1(x-\mu)^2$$

$$= 0$$

Lagrangian

$$\mathcal{L} = -\int_{-\infty}^{\infty} \underbrace{p(x) \cdot \lg\left(p(x)\right)}_{\frac{1}{\ln(2)} \cdot p(x) \cdot \ln\left(p(x)\right)} \cdot dx$$

$$+\lambda_0\left(\int_{-\infty}^{\infty} p(x) \cdot dx \; - 1\right)$$

$$+\lambda_1\left(\int_{-\infty}^{\infty} (x-\mu)^2 \cdot p(x) \cdot dx \; - \sigma^2\right)$$

Maximum Entropy Distributions

EXAMPLE: Suppose a continuous random variable X has given mean (1st moment) μ and variance (2nd moment) σ^2. Which PDF $p(x)$ has the maximum entropy $H(x)$?

$$-\frac{1}{\ln(2)}\big(1+\ln(p(x))\big)+\lambda_0+\lambda_1(x-\mu)^2=0$$

$$-\big(1+\ln(p(x))\big)+\lambda_0'+\lambda_1'(x-\mu)^2=0$$

$$p(x)=e^{\overbrace{\lambda_0''}^{\lambda_0'-1}+\lambda_1'(x-\mu)^2}$$

Constraints

?

Maximum Entropy Distributions

EXAMPLE: Suppose a continuous random variable X has given mean (1st moment) μ and variance (2nd moment) σ^2. Which PDF $p(x)$ has the maximum entropy $H(x)$?

$$-\frac{1}{\ln(2)}\big(1+\ln\big(p(x)\big)\big)+\lambda_0+\lambda_1(x-\mu)^2=0$$

$$-\big(1+\ln\big(p(x)\big)\big)+\lambda_0'+\lambda_1'(x-\mu)^2=0$$

$$p(x)=e^{\lambda_0''+\lambda_1'(x-\mu)^2}$$

Constraints

$$\left.\begin{array}{lll}\int_{-\infty}^{\infty} p(x)\cdot dx=1 & \Rightarrow & \int_{-\infty}^{\infty} e^{\lambda_0''+\lambda_1'(x-\mu)^2}\cdot dx=1 \\ \int_{-\infty}^{\infty}(x-\mu)^2\cdot p(x)\cdot dx=\sigma^2 & \Rightarrow & \int_{-\infty}^{\infty}(x-\mu)^2\cdot e^{\lambda_0''+\lambda_1'(x-\mu)^2}\cdot dx=\sigma^2\end{array}\right\}\Rightarrow \begin{array}{l}\lambda_1'=-\frac{1}{2\sigma^2} \\ e^{\lambda_0''}=\sqrt{-\frac{\lambda_1'}{\pi}}=\frac{1}{\sigma\sqrt{2\pi}}\end{array}$$

For details, see next page

$$\boxed{p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

The maximum entropy principle is empirically justified ☺

Maximum Entropy Distribution: DETAILS

BACKUP

$$\int_{-\infty}^{\infty} e^{\lambda_0'' + \lambda_1'(x-\mu)^2} \cdot dx = 1$$

$$e^{\lambda_0''} \cdot \int_{-\infty}^{\infty} e^{\lambda_1'(x-\mu)^2} \cdot dx = 1$$

$$\underbrace{\int_{-\infty}^{\infty} e^{\lambda_1'(x-\mu)^2} \cdot dx}_{} = e^{-\lambda_0''}$$

$$\sqrt{\frac{\pi}{-\lambda_1'}} = e^{-\lambda_0''}$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 \cdot e^{\lambda_0'' + \lambda_1'(x-\mu)^2} \cdot dx = \sigma^2$$

$$e^{\lambda_0''} \cdot \underbrace{\int_{-\infty}^{\infty} z^2 \cdot e^{\lambda_1' z^2} \cdot dz}_{} = \sigma^2$$

$$\frac{1}{2}\sqrt{\frac{\pi}{-\lambda_1'^{\,3}}} = \sigma^2 \cdot e^{-\lambda_0''}$$

$$\frac{1}{2\lambda_1'} \cancel{\sqrt{\frac{\pi}{-\lambda_1'}}} = \sigma^2 \cdot \cancel{\sqrt{\frac{\pi}{-\lambda_1'}}}$$

$$\lambda_1' = -\frac{1}{2\sigma^2}$$

$$e^{\lambda_0''} = \sqrt{-\frac{\lambda_1'}{\pi}} = \frac{1}{\sigma\sqrt{2\pi}}$$

Calculus cheat sheet $\displaystyle\int_{-\infty}^{\infty} e^{-a(x+b)^2}\, dx = \sqrt{\frac{\pi}{a}} \quad (a > 0)$

https://en.wikipedia.org/wiki/Gaussian_integral

Calculus cheat sheet $\displaystyle\int_{-\infty}^{\infty} x^2 e^{-ax^2}\, dx = \frac{1}{2}\sqrt{\frac{\pi}{a^3}} \quad (a > 0)$

https://en.wikipedia.org/wiki/List_of_integrals_of_exponential_functions