Hypergeometric, Poisson & Joint Distributions
Sec 4.7 - 4.9

Cathy Poliak, Ph.D.
cathy@math.uh.edu
Office in Fleming 11c

Department of Mathematics
University of Houston

Lecture 6 - 3339
Outline

1. Hypergeometric Distribution
2. Poisson Distribution
3. Joint Distribution
Each of 12 refrigerators of a certain type has been returned to a distributor because of an audible, high-pitched, oscillating noise when the refrigerator is running. Suppose that 7 of these refrigerators have a defective compressor and the other 5 have less serious problems. The technition looks at 6 refrigerators, what is the probability that exactly 5 have a defective compressor?

\[
P(5 \text{ bad and 1 good}) = \frac{\binom{7}{5} \binom{5}{1}}{\binom{12}{6}} = 0.1136
\]
Conditions for a Hypergeometric Distribution

1. The population or set to be sampled consists of \( N \) individuals, objects or elements (a finite population).
2. Each individual can be characterized as a "success" or "failure." There are \( m \) successes in the population, and \( n \) failures in the population. Notice: \( m + n = N \).
3. A sample size of \( k \) individuals is selected without replacement in such a way that each subset of size \( k \) is equally likely to be chosen.

The parameters of a hypergeometric distribution is \( m, n, k \). We write \( X \sim \text{Hyper}(m, n, k) \). The probability mass function for a hypergeometric is:

\[
f_X(x) = P(X = x) = \binom{m}{x} \binom{n}{k-x} / \binom{m+n}{k} = \frac{C(m, x)C(n, k-x)}{C(m+n, k)}
\]

Example: \( m=7, n=5, k=6, x=5 \):

\[
f_X(5) = \frac{C(7, 5)C(5, 1)}{C(12, 6)}
\]
Using R

- R commands: $P(X = x) = \text{dhyper}(x, m, n, k)$ and $P(X \leq x) = \text{phyper}(x, m, n, k)$

- Example going back to the refrigerator example, $m = 7$, $n = 5$, $k = 6$. $P(X = 5)$
  
  ```r
  > dhyper(5, 7, 5, 6)
  [1] 0.1136364
  ```

- $P(X \leq 4)$
  
  ```r
  > phyper(4, 7, 5, 6)
  [1] 0.8787879
  ```

- $P(X \geq 5) = 1 - P(X \leq 4)$
  
  
  $= 1 - 0.8788$
  
  $= 0.1212$
Mean and Variance of a Hypergeometric Distribution

Let $Y$ have a hypergeometric distribution with parameter, $m, n,$ and $k$.

- The mean of $Y$ is:
  \[ \mu_Y = E(Y) = k \left( \frac{m}{m+n} \right) = kp. \]

- The variance of $Y$ is:
  \[ \sigma^2_Y = \text{var}(Y) = kp(1 - p) \left( 1 - \frac{k - 1}{m+n-1} \right). \]

- $1 - \frac{k-1}{m+n-1}$ is called the finite population correction factor. As the population increases, this factor will get closer to 1.
A small voting district has 101 female voters and 95 male voters. A random sample of 10 voters is drawn.

1. What is the probability exactly 7 of the voters will be female?

\[
P(X = 7) = \text{hypergeom}(7, 101, 95, 10) = 0.1304
\]

2. What is the probability that at most 7 of the voters will be female?

\[
P(X \leq 7) = \text{hypergeom}(7, 101, 95, 10) = 0.9387
\]

3. What is the probability that at least 7 of the voters will be female?

\[
P(X \geq 7) = 1 - P(X \leq 6) = 1 - \text{hypergeom}(6, 101, 95, 10) = 0.1917
\]
Example

Suppose that the average number of emails received by a particular student at UH is five emails per hour.

We want to know what is the probability that a particular student will get $X$ number of emails in any given hour.

Suppose we want to know $P(X = 3)$, the probability of getting exactly 3 emails in one hour.

What about in two hours?

$\mu = 5$ per hour

$X = \# \text{ of emails received}$
The Poisson Distribution

The Poisson distribution is appropriate under the following conditions.

