11: Joint (Multivariate) Distributions

Lisa Yan

April 29, 2020

Quick slide reference

Normal Approximation

Normal RVs

mean variance

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

- Used to model many real-life situations because it maximizes entropy (i.e., randomness) for a given mean and variance

- Also useful for approximating the Binomial random variable!

Website testing

- 100 people are given a new website design.
- X = # people whose time on site increases
- The design actually has no effect, so P(time on site increases) = 0.5 independently.
- CEO will endorse the new design if $X \geq 65$.

What is P(CEO endorses change)? *Give a numerical approximation*.

Approach 1: Binomial

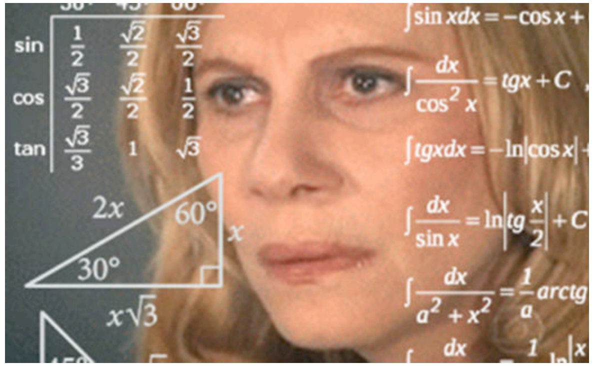
Define

$$X \sim \text{Bin}(n = 100, p = 0.5)$$

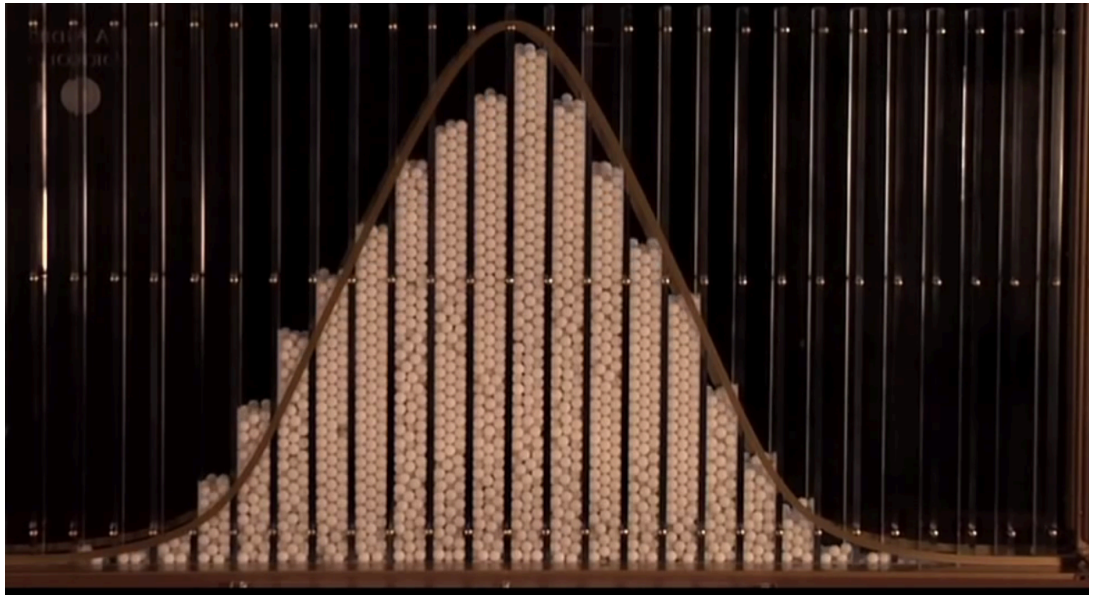
Want: $P(X \geq 65)$

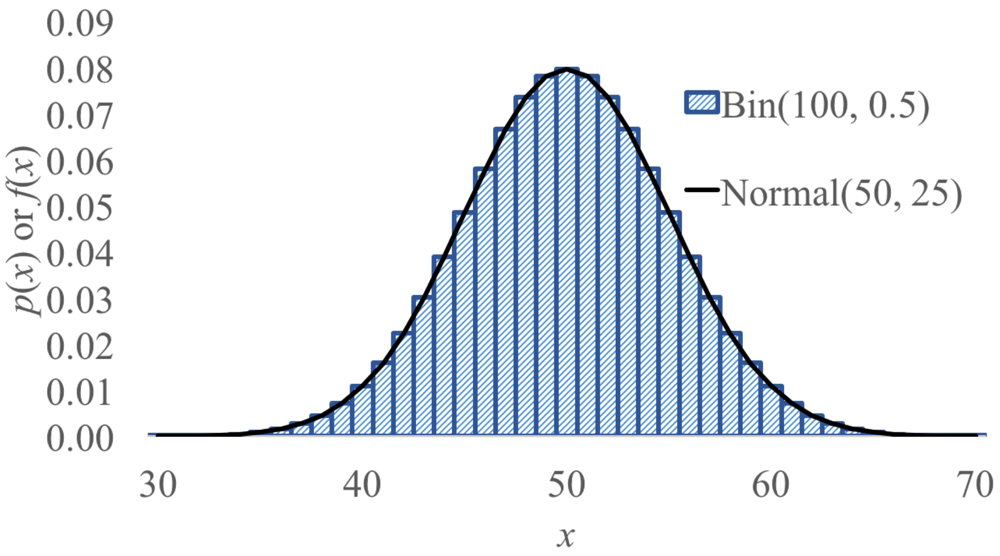
Solve

$$P(X \geq 65) = \sum_{i=65}^{100} \binom{100}{i} 0.5^i (1 - 0.5)^{100-i}$$



Don't worry, Normal approximates Binomial





Galton Board

(We'll explain *why* in 2 weeks' time)

Website testing

- 100 people are given a new website design.
- X = # people whose time on site increases
- The design actually has no effect, so P(time on site increases) = 0.5 independently.
- CEO will endorse the new design if $X \geq 65$.

What is $P(\text{CEO endorses change})$? *Give a numerical approximation*.

