Multiple random variables, Correlations
What we learned so far...

• **Random Events:**
  – Working with events as sets: union, intersection, etc.
    • Some events are simple: Head vs Tails, Cancer vs Healthy
    • Some are more complex: $10 < \text{Gene expression} < 100$
    • Some are even more complex: Series of dice rolls: 1,3,5,3,2
  – Conditional probability: $P(A \mid B) = P(A \cap B)/P(B)$
  – Independent events: $P(A \mid B) = P(A)$ or $P(A \cap B) = P(A) \cdot P(B)$
  – Bayes theorem: relates $P(A \mid B)$ to $P(B \mid A)$

• **Random variables:**
  – Mean, Variance, Standard deviation. How to work with $E(g(X))$
  – Discrete (Uniform, Bernoulli, Binomial, Poisson, Geometric, Negative binomial, Power law);
    PMF: $f(x) = \text{Prob}(X=x)$; CDF: $F(x) = \text{Prob}(X \leq x)$;
  – Continuous (Uniform, Exponential, Erlang, Gamma, Normal, Log-normal);
    PDF: $f(x)$ such that $\text{Prob}(X \text{ inside } A) = \int_A f(x)dx$; CDF: $F(x) = \text{Prob}(X \leq x)$

• **Next step:** work with multiple random variables measured together in the same series of random experiments
Concept of Joint Probabilities

• Biological systems are usually described not by a single random variable but by many random variables
• Example: The expression state of a human cell: 20,000 random variables $X_i$ for each of its genes
• A joint probability distribution describes the behavior of several random variables
• We will start with just two random variables $X$ and $Y$ and generalize when necessary
Joint Probability Mass Function Defined

The joint probability mass function of the discrete random variables $X$ and $Y$, denoted as $f_{XY}(x, y)$, satisfies:

1. $f_{XY}(x, y) = P$
2. $f_{XY}(x, y) \geq 0$ \hspace{1cm} All probabilities are non-negative
3. $\sum_x \sum_y f_{XY}(x, y) = 1$ \hspace{1cm} The sum of all probabilities is 1

Montgomery Runger 5th edition Equation (5–1)
Example 5-1: # Repeats vs. Signal Bars

You use your cell phone to check your airline reservation. It asks you to speak the name of your departure city to the voice recognition system.

• Let $Y$ denote the number of times you have to state your departure city.
• Let $X$ denote the number of bars of signal strength on your cell phone.

<table>
<thead>
<tr>
<th>$y$ = number of times city name is stated</th>
<th>$x$ = number of bars of signal strength</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**Figure 5-1** Joint probability distribution of $X$ and $Y$. The table cells are the probabilities. Observe that more bars relate to less repeating.
Marginal Probability Distributions (discrete)

For a discrete joint PDF, there are marginal distributions for each random variable, formed by summing the joint PMF over the other variable.

\[
    f_X(x) = \sum_y f_{XY}(x, y)
\]

\[
    f_Y(y) = \sum_x f_{XY}(x, y)
\]

Called **marginal** because they are written in the margins

**Figure 5-6** From the prior example, the joint PMF is shown in green while the two marginal PMFs are shown in purple.
Mean & Variance of X and Y are calculated using marginal distributions

<table>
<thead>
<tr>
<th>y = number of times city name is stated</th>
<th>x = number of bars of signal strength</th>
<th>f(y)</th>
<th>y*f(y)</th>
<th>y^2*f(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.01</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.02</td>
<td>0.25</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.02</td>
<td>0.17</td>
<td>0.51</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.15</td>
<td>0.30</td>
<td>1.20</td>
</tr>
</tbody>
</table>

\[ f(x) = \begin{array}{c}
0.20 \\
0.25 \\
0.55 \\
\end{array} \]

\[ x*f(x) = \begin{array}{c}
0.20 \\
0.50 \\
1.65 \\
\end{array} \]

\[ x^2*f(x) = \begin{array}{c}
0.20 \\
1.00 \\
4.95 \\
\end{array} \]

\[ \mu_X = E(X) = 2.35; \quad \sigma_X^2 = V(X) = 6.15 - 2.35^2 = 6.15 - 5.52 = 0.6275 \]

\[ \mu_Y = E(Y) = 2.49; \quad \sigma_Y^2 = V(Y) = 7.61 - 2.49^2 = 7.61 - 16.20 = 1.4099 \]

Sec 5-1.2 Marginal Probability Distributions
Conditional Probability Distributions