1. Let $X$ be the number of successes occurring per unit of measure. $X = 0, 1, 2, 3, \ldots$

2. Let $\mu$ be the mean number of successes occurring per unit of measure.

3. The number of successes that occur in two non-overlapping units of measure are independent.

4. The probability that success will occur in a unit of measure is the same for all units of equal size and is proportional to the size of the unit. $\mu = 5$ per hour $\Rightarrow \mu = 2(5) = 10$ per two hours

5. The probability that more than one event occurs in a unit of measure is negligible for very small-sized units. In other words, the events occur one at a time.
A random variable $X$ with non-negative integer values has a Poisson distribution if its frequency function is:

$$f(x) = P(X = x) = e^{-\mu} \frac{\mu^x}{x!}$$

for $x = 0, 1, 2, \ldots$, where $\mu > 0$ is a constant. If $X$ has a Poisson distribution with parameter $\mu$, we can write $X \sim \text{Pois}(\mu)$.

**Example:** $\mu = 5$ emails per hour

$$P(X=3) = e^{-5} \left( \frac{5^3}{3!} \right)$$

```r
> exp(-5)*5^3/factorial(3)
[1] 0.1403739
> dpois(3,5)
[1] 0.1403739
```
Let $X \sim \text{Pois}(\mu)$

- The mean of $X$ is $\mu$ per unit of measure. By the conditions of the Poisson distribution.

- The variance of $X$ is also $\mu$ per unit of measure.

- The standard deviation of $X$ is $\sqrt{\mu}$. 
Example

The number of people arriving for treatment at an emergency room can be modeled by a Poisson process with a mean of five people per hour.

1. What is the probability that exactly four arrivals occur at a particular hour?

\[ P(X = 4) = \text{dpois}(4, 5) = 0.1754 \]

2. What is the probability that at least four people arrive during a particular hour?

\[ P(X \geq 4) = 1 - P(X \leq 3) = 1 - \text{dpois}(3, 5) \]

\[ = 0.735 \]

3. How many people do you expect to arrive during a 45-min period?

\[ E(X) = 5 \cdot \left( \frac{45}{60} \right) = 3.75 \]
Using R to Compute Probabilities of the Poisson Distribution

R commands: \( P(X = x) = \text{dpois}(x, \mu) \) and \( P(X \leq x) = \text{ppois}(x, \mu) \)

- \( P(X = 4) \)
  
  \[
  > \text{dpois}(4, 5) \\
  [1] 0.1754674
  \]

- \( P(X \geq 4) = 1 - P(X \leq 3) \)
  
  \[
  > 1 - \text{ppois}(3, 5) \\
  [1] 0.7349741
  \]
Review Questions

For the following information indicate the type of distribution.


1. A company makes sports bikes. 90% pass final inspection (and 10% fail and need to be fixed). We randomly select 10 bikes and want to know what is the probability that 7 will pass.

   Binomial \( n=10, \ p=0.9, \ \text{independent} \)

2. A company makes sports bikes. 90% pass final inspection (and 10% fail and need to be fixed). What is the probability that the 10\( ^{th} \) bike will not pass final inspection?

   Geometric \( n \text{ is not fixed} \)

3. A small company made 100 bikes in a month. 10 of these bikes will need to be fixed. Suppose we select 4 bikes, what is the probability that at least one needs to be fixed?

   Hypergeometric \( m=10, \ n=90, \ k=4 \ \text{P}(x \geq 1) \)

4. A small company makes bicycles. The average amount that this company makes in a month is 100 bikes. What is the probability that the company will make less than 30 bikes in a week?

   Poisson \( \mu=100 \text{ per month} \ \text{P}(X<30)\text{pois} \)
\[ P(x < 30) = P(x \leq 29) = \text{ppois}(29, 100 \times \frac{1}{4}) \]
Beginning Example

Suppose two random variables $X$ and $Y$ have a joint function $f(x, y) = \frac{2x-y}{5}$ for $X = 0.75$ and 1 and $Y = 0$ and 1. What are the possible values of $f(x, y)$?

\[
\begin{align*}
  f(0.75, 0) &= \frac{2(0.75)-0}{5} = 0.3 \\
  f(0.75, 1) &= \frac{2(0.75)-1}{5} = 0.4 \\
  f(1, 0) &= \frac{2(1)-0}{5} = 0.1 \\
  f(1, 1) &= \frac{2(1)-1}{5} = 0.2
\end{align*}
\]
Joint Function Table

<table>
<thead>
<tr>
<th></th>
<th>0.75</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

\[ f(x, y) = \text{joint probability} \]
\[ f_x(x) \text{ or } f_y(y) = \text{ marginal probabilities} \]
Joint Probability Distribution For Discrete Random Variables