Approach 1: Binomial

Define

$X \sim \text{Bin}(n = 100, p = 0.5)$

Want: $P(X \geq 65)$

Solve

$P(X \geq 65) \approx 0.0018$

Approach 2: approximate with Normal

Define

mean variance

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

$\mu = np = 50$

$\sigma^2 = np(1-p) = 25$

$\sigma = \sqrt{25} = 5$

Solve

$P(X \geq 65) \approx P(Y \geq 65) = 1 - F_Y(65)$

$= 1 - \Phi\left(\frac{65-50}{5}\right) = 1 - \Phi(3) \approx 0.0013$?

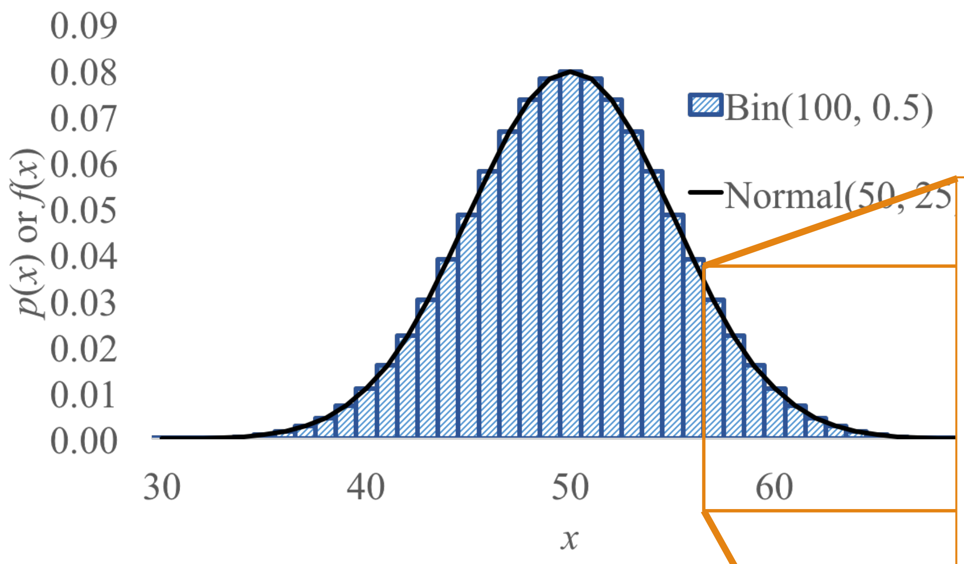
⚠️ ⚠️ (this approach is actually missing something)

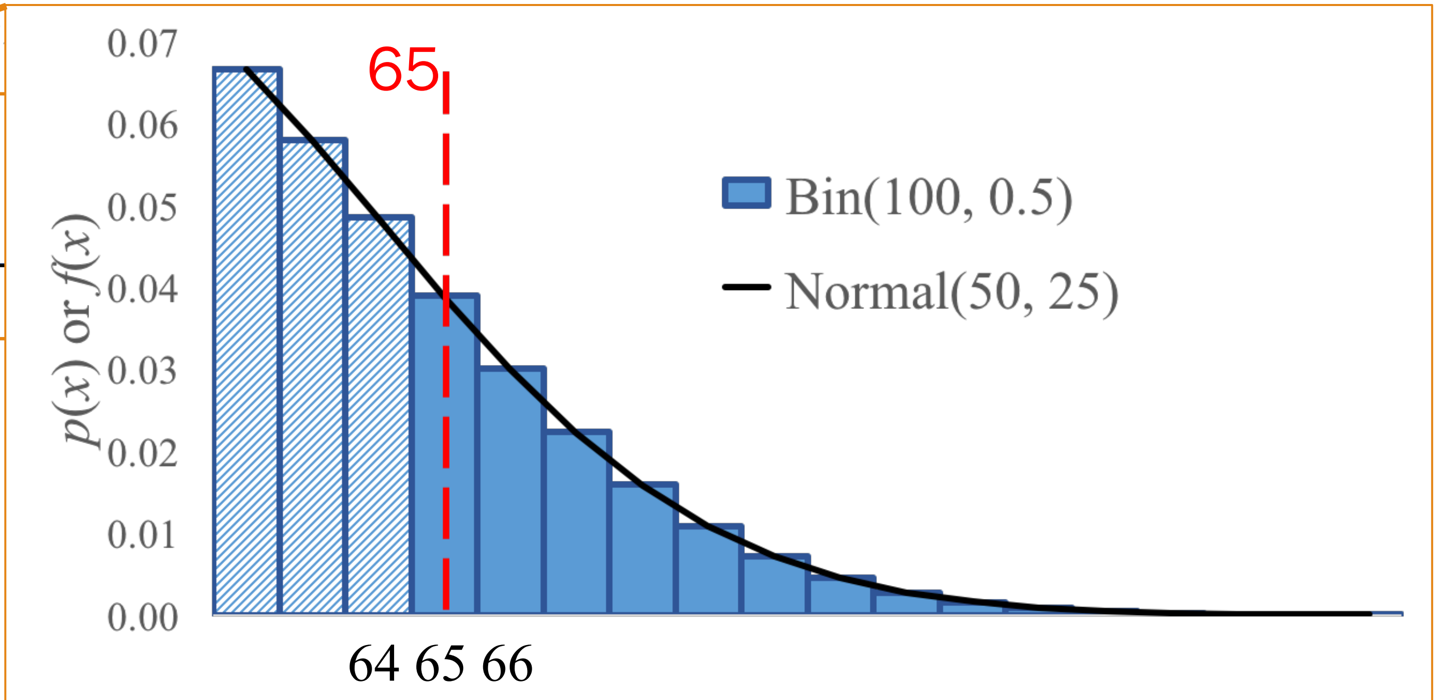
Website testing (with continuity correction)

In our website testing, $Y \sim \mathcal{N}(50, 25)$ approximates $X \sim \text{Bin}(100, 0.5)$.

$P(X = 65) \approx P(Y = 65) = 0$????

