11: Joint (Multivariate) Distributions

Jerry Cain
February 2, 2024
Normal Approximation
Normal Random Variables

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

- 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 presented with a new website design.
- $X = \#$ people whose time on site increases
- PM assumes design has no effect, so assume $P($stickier$) = 0.5$ 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

$$P(X \geq 65) = \sum_{k=65}^{100} \binom{100}{k} 0.5^k (1 - 0.5)^{100-k}$$
Don’t worry, Normal approximates Binomial

Galton Board

(We’ll explain why in 2 weeks)
Website testing

- 100 people are given a new website design.
- \( X = \# \) people whose time on site increases
- PM assumes design has no effect, so \( P(\text{stickier}) = 0.5 \) 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

\[ P(X \geq 65) \approx 0.0018 \]

⚠⚠⚠ (this approach is missing something important)

**Approach 2: approximate with Normal**

Define

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

\[ \mu = np = 50 \]
\[ \sigma^2 = np(1 - p) = 25 \]
\[ \sigma = \sqrt{25} = 5 \]

Solve

\[ P(X \geq 65) \approx P(Y \geq 65) = 1 - F_Y(65) \]
\[ = 1 - \Phi \left( \frac{65 - 50}{5} \right) = 1 - \Phi(3) \approx 0.0013 \]

(jerry$ python
>>> from scipy.stats import binom, norm
>>> binom.pmf(range(65, 101), n, p).sum()
0.001758820861485058
>>> 1 - norm(50, 5).cdf(65)
0.0013498980316301035
Website testing (with continuity correction)

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

$P(X \geq 65) \quad \text{Binomial}$

$\approx P(Y \geq 64.5) \quad \text{Normal}$

$\approx 0.0018 \quad \square$ 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?

<table>
<thead>
<tr>
<th>Discrete (e.g., Binomial) probability question</th>
<th>Continuous (Normal) probability question</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X = 6) )</td>
<td></td>
</tr>
<tr>
<td>( P(X \geq 6) )</td>
<td></td>
</tr>
<tr>
<td>( P(X &gt; 6) )</td>
<td></td>
</tr>
<tr>
<td>( P(X &lt; 6) )</td>
<td></td>
</tr>
<tr>
<td>( P(X \leq 6) )</td>
<td></td>
</tr>
</tbody>
</table>
Continuity correction

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

<table>
<thead>
<tr>
<th>Discrete (e.g., Binomial) probability question</th>
<th>Continuous (Normal) probability question</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X = 6) )</td>
<td>( P(5.5 \leq Y \leq 6.5) )</td>
</tr>
<tr>
<td>( P(X \geq 6) )</td>
<td>( P(Y \geq 5.5) )</td>
</tr>
<tr>
<td>( P(X &gt; 6) )</td>
<td>( P(Y \geq 6.5) )</td>
</tr>
<tr>
<td>( P(X &lt; 6) )</td>
<td>( P(Y \leq 5.5) )</td>
</tr>
<tr>
<td>( P(X \leq 6) )</td>
<td>( P(Y \leq 6.5) )</td>
</tr>
</tbody>
</table>
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?

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

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
Stanford Admissions (a while back)

Stanford accepts 2480 students.
- Each admitted student matriculates with \( p = 0.68 \) (independently)
- Let \( X = \# \) of students who will attend

What is \( \mathbb{P}(X > 1745) \)? *Give a numerical approximation.*

Strategy:
- A. Just Binomial
- B. Poisson
- C. Normal
- D. None/other
Stanford Admissions (a while back)

Stanford accepts 2480 students.