The function $f(x, y)$ is a **joint probability distribution** of the discrete random variables $X$ and $Y$ if

1. $f(x, y) \geq 0$ for all $(x, y)$.
2. $\sum_x \sum_y f(x, y) = 1$.
3. $P(X = x, Y = y) = f(x, y) = P(X=x \land Y=y)$

For any region in the $xy$ plane, $P[(X, Y) \in A] = \sum_A \sum f(x, y)$. 
Joint Probability Distribution

Let $X$ and $Y$ be discrete random variables that have the joint probability distribution $f(x, y)$. Then

1. $f_Y(y) = \sum_x f(x, y)$ for all $y$ is the marginal probability mass function of $Y$.

2. $f_X(x) = \sum_y f(x, y)$ for all $x$ is the marginal probability mass function of $X$.

3. $f_{Y|X}(y|x) = \frac{f(x,y)}{f(x)}$ if $f_X(x) > 0$. This is the conditional frequency function.

4. $X$ and $Y$ are independent if and only if $f(x, y) = f_X(x)f_Y(y)$ for all $x, y$.

5. $\sum_x \sum_y f(x, y) = 1$. 

Cathy Poliak, Ph.D. cathy@math.uh.edu Office in Fleming 11c (Department of Mathematics University of Houston)
Example 2

Suppose that the following table represents the joint pmf of X and Y.

<table>
<thead>
<tr>
<th>Y/X</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2/k</td>
<td>3/k</td>
</tr>
<tr>
<td>2</td>
<td>3/k</td>
<td>4/k</td>
</tr>
<tr>
<td>3</td>
<td>4/k</td>
<td>5/k</td>
</tr>
</tbody>
</table>

1. What should be the value of $k$?
\[
\frac{2}{k} + \frac{3}{k} + \frac{3}{k} + \frac{4}{k} + \frac{4}{k} + \frac{5}{k} = 1
\]

2. Determine $P(X = 2, Y \geq 2)$
\[
P(X = 2, Y = 2) + P(X = 2, Y = 3) = \frac{4}{21} + \frac{5}{21} = \frac{9}{21} = 0.4285
\]

3. Determine $P(X = 2 | Y = 2)$
\[
\frac{P(X = 2, Y = 2)}{P(Y = 2)} = \frac{4/21}{7/21} = \frac{4}{7} = 0.5714
\]
Rule for Means

- If $X$ and $Y$ are two different random variables, then the mean of the sums of the pairs of the random variable is the same as the sum of their means:

$$E[X + Y] = E[X] + E[Y].$$

This is called the addition rule for means.

- The mean of the difference of the pairs of the random variable is the same as the difference of their means:

$$E[X - Y] = E[X] - E[Y].$$
Rule for Variances

If $X$ and $Y$ are independent random variables

$$\sigma_{X+Y}^2 = \text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y].$$

and

$$\sigma_{X-Y}^2 = \text{Var}[X - Y] = \text{Var}[X] + \text{Var}[Y].$$

$$\text{Var}(X - Y) = \text{Var}(X + (-1)Y)$$

$$= \text{Var}(X) + \text{Var}(-1Y)$$

$$= \text{Var}(X) + (-1)^2 \text{Var}(Y)$$

$$= \text{Var}(X) + \text{Var}(Y).$$
Example of Rules

Tamara and Derek are sales associates in a large electronics and appliance store. The following table shows their mean and standard deviation of daily sales. Assume that daily sales among the sales associates are independent.

<table>
<thead>
<tr>
<th>Sales associate</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamara $X$</td>
<td>$E[X] = 1100$</td>
<td>$\sigma_X = 100$</td>
</tr>
<tr>
<td>Derek $Y$</td>
<td>$E[Y] = 1000$</td>
<td>$\sigma_Y = 80$</td>
</tr>
</tbody>
</table>

1. Determine the mean of the total of Tamara’s and Derek’s daily sales, $E[X + Y]$.
   \[ E[X + Y] = 1100 + 1000 = 2100 \]

2. Determine the variance of the total of Tamara’s and Derek’s daily sales, $\sigma^2_{(X+Y)}$.
   \[ \sigma^2_{(X+Y)} = 10000 + 6400 = 16400 \]