$\approx P(64.5 \le Y \le 65.5)$ ✓





$P(X \ge 65)$ Binomial

$\approx P(Y \ge 64.5)$ Normal

≈ 0.0018

the better

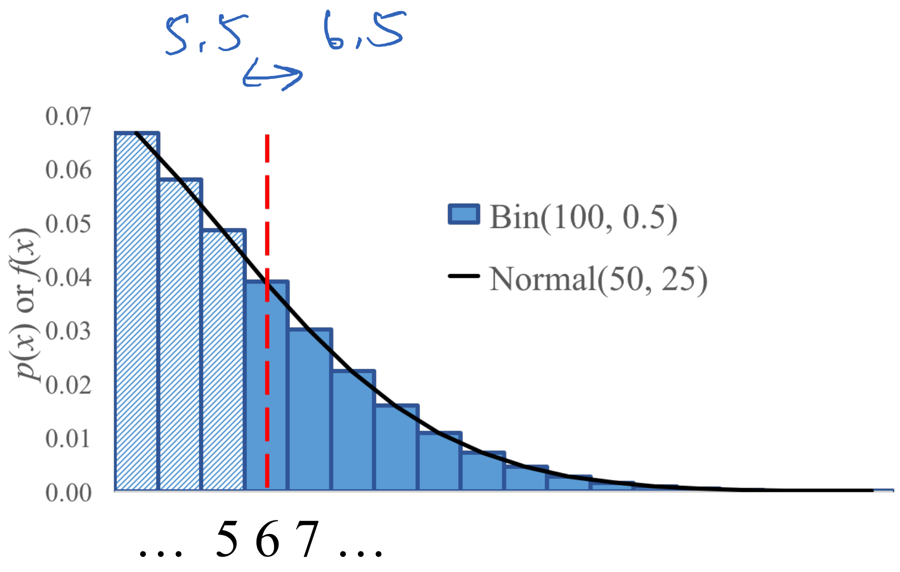
Approach 2

You must perform a **continuity correction** when approximating a Binomial RV with a Normal RV.

Continuity correction

If $Y \sim \mathcal{N}(np, np(1-p))$ approximates $X \sim \text{Bin}(n, p)$, how do we approximate the following probabilities?

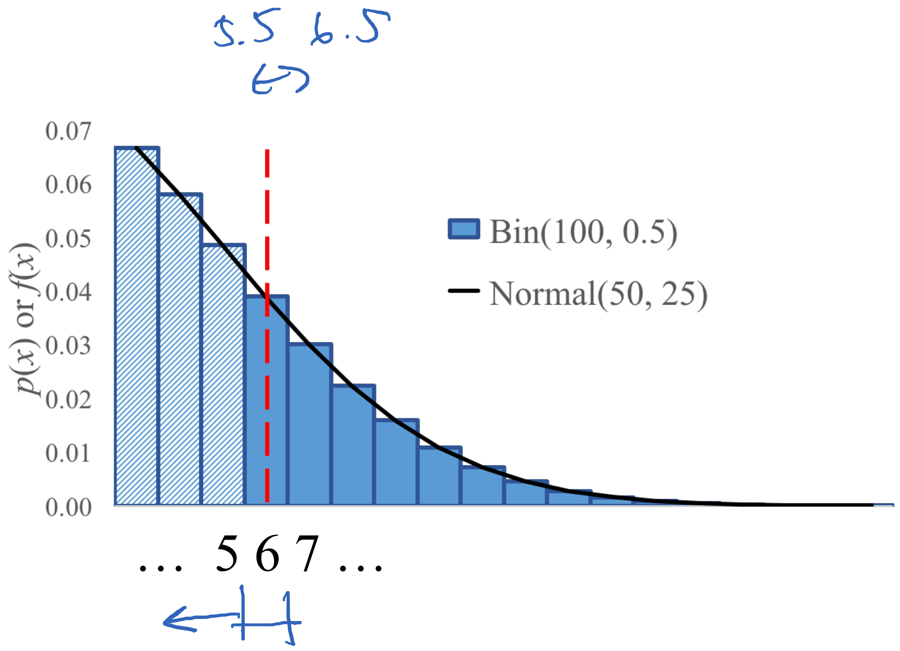
Discrete (e.g., Binomial) probability question	Continuous (Normal) probability question
$P(X = 6)$	
$P(X \geq 6)$	
$P(X > 6)$	
$P(X < 6)$	
$P(X \leq 6)$	



Continuity correction

If $Y \sim \mathcal{N}(np, np(1-p))$ approximates $X \sim \text{Bin}(n, p)$, how do we approximate the following probabilities?

Discrete (e.g., Binomial) probability question	Continuous (Normal) probability question
$P(X = 6)$	$P(5.5 \leq Y \leq 6.5)$
$P(X \geq 6)$	$P(Y \geq 5.5)$
$P(X > 6)$	$P(Y \geq 6.5)$
$P(X < 6)$	$P(Y \leq 5.5)$
$P(X \leq 6)$	$P(Y \leq 6.5)$



Who gets to approximate?