- Each admitted student matriculates with $p = 0.68$ (independently)
- Let $X =$ # of students who will attend

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

Strategy:

A. Just Binomial  
   computationally expensive (also not an approximation)
B. Poisson  
   $p = 0.68$, not small enough
C. Normal  
   ✓ Variance $np(1 - p) = 540 > 10$
D. None/other

Define an approximation

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

- $E[X] = np = 1686$
- $\text{Var}(X) = np(1 - p) \approx 540 \rightarrow \sigma = 23.3$

Solve

- $P(Y \geq 1745.5) = 1 - F(1745.5)$
- $= 1 - \Phi\left(\frac{1745.5 - 1686}{23.3}\right)$
- $= 1 - \Phi(2.54) \approx 0.0055$

⚠ Continuity correction

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Winter 2024
Discrete Joint RVs
From last slide deck

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 $Y$.

<table>
<thead>
<tr>
<th>$X$</th>
<th>$P(X = 1)$</th>
<th>$P(X = k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>random variable</td>
<td>probability of an event</td>
<td>probability mass function</td>
</tr>
</tbody>
</table>
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

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

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

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

The marginal distributions of the joint PMF are defined as:

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

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

Use marginal distributions to extract a 1D 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 $Y$?

$$p_{X,Y}(a, b) = \frac{1}{36} \quad (a, b) \in \{(1,1), \ldots, (6,6)\}$$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1/36</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1/36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1/36</td>
</tr>
</tbody>
</table>

$P(X = 4, Y = 3)$

Probability table
- All possible outcomes for several discrete RVs
- Not parametric (e.g., parameter $p$ in $\text{Ber}(p)$)
Two dice

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

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

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

2. What is the marginal PMF of $X$?

$$p_X(a) = P(X = a) = \sum_y p_{X,Y}(a, y) = \sum_{y=1}^{6} \frac{1}{36} = \frac{1}{6} \quad a \in \{1, \ldots, 6\}$$
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 total (Macs + PCs).

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

<table>
<thead>
<tr>
<th>$Y$ (# PCs)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.16</td>
<td>?</td>
<td>.07</td>
<td>.04</td>
<td>.03</td>
</tr>
<tr>
<td>1</td>
<td>.12</td>
<td>.14</td>
<td>.12</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.07</td>
<td>.12</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Consider households in Silicon Valley.

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

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

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.16</td>
<td>.12</td>
<td>.07</td>
<td>.04</td>
</tr>
<tr>
<td>1</td>
<td>.12</td>
<td>.14</td>
<td>.12</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>.07</td>
<td>.12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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 total (Macs + PCs).

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

<table>
<thead>
<tr>
<th></th>
<th>$X$ (# Macs)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>.16</td>
<td>.12</td>
<td>.07</td>
<td>.04</td>
</tr>
<tr>
<td>1</td>
<td>.12</td>
<td>.14</td>
<td>.12</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>.07</td>
<td>.12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Sum rows here: 0.39, 0.38, 0.19, 0.04

Sum cols here: 0.39, 0.38, 0.19, 0.04
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 total (Macs + PCs).

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

<table>
<thead>
<tr>
<th>$Y$ (# PCs)</th>
<th>$X$ (# Macs)</th>
<th>[0]</th>
<th>[1]</th>
<th>[2]</th>
<th>[3]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0]</td>
<td>A</td>
<td>.16</td>
<td>.12</td>
<td>.07</td>
<td>.04</td>
</tr>
<tr>
<td>[1]</td>
<td></td>
<td>.12</td>
<td>.14</td>
<td>.12</td>
<td>.00</td>
</tr>
<tr>
<td>[2]</td>
<td></td>
<td>.07</td>
<td>.12</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>[3]</td>
<td></td>
<td>.04</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
</tbody>
</table>

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

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

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

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 total (Macs + PCs).

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

<table>
<thead>
<tr>
<th>$X$ (# Macs)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$ (# PCs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.16</td>
<td>.12</td>
<td>.07</td>
<td>.04</td>
</tr>
<tr>
<td>1</td>
<td>.12</td>
<td>.14</td>
<td>.12</td>
<td>.00</td>
</tr>
<tr>
<td>2</td>
<td>.07</td>
<td>.12</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>3</td>
<td>.04</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
</tbody>
</table>
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 total (Macs + PCs).