3. Determine the standard deviation of the total of Tamara’s and Derek’s daily sales, $\sigma_{(X+Y)}$.
   \[ \sigma_{(X+Y)} = \sqrt{16400} = 128.6425 \]
Let $X$ and $Y$ be two random variables, then:

- $E(X + Y) = E(X) + E(Y)$

- $E(XY) = x_1 y_1 p_{11} + x_1 y_1 p_{12} + \ldots + x_n y_n p_{nn}$

- For more than two random variables $X_1, X_2, X_3, \ldots, X_n$, $E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$
Covariance of $X$ and $Y$

Let $X$ and $Y$ be jointly distributed random variables with respective means $\mu_x$ and $\mu_y$ and standard deviation $\sigma_x$, $\sigma_y$. The **covariance** of $X$ and $Y$ is

$$\text{cov}(X, Y) = E((x - \mu_x)(Y - \mu_y)).$$

or an easier calculation:

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y).$$

The **correlation** of $X$ and $Y$ is;

$$\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}.$$
Properties of Covariance

1. \( \text{cov}(X, Y) = \text{cov}(Y, X) \)

2. \( \text{cov}(X, X) = \text{var}(X) \)

3. If \( X, Y, \) and \( Z \) are jointly distributed and \( a \) and \( b \) are constants
   \[
   \text{cov}(X, aY + bZ) = a[\text{cov}(X, Y)] + b[\text{cov}(X, Z)].
   \]

4. If \( X \) and \( Y \) are jointly distributed,
   \[
   \text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cor}(X, Y)\text{sd}(X)\text{sd}(Y)
   \]
   \[
   \text{var}(X - Y) = \text{var}(X) + \text{var}(Y) - 2\text{cor}(X, Y)\text{sd}(X)\text{sd}(Y)
   \]

5. If \( X \) and \( Y \) are independent, \( \text{cov}(X, Y) = 0. \)

6. If jointly distributed random variables \( X_1, X_2, \cdots, X_n \) are pairwise uncorrelated, then
   \[
   \text{var}(X_1 + X_2 + \cdots + X_n) = \text{var}(X_1) + \text{var}(X_2) + \cdots + \text{var}(X_n)
   \]
\[
\text{Cov}(x, y) = \mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y)
\]

\[
= 0.275 - 0.9(0.3) = 0.005
\]

\[
\text{var}(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2
\]

\[
\text{Cor}(x, y) = \frac{\text{Cov}(x, y)}{\text{SD}(x)\text{SD}(y)}
\]
\[
\begin{array}{c|cc}
X & 0.75 & 1 \\
\hline
y \quad 0 & 0.3 & 0.4 & 0.7 \\
1 & 0.1 & 0.2 & 0.3 \\
\hline
0.4 & 0.4
\end{array}
\]

\[
E(X) = 0.75(0.4) + 1(0.4) = 0.9 \\
E(Y) = 0(0.7) + 1(0.3) = 0.3 \\
E(X+Y) = 0.9 + 0.3 = 1.2 \\
E(XY) = 0(0.75)(0.3) + 0(1)(0.4) \\
+ 1(0.75)(0.1) + 1(1)(0.2) = 0.275 \\
\ne E(X)E(Y)
\]
\[ x = c(0.75, 1, 0.75, 1) \]
\[ y = c(0, 0, 1, 1) \]
\[ f_{xy} = (2x - y)/5 \]
\[ \text{cbind}(x, y, f_{xy}) \]
\[ \begin{array}{ccc}
1 & 0.75 & 0.3 \\
2 & 1.00 & 0.4 \\
3 & 0.75 & 1.01 \\
4 & 1.00 & 0.2 \\
\end{array} \]
\[ \sum(x^2)f_{xy} \]
\[ [1] \ 0.825 = E(x^2) \]
\[ \sum(y^2)f_{xy} \]
\[ [1] \ 0.3 = E(y^2) \]
\[ \text{sdx} = 0.835 - 0.9^2 \]
\[ \text{sdx} = 0.855 - 0.9^2 \]
\[ \text{sdy} = 0.3 - 0.3^2 \]
\[ \text{cor}_{xy} = 0.005 / (\text{sdx} \times \text{sdy}) \]
\[ \text{cor}_{xy} \]
\[ [1] \ 0.5291005 \]
\[ \text{cov}(x, y) = E(xy) - E(x)E(y) \]
Example of a Joint Probability Distribution

The following table is a joint probability table of $X =$ number of sports involved and $Y =$ age of high school students.