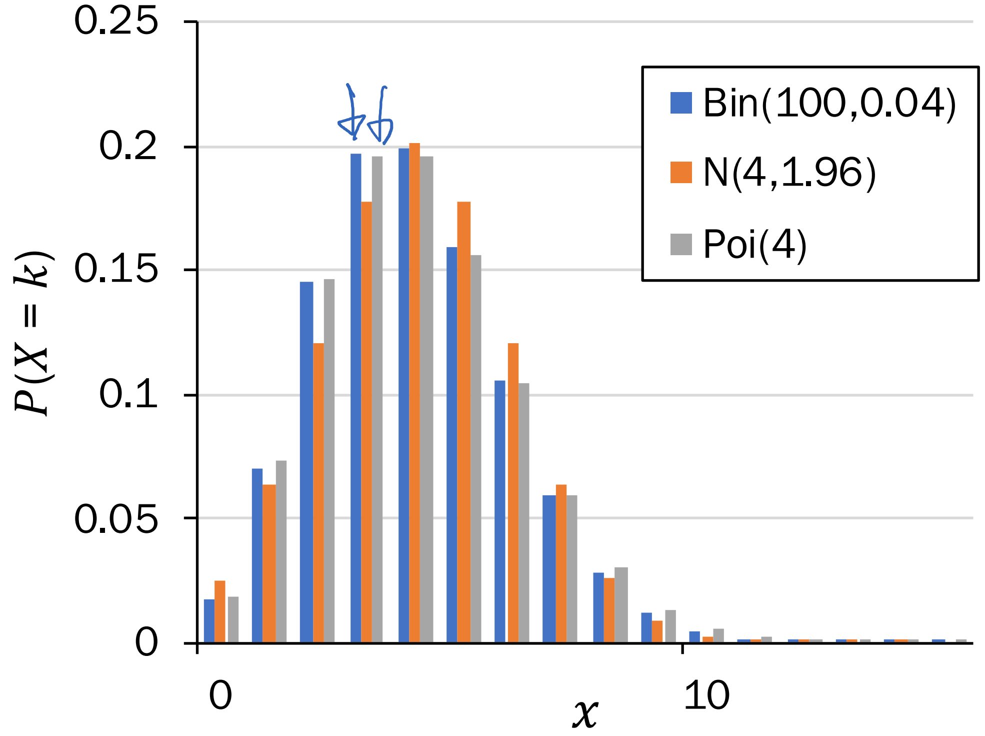
$$X \sim \text{Bin}(n, p)$$
$$E[X] = np$$
$$\text{Var}(X) = np(1-p)$$

$$Y \sim \text{Poi}(\lambda)$$
$$\lambda = np$$

?

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$
$$\mu = np$$
$$\sigma^2 = np(1-p)$$

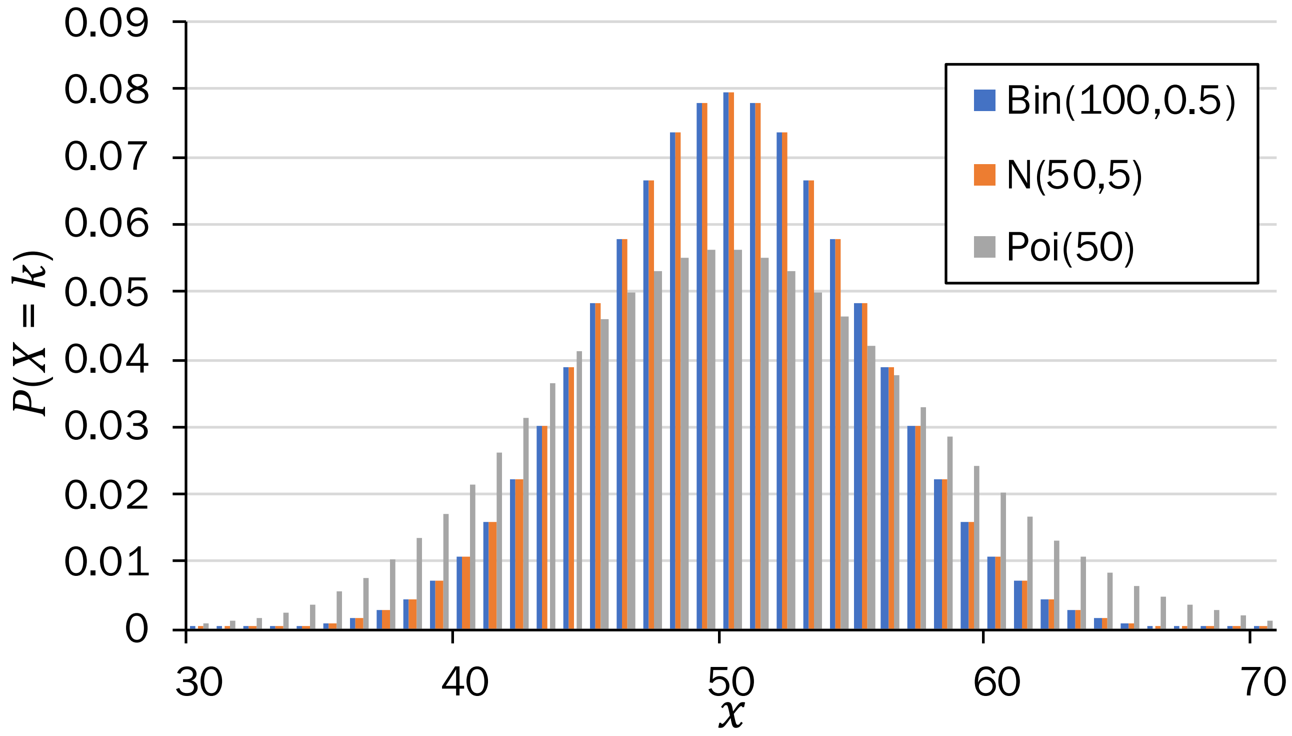
Who gets to approximate?



Poisson approximation

n large (> 20), p small (< 0.05)

slight dependence okay



Normal approximation

n large (> 20), p mid-ranged ($np(1-p) > 10$)

independence

1. If there is a choice, use Normal to approximate.
2. When using Normal to approximate a discrete RV, use a continuity correction.

Discrete Joint RVs

From last time

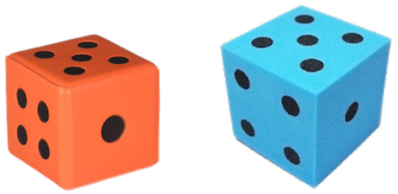
What is the probability that the Warriors win?
How do you model zero-sum games?

$$P(A_W > A_B)$$

This is a probability of an event involving ***two*** random variables!

Joint probability mass functions

Roll two 6-sided dice, yielding values X and Y.



X

random variable

$P(X = 1)$

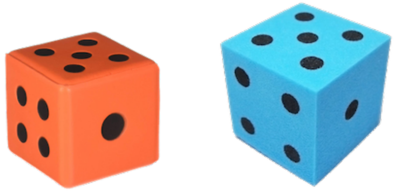
probability of an event

$P(X = k)$

probability mass function

Joint probability mass functions

Roll two 6-sided dice, yielding values X and Y.