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

<table>
<thead>
<tr>
<th>$Y$ (# PCs)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.16</td>
<td>.12</td>
<td>.07</td>
<td>.04</td>
</tr>
<tr>
<td>1</td>
<td>.12</td>
<td>.14</td>
<td>.12</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>.07</td>
<td>.12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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

Law of Total Probability

$$= \sum_{x} \sum_{y} P(X + Y = 3 | X = x, Y = y) P(X = x, Y = y)$$

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

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?

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

**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!
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**

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

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 fair, six-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 fair, six-sided die is rolled 7 times. What is the probability of getting:
- 1 one
- 0 threes
- 0 fives
- 1 two
- 2 fours
- 3 sixes

\[
P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3) = \binom{7}{1,1,0,2,0,3} \left( \frac{1}{6} \right)^1 \left( \frac{1}{6} \right)^1 \left( \frac{1}{6} \right)^0 \left( \frac{1}{6} \right)^2 \left( \frac{1}{6} \right)^0 \left( \frac{1}{6} \right)^3 = 420 \left( \frac{1}{6} \right)^7
\]
Hello dice rolls, my old friends

A fair, six-sided die is rolled 7 times.
What is the probability of getting:

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

\[ P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3) \]

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

\[ = 420 \left( \frac{1}{6} \right)^7 \]
Probabilistic text analysis

Ignoring the order of words...

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

- $P(\text{word} = "\text{the}" ) > P(\text{word} = "\text{susurration}" )$
- $P(\text{word} = "\text{Stanford}" ) > P(\text{word} = "\text{Cal}" )$

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:

"When my late husband was alive he deposited some amount of Money with overseas 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 Heavens work as my wish."

\[ P \left( \text{spam} \right) = \frac{48!}{1! 1! 1! 1! \cdots 3!} p_\text{bank}^1 p_\text{fund}^1 \cdots p_\text{to}^3 \]

Note: \( P \left( \text{bank} \mid \text{spam writer} = \text{you} \right) \gg P \left( \text{bank} \mid \text{writer} = \text{you} \right) \)
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
• Curious what the analysis is? Read this!
Statistics of Two RVs
Expectation and Covariance

In real life, we often have many RVs interacting at once.

- We’ve seen some simpler cases (e.g., sum of independent Bernoullis).
- Come Monday, we’ll discuss sums of Binomials, Poissons, etc.
- In general, manipulating joint PMFs is difficult.
- Fortunately, **you don’t need to model** joint RVs completely all the time.

Instead, we’ll focus next on reporting **statistics** of multiple RVs:

- **Expectation of sums** (you’ve seen some of this, more on Monday)
- **Covariance**: measure of how two random variable vary with each other (more next Wednesday and Friday)
Properties of Expectation, extended to two RVs

1. Linearity:
   \[ E[aX + bY + c] = aE[X] + bE[Y] + c \]

2. Expectation of a sum = sum of expectation:
   \[ E[X + Y] = E[X] + E[Y] \]
   (we’ve seen this; we’ll prove today!)

3. Unconscious statistician:
   \[ E[g(X,Y)] = \sum_x \sum_y g(x,y)p_{X,Y}(x,y) \]
   True for both independent and dependent random variables!
Proof of expectation of a sum of RVs

\[ E[X + Y] = \sum_{x} \sum_{y} (x + y)p_{X,Y}(x, y) \]

\[ = \sum_{x} \sum_{y} xp_{X,Y}(x, y) + \sum_{x} \sum_{y} yp_{X,Y}(x, y) \]

\[ = \sum_{x} x \sum_{y} p_{X,Y}(x, y) + \sum_{y} y \sum_{x} p_{X,Y}(x, y) \]

\[ = \sum_{x} xp_{X}(x) + \sum_{y} yp_{Y}(y) \]

\[ = E[X] + E[Y] \]