<table>
<thead>
<tr>
<th>$Y =$ Age</th>
<th>$X =$ Number of sports involved</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0.03</td>
</tr>
<tr>
<td>15</td>
<td>0.015</td>
</tr>
<tr>
<td>16</td>
<td>0.045</td>
</tr>
<tr>
<td>17</td>
<td>0.03</td>
</tr>
<tr>
<td>18</td>
<td>0.03</td>
</tr>
</tbody>
</table>

1. Determine $E(X)$.
2. Determine $E(Y)$.
3. Determine $E(XY)$.
4. Determine $cov(X, Y)$.
R code

I put the values in an Excel .csv file named "jointprob."

```r
> library(readr)
> jointprob <- read_csv("F:/Lectures uh/MATH 3339/jointprob.csv")
> View(jointprob)
> sum(jointprob$x*jointprob$y*jointprob$pxy)
[1] 21.735
> 1.35*16.1
[1] 21.735
> sum(jointprob$x*jointprob$pxy)
[1] 1.35
> sum(jointprob$y*jointprob$pxy)
[1] 16.1
> sum(jointprob$x*jointprob$y*jointprob$pxy)
[1] 21.735

\[
\text{Corr}(X,Y) = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y} = \frac{21.735 - 1.35 \cdot 16.1}{\sigma_X \sigma_Y} = 0 = \text{independent.}
\]
You Try Question

Suppose we have two random variables, $X$ and $Y$ with frequency functions, $f_X(x)$, $f_Y(y)$ and joint frequency function $f_{X,Y}(x,y)$. If $X$ and $Y$ are independent, which statement is false.

a) $f_{X,Y}(x,y) = 0$

b) $f_{X,Y}(x,y) = f_X(x)f_Y(y)$

c) $f_{X|Y}(x|y) = f_X(x)$

d) $\text{cov}(X, Y) = 0$

e) $E(XY) = E(X)E(Y)$

\[ P(X \cap Y) = 0 \Rightarrow P(X \cap Y) = P(X)P(Y) \]
Example of Joint Probability Distribution

Suppose that a fair, 6 sided die is rolled. Let $X$ indicate the event that an even number is rolled (in other words, $X = 1$ if an even number is rolled and $X = 0$ otherwise). Let $Y$ indicate the event that 3, 4, or 5 is rolled (in other words, $Y = 1$ if 3, 4, or 5 is rolled and $Y = 0$ otherwise). Find $P(X = 0, Y = 1)$.

\[
P(X = 0 \cap Y = 1) = P(\{1, 3, 4, 5\} \cap \{3, 4, 5\}) = P(\{3, 5\}) = P(3 \text{ or } 5) = \frac{2}{6} = 0.33
\]
A class has 10 mathematics majors, 6 computer science majors and 4 statistics majors. A committee of two is selected at random to work an a problem. Let \( X \) be the number of mathematics majors and let \( Y \) be the number of computer science majors chosen.

1. Determine all possible \((X, Y)\) pairs for randomly selecting two students.

\[ (0,0), (0,1), (0,2), (1,0), (1,1), (2,0) \]

2. Determine \( f_{X,Y}(0, 0) \), that is the probability that the two pick are neither a mathematics major nor a computer science major.

\[
f_{X,Y}(0, 0) = \frac{C(10,0) \times C(6,0) \times C(4,2)}{C(20,2)} = \frac{6}{190}
\]

3. Determine \( f_X(0) \), that is the probability that we do not pick any mathematics majors.

\[
f_X(0) = \frac{C(10,0) \times C(10,0)}{C(20,2)} = \frac{45}{190}
\]
1. Determine $E(X)$.  
   $$E(X) = \frac{45}{190} \approx 0.6882$$

2. Determine $E(Y)$.  
   $$E(Y) = \frac{60}{190} \approx 0.6380$$

3. Determine $E(XY)$.  
   $$E(XY) = \frac{45}{190}$$

4. Determine $cov(X, Y)$.  
   $$cov(X, Y) = \frac{60}{190} - \frac{45}{190} \times \frac{45}{190} = -0.2842$$

5. Determine $cor(X, Y)$.  
   $$cor(X, Y) = \frac{-0.2842}{(0.6882) \times (0.6380)}$$