X	$P(X = 1)$	$P(X = k)$
random variable	probability of an event	probability mass function
X, Y	$P(X = 1 \cap Y = 6)$ $P(X = 1, Y = 6)$ new notation: the comma	$P(X = a, Y = b)$
random variables	probability of the intersection of two events	joint probability mass function

Discrete joint distributions

For two discrete joint random variables X and Y, the joint probability mass function is defined as:

Valid joint PMFs

$$p_{X,Y}(a,b) = P(X = a, Y = b)$$

$$1 = \sum_a \sum_b p_{X,Y}(a,b)$$

joint PMF ↑ ↑ 2 RVs

The marginal distributions of the joint PMF are defined as:

$$p_X(a) = P(X = a) = \sum_y p_{X,Y}(a,y)$$

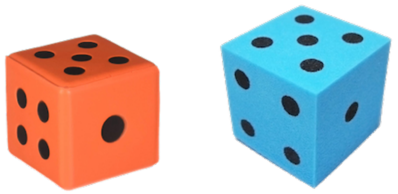
$$\rightarrow \sum_a p_X(a) = 1$$

$$p_Y(b) = P(Y = b) = \sum_x p_{X,Y}(x,b)$$

Use marginal distributions to get a 1-D RV from a joint PMF.

Two dice

Roll two 6-sided dice, yielding values X and Y.



1. What is the joint PMF of X and Y?

$$p_{X,Y}(a,b) = 1/36 \qquad (a,b) \in \{(1,1), \dots, (6,6)\}$$

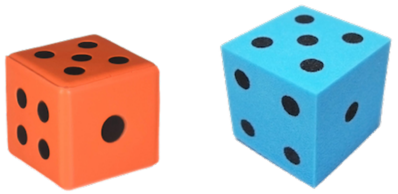
Y \ X	1	2	3	4	5	6
1	1/36	…	…	…	…	1/36
2	…	…	…	…	…	…
3	…	…	…	…	…	…
4	…	…	…	…	…	…
5	…	…	…	…	…	…
6	1/36	…	…	…	…	1/36

$P(X = 4, Y = 2)$

Probability table
- All possible outcomes for several discrete RVs
- Not parametric (e.g., parameter p in Ber(p))

Two dice

Roll two 6-sided dice, yielding values X and Y.



1. What is the joint PMF of X and Y?

$$p_{X,Y}(a,b) = 1/36 \qquad (a,b) \in \{(1,1),\dots,(6,6)\}$$

2. What is the marginal PMF of X?

$$p_X(a) = P(X=a) = \sum_y p_{X,Y}(a,y) = \sum_{y=1}^{6} \frac{1}{36} = \frac{1}{6} \qquad a \in \{1,\dots,6\}$$

$P(X=1) = P(X=1, Y=1) + \dots + P(X=1, Y=6)$

A computer (or three) in every house.

Consider households in Silicon Valley.

- A household has X Macs and Y PCs.
- Each house has a maximum of 3 computers (Macs + PCs) in the house.

1. What is $P(X = 1, Y = 0)$, the missing entry in the probability table?

Y (# PCs) \ X (# Macs)	0	1	2	3	
0	.16	?	.07	.04	
1	.12	.14	.12	0	
2	.07	.12	0	0	
3	.04	0	0	0	

A computer (or three) in every house.

Consider households in Silicon Valley.

- A household has X Macs and Y PCs.
- Each house has a maximum of 3 computers (Macs + PCs) in the house.

1. What is $P(X = 1, Y = 0)$, the missing entry in the probability table?

Y (# PCs) \ X (# Macs)	0	1	2	3	
0	.16	.12	.07	.04	
1	.12	.14	.12	0	
2	.07	.12	0	0	
3	.04	0	0	0	

A joint PMF must sum to 1:

$$\sum_x \sum_y p_{X,Y}(x, y) = 1$$

A computer (or three) in every house.

Consider households in Silicon Valley.

- A household has X Macs and Y PCs.
- Each house has a maximum of 3 computers (Macs + PCs) in the house.

2. How do you compute the marginal PMF of X?

X (# Macs)

Y (# PCs)	0	1	2	3	sum cols here
0 A	.16	.12	.07	.04	.39
1	.12	.14	.12	0	38
2	.07	.12	0	0	.19
3	.04	0	0	0	.04
B (sum rows here)	.39	.38	.19	.04	

C

A computer (or three) in every house.

Consider households in Silicon Valley.

- A household has X Macs and Y PCs.
- Each house has a maximum of 3 computers (Macs + PCs) in the house.

2. How do you compute the marginal PMF of X?

Y (# PCs) \ X (# Macs)	0	1	2	3	sum cols here
0 A	.16	.12	.07	.04	.39
1	.12	.14	.12	0	38
2	.07	.12	0	0	.19
3	.04	0	0	0	.04
B (sum rows here)	.39	.38	.19	.04	

C

A. $p_{X,Y}(x, 0) = P(X = x, Y = 0)$

B. Marginal PMF of X $\quad p_X(x) = \sum_y p_{X,Y}(x, y)$

C. Marginal PMF of Y $\quad p_Y(y) = \sum_x p_{X,Y}(x, y)$

To find a marginal distribution over one variable, sum over all other variables in the joint PMF.

A computer (or three) in every house.

Consider households in Silicon Valley.

- A household has X Macs and Y PCs.
- Each house has a maximum of 3 computers (Macs + PCs) in the house.

3. Let $C = X + Y$. What is $P(C = 3)$?

Y (# PCs) \ X (# Macs)	0	1	2	3	
0	.16	.12	.07	.04	
1	.12	.14	.12	0	
2	.07	.12	0	0	
3	.04	0	0	0	

A computer (or three) in every house.

Consider households in Silicon Valley.

- A household has X Macs and Y PCs.
- Each house has a maximum of 3 computers (Macs + PCs) in the house.

3. Let $C = X + Y$. What is $P(C = 3)$?

Y (# PCs) \ X (# Macs)	0	1	2	3
0	.16	.12	.07	.04
1	.12	.14	.12	0
2	.07	.12	0	0
3	.04	0	0	0

$$P(C = 3) = P(X + Y = 3)$$

1 or 0

Law of Total Probability

$$= \sum_x \sum_y P(X + Y = 3 | X = x, Y = y) P(X = x, Y = y)$$

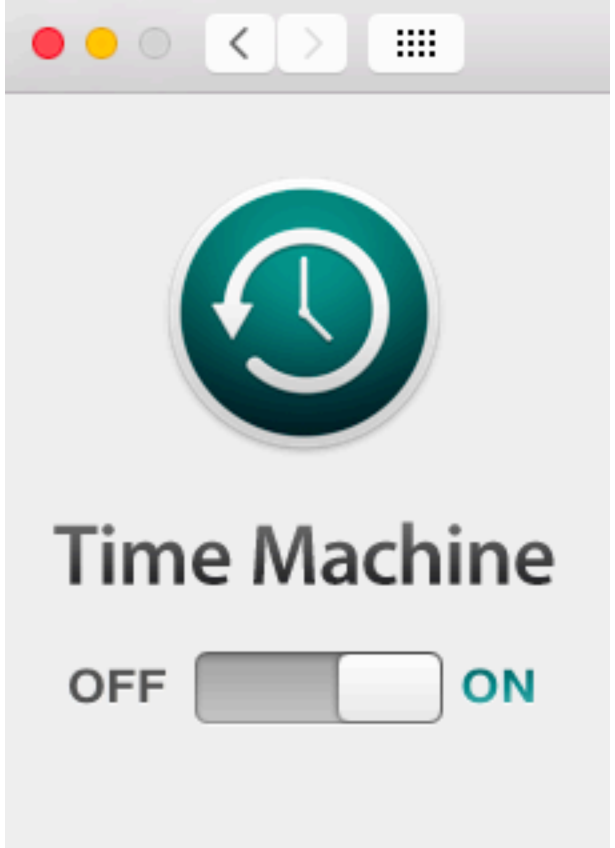
$$= P(X = 0, Y = 3) + P(X = 1, Y = 2) + P(X = 2, Y = 1) + P(X = 3, Y = 0)$$

$$= 0.32$$

We'll come back to sums of RVs next lecture!

Multinomial RV

Recall the good times



Permutations

$n!$

How many ways are there to order n objects?

Counting unordered objects

Binomial coefficient

How many ways are there to group n objects into two groups of size k and $n-k$, respectively?

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Called the binomial coefficient because of something from Algebra

Multinomial coefficient

How many ways are there to group n objects into r groups of sizes $n_1, n_2, \ldots, n_r$ respectively?

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1!\,n_2!\cdots n_r!}$$

Multinomials generalize Binomials for counting.

Probability

Binomial RV

What is the probability of getting k successes and $n - k$ failures in n trials?

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Binomial # of ways of ordering the successes

Probability of each ordering of k successes is equal + mutually exclusive

Multinomial RV

What is the probability of getting c_1 of outcome 1, c_2 of outcome 2, …, and c_m of outcome m in n trials?

Multinomial RVs also generalize Binomial RVs for probability!

Multinomial Random Variable

Consider an experiment of n independent trials:

- Each trial results in one of m outcomes. $P(\text{outcome } i) = p_i$, $\sum_{i=1}^{m} p_i = 1$
- Let X_i= # trials with outcome i

Joint PMF

$$P(X_1 = c_1, X_2 = c_2, \dots, X_m = c_m) = \binom{n}{c_1, c_2, \dots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$$

where $\sum_{i=1}^{m} c_i = n$ and $\sum_{i=1}^{m} p_i = 1$

Multinomial # of ways of ordering the outcomes

Probability of each ordering is equal + mutually exclusive

Hello dice rolls, my old friends

A 6-sided die is rolled 7 times.

What is the probability of getting:

- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes

Hello dice rolls, my old friends

A 6-sided die is rolled 7 times.

What is the probability of getting:

$X_1 = 1$ $X_3 = 0$

- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes $X_6 = 3$

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3)$$

$$= \binom{7}{1,1,0,2,0,3}\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^3 = 420\left(\frac{1}{6}\right)^7$$

Hello dice rolls, my old friends

A 6-sided die is rolled 7 times.

What is the probability of getting:

- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes

# of times a six appears

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3)$$

$$= \binom{7}{1,1,0,2,0,3} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^3 = 420 \left(\frac{1}{6}\right)^7$$

choose where the sixes appear

probability of rolling a six

this many times

(live)
11: Joint (Multivariate) Distributions

Lisa Yan

April 29, 2020

Normal RVs

mean variance

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

- Used to model many real-life situations because it maximizes entropy (i.e., randomness) for a given mean and variance
- Also useful for approximating the Binomial random variable!

Who gets to approximate?

$$X \sim \text{Bin}(n, p)$$
$$E[X] = np$$
$$\text{Var}(X) = np(1-p)$$

- Computing probabilities on Binomial RVs is often computationally expensive.
- Two reasonable approximations, but when to use which?

$$Y \sim \text{Poi}(\lambda)$$
$$\lambda = np$$

n large (> 20)
p small (< 0.05)
slight dependence okay

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$
$$\mu = np$$
$$\sigma^2 = np(1-p)$$

n large (> 20), p mid-ranged ($np(1-p) > 10$)
independence
need continuity correction

Think

Check out the question on the next slide (Slide 38). Post any clarifications here!

https://us.edstem.org/courses/109/discussion/46501

(by yourself)

Stanford Admissions (a while back)

Stanford accepts 2480 students.

- Each accepted student has 68% chance of attending (independent trials)
- Let X = # of students who will attend

What is $P(X > 1745)$? *Give a numerical approximation*.

Strategy:

A. Just Binomial
B. Poisson
C. Normal
D. None/other

🤨 (by yourself)

Stanford Admissions (a while back)

Stanford accepts 2480 students.

- Each accepted student has 68% chance of attending (independent trials)
- Let X = # of students who will attend

What is $P(X > 1745)$? *Give a numerical approximation*.

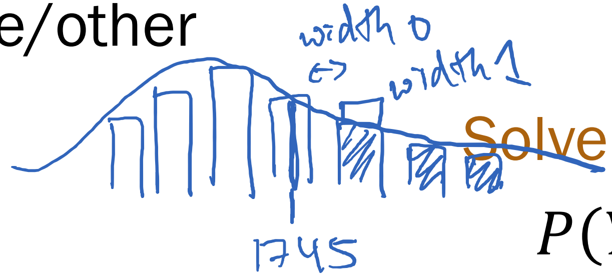
Strategy:

A. Just Binomial — not an approximation (also computationally expensive) $\binom{2480}{1746}$

B. Poisson — $p = 0.68$, not small enough

C. Normal ✅ Variance $np(1-p) = 540 > 10$

D. None/other



Define an approximation

Let $Y \sim \mathcal{N}(E[X], \text{Var}(X))$

$E[X] = np = 1686$

$\text{Var}(X) = np(1-p) \approx 540 \rightarrow \sigma = 23.3$

$P(X > 1745) \approx P(Y \geq 1745.5)$ ⚠️ Continuity correction

Solve

$$P(Y \geq 1745.5) = 1 - F(1745.5)$$

$$= 1 - \Phi\left(\frac{1745.5 - 1686}{23.3}\right)$$

$$= 1 - \Phi(2.54) \approx 0.0055$$

Changes in Stanford Admissions

Stanford accepts 2480 students.

- Each accepted student has 68% chance of attending (independent trials)
- Let X = # of students who will attend

What is $P(X > 1745)$? *Give a numerical approximation.*

Yield rate 20 years ago

The Stanford Daily

NEWS | SPORTS | OPINIONS | ARTS & LIFE | THE GRIND | MULTIMEDIA | FEATURES | ARCHIVES

Class of 2018 admit rates lowest in University history

March 28, 2014 16 Comments Tweet Like 901

Alex Zivkovic
Desk Editor

Stanford admitted 2,138 students to the Class of 2018 in this year's admissions cycle, producing – at 5.07 percent – the lowest admit rate in University history.

The University received a total of 42,167 applications this year, a record total and a 8.6 percent increase over last year's figure of 38,828. Stanford accepted 748 students

Overview for the Class of 2022

- Total Applicants: 47,451
- Total Admits: 2,071
- Total Enrolled: 1,706

Admit rate: 4.3%

Yield rate: 81.9%

People love coming to Stanford!

Multinomial Random Variable

Review

Consider an experiment of n independent trials:

- Each trial results in one of m outcomes. $P(\text{outcome } i) = p_i$, $\sum_{i=1}^{m} p_i = 1$
- Let X_i= # trials with outcome i

Joint PMF

$$P(X_1 = c_1, X_2 = c_2, \dots, X_m = c_m) = \binom{n}{c_1, c_2, \dots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$$

where $\sum_{i=1}^{m} c_i = n$ and $\sum_{i=1}^{m} p_i = 1$

Example:

- Rolling 2 twos, 3 threes, and 5 fives on 10 rolls of a fair-sided die
- Generating a random 5-word phrase with 1 “the”, 2 “bacon”, 1 “put”, 1 “on”

Hello dice rolls, my old friends

Review

A 6-sided die is rolled 7 times.

What is the probability of getting:

- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes

# of times a six appears

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3)$$

$$= \binom{7}{1,1,0,2,0,3} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^3 = 420 \left(\frac{1}{6}\right)^7$$

choose where the sixes appear

probability of rolling a six

this many times

Parameters of a Multinomial RV?

$X \sim \text{Bin}(n, p)$ has parameters n, p...

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

p: probability of success outcome on a single trial

A Multinomial RV has parameters $n, p_1, p_2, \ldots, p_m$ (Note $p_m = 1 - \sum_{i=1}^{m-1} p_i$)

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) = \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$$

p_i: probability of outcome i on a single trial

Where do we get p_i from?

$$\int \frac{d\,\text{cabin}}{\text{cabin}} \;?$$

log cabin

Interlude for jokes/announcements

Announcements

<u>Quiz #1</u>

Time frame: Thursday 4/30 12:00am-11:59pm PT

Covers: Up to end of Week 3 (including Lecture 9)

Info and practice: https://web.stanford.edu/class/cs109/exams/quizzes.html

24 hours

Thoughts pre-quiz:

- A checkpoint for <u>you</u>, not other people
- We are all here to learn. This exam was designed for a range of students.
- Typesetting will take a bit of time (total: ~2 hr + typeset)
- Take breaks, stretch, sleep
- The staff and I are here for <u>you</u>.

<u>Other things this week</u>

- Section optional (not graded), attend any section
- Friday's concept check #12 EC

Interesting probability news

Estimating Coronavirus Prevalence by Cross-Checking Countries

"We'll make the modeling assumption that N_{ij} is a **Poisson distribution with rate parameter** $A_{ij} * \lambda_i * \alpha_j$. What this means is that the **expected number of cases** should be equal to the total amount of travel, times some source-dependent multiplier α_j ..., times some country-dependent multiplier λ_i (the infection prevalence in country i)."

https://medium.com/@jsteinhardt/estimating-coronavirus-prevalence-by-cross-checking-countries-c7e4211f0e18

[CS109 Current Events Spreadsheet]

LIVE

The Federalist Papers

Probabilistic text analysis

Ignoring the order of words…

What is the probability of any given word that you write in English?

- $P(\text{word} = \text{“the”}) > P(\text{word} = \text{“pokemon”})$
- $P(\text{word} = \text{“Stanford”}) > P(\text{word} = \text{“Cal”})$

Probabilities of *counts* of words = Multinomial distribution

A document is a large multinomial.

(according to the Global Language Monitor, there are 988,968 words in the English language used on the internet.)

Probabilistic text analysis

Probabilities of *counts* of words = Multinomial distribution

Example document:

#words: $n = 48$

"When my late husband was alive he deposited some amount of Money with china Bank in which the amount will be declared to you once you respond to this message indicating your interest in helping to receive the fund and use it for Gods work as my wish."

$$P\left(\begin{array}{l}\text{bank} = 1\\ \text{fund} = 1\\ \text{money} = 1\\ \text{wish} = 1\\ \ldots\\ \text{to} = 3\end{array} \middle| \text{spam}\right) = \frac{n!}{1!\,1!\,1!\,1!\cdots 3!}\, p_{\text{bank}}^{1}\, p_{\text{fund}}^{1} \cdots p_{\text{to}}^{3}$$

Note: $P\left(\text{bank} \middle| \begin{array}{c}\text{spam}\\ \text{writer}\end{array}\right) \gg P\left(\text{bank} \middle| \begin{array}{c}\text{writer=}\\ \text{you}\end{array}\right)$

Old and New Analysis

Authorship of the Federalist Papers

- 85 essays advocating ratification of the US constitution
- Written under the pseudonym “Publius” (really, Alexander Hamilton, James Madison, John Jay)

Who wrote which essays?

- Analyze probability of words in each essay and compare against word distributions from known writings of three authors

Let’s write a program!

website demo

Probabilistic text analysis

Probabilities of *counts* of words = Multinomial distribution

What about probability of those same words in someone else's writing?

- $P\left(\text{word} = \text{“probability”} \,\middle|\, \begin{matrix}\text{writer =}\\ \text{you}\end{matrix}\right) > P\left(\text{word} = \text{“probability”} \,\middle|\, \begin{matrix}\text{writer =}\\ \text{non−CS109 student}\end{matrix}\right)$

To determine authorship:

1. Estimate $P(\text{word}|\text{writer})$ from known writings
2. Use Bayes' Theorem to determine $P(\text{writer}|\text{document})$ for a new writing!

madison.txt : get probs of Madison writing

unknown.txt

hamilton.txt : " Hamilton writing

$P(\text{Madison} \mid \text{unknown.txt})$

Who wrote the Federalist Papers?

$P_{\text{Congress, Madison}}$ $P_{\text{Congress, Hamilton}}$

Step 1. Generate probability lookups

$$\text{frequency}_{\text{congress, madison}} = \frac{\#\text{ times Congress appears}}{\#\text{ total words in madison.txt}}$$

only one document?

$$P_{\text{congress, madison}}? = \text{freq} \cdot K, \quad K = \frac{\#\text{ total words in madison.txt}}{\#\text{ words in English}}$$

Samplespace: English dictionary

freq: stored in madisonWordProb

$$P_{\text{congress, mad}} \propto \text{freq}$$

$\propto$

Step 1. Generate probability lookups

m_i Frequency of word i in Madison's writing, $\propto P(\text{word } i \mid \text{Madison})$

h_i Frequency of word i in Hamilton's writing, $\propto P(\text{word } i \mid \text{Hamilton})$

4. How will these values help us compute probabilities on a sentence being written by Hamilton or Madison?
    - "The People The Congress"
    - "People Congress The Rambutans"

5. [reach] Why don't the total numbers for just Madison add up to *exactly* one?

6. [reach] How does returning EPSILON for unknown words help us?

Step 1. Generate probability lookups

$P(\text{"The Congress The People"} \mid \text{Madison})$

$= P(\text{counts of words} \mid \text{Madison})$

$= \binom{4}{2,1,1} m_{\text{Congress}}^{1} \, m_{\text{the}}^{2} \, m_{\text{people}}^{1} \, m_{\text{state}}^{0} \, m_1^{0} \, m_2^{0}$

$P(\text{"People Congress The Rambutans"} \mid \text{Hamilton})$

$= \binom{4}{1,1,1,1} h_{\text{ppl}}^{1} \, h_{\text{cong}}^{1} \, h_{\text{the}}^{1} \, \varepsilon^{1}$

$\uparrow$

$10^{-6}, P(\text{rambutan} \mid \text{Hamilton})$

Step 2. Unknown document counts

2. How would you represent the probability of Madison writing this document with a Multinomial? Let c_i be the count of word i.

$$\binom{n = 2170}{c_1, c_2, \ldots, c_m} m_1^{c_1} m_2^{c_2} \cdots = P\left(\text{these document counts } c_1, c_2, \ldots, c_m \mid \text{Madison}\right)$$

(m ↑ # unique words)

Step 3. Bayes' Theorem

$$P(\text{Madison}|\text{unknownDoc}) = \frac{P(\text{unknownDoc}|\text{Madison})P(\text{Madison})}{P(\text{unknownDoc})} \qquad \text{(Bayes)}$$

Assume that $P(\text{writer}) = 0.5$. We can rewrite this into a decision:

$$\frac{P(\text{unknownDoc}|\text{Madison})}{P(\text{unknownDoc}|\text{Hamilton})} > 1 \qquad \text{(If true, Madison is writer)}$$

$$P(M|D) > P(H|D) \Rightarrow \frac{P(M|D)}{P(H|D)} > 1$$

$$\Rightarrow \frac{\frac{P(D|M)P(M)}{P(D)}}{\frac{P(D|H)P(H)}{P(D)}} \Rightarrow \frac{P(D|M)}{P(D|H)} > 1$$

if true, report Madison.

Step 3. Bayes' Theorem

Assume that $P(\text{writer}) = 0.5$. We can rewrite this into a decision:

$$\frac{P(\text{unknownDoc}|\text{Madison})}{P(\text{unknownDoc}|\text{Hamilton})} > 1$$

(If true, Madison is writer)

$$\frac{\binom{2170}{c_1, c_2, c_3, \dots} m_1^{c_1} m_2^{c_2} \cdots m_m^{c_m}}{\binom{2170}{c_1, c_2, c_3, \dots} h_1^{c_1} h_2^{c_2} \cdots h_m^{c_m}} > 1$$

$$\Rightarrow \frac{\prod_{i=1}^{m} m_i^{c_i}}{\prod_{i=1}^{m} h_i^{c_i}} > 1$$

$$\log\left(\frac{\prod_{i=1}^{m} m_i^{c_i}}{\prod_{i=1}^{m} h_i^{c_i}}\right) > \log(1)$$

$$\log\left(\prod_{i=1}^{m} m_i^{c_i}\right) - \log\left(\prod_{i=1}^{m} h_i^{c_i}\right) > 0$$

$$\sum_{i=1}^{m} c_i \log m_i - \sum_{i=1}^{m} c_i \log h_i > 0$$