Recall that $P(B|A) = \frac{P(A \cap B)}{P(A)}$

$$P(Y=y|X=x)=\frac{P(X=x,Y=y)}{P(X=x)}=\frac{f(x,y)}{f_X(x)}$$

From Example 5-1

$P(Y=1|X=3) = 0.25/0.55 = 0.455$

$P(Y=2|X=3) = 0.20/0.55 = 0.364$

$P(Y=3|X=3) = 0.05/0.55 = 0.091$

$P(Y=4|X=3) = 0.05/0.55 = 0.091$

Sum = 1.00

Note that there are 12 probabilities conditional on $X$, and 12 more probabilities conditional upon $Y$. 
Joint Random Variable Independence

- Random variable independence means that knowledge of the value of $X$ does not change any of the probabilities associated with the values of $Y$.

- Opposite: Dependence implies that the values of $X$ are influenced by the values of $Y$. 
Independence for Discrete Random Variables

• Remember independence of events (slide 13 lecture 4): Events are independent if any one of the three conditions are met:
  1) \( P(A | B) = P(A \cap B) / P(B) = P(A) \) or
  2) \( P(B | A) = P(A \cap B) / P(A) = P(B) \) or
  3) \( P(A \cap B) = P(A) \cdot P(B) \)

• Random variables independent if all events \( A \) that \( Y=y \) and \( B \) that \( X=x \) are independent if any one of these conditions is met:
  1) \( P(Y=y | X=x) = P(Y=y) \) for any \( x \) or
  2) \( P(X=x | Y=y) = P(X=x) \) for any \( y \) or
  3) \( P(X=x, Y=y) = P(X=x) \cdot P(Y=y) \) for every pair \( x \) and \( y \)
X and Y are Bernoulli variables

<table>
<thead>
<tr>
<th></th>
<th>Y=0</th>
<th>Y=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=0</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>X=1</td>
<td>2/6</td>
<td>1/6</td>
</tr>
</tbody>
</table>

What is the marginal $P_Y(Y=0)$?

A. 1/6
B. 2/6
C. 3/6
D. 4/6
E. I don’t know

Get your i-clickers
**X and Y are Bernoulli variables**

<table>
<thead>
<tr>
<th></th>
<th>Y=0</th>
<th>Y=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=0</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>X=1</td>
<td>2/6</td>
<td>1/6</td>
</tr>
</tbody>
</table>

What is the conditional $P(X=0|Y=1)$?

A. 2/6
B. 1/2
C. 1/6
D. 4/6
E. I don’t know

Get your i-clickers
X and Y are Bernoulli variables

<table>
<thead>
<tr>
<th></th>
<th>Y=0</th>
<th>Y=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=0</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>X=1</td>
<td>2/6</td>
<td>1/6</td>
</tr>
</tbody>
</table>

Are they independent?

A. yes
B. no
C. I don’t know

Get your i-clickers
X and Y are Bernoulli variables

<table>
<thead>
<tr>
<th></th>
<th>Y=0</th>
<th>Y=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=0</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>X=1</td>
<td>0</td>
<td>1/2</td>
</tr>
</tbody>
</table>

Are they independent?

A. yes  
B. no  
C. I don’t know

Get your i-clickers
Joint Probability Density Function Defined

The joint probability density function for the continuous random variables $X$ and $Y$, denotes as $f_{XY}(x,y)$, satisfies the following properties:

1. $f_{XY}(x,y) \geq 0$ for all $x, y$

2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) \, dx \, dy = 1$

3. $P((X, Y) \in R) = \int_{R} f_{XY}(x,y) \, dx \, dy \quad (5-2)$

Figure 5-2 Joint probability density function for the random variables $X$ and $Y$. Probability that $(X, Y)$ is in the region $R$ is determined by the volume of $f_{XY}(x,y)$ over the region $R$. 

Sec 5-1.1 Joint Probability Distributions 16
Figure 5-3 Joint probability density function for the continuous random variables $X$ and $Y$ of expression levels of two different genes. Note the asymmetric, narrow ridge shape of the PDF – indicating that small values in the $X$ dimension are more likely to occur when small values in the $Y$ dimension occur.
Marginal Probability Distributions (continuous)

• Rather than summing a discrete joint PMF, we integrate a continuous joint PDF.
• The marginal PDFs are used to make probability statements about one variable.
• If the joint probability density function of random variables $X$ and $Y$ is $f_{XY}(x,y)$, the marginal probability density functions of $X$ and $Y$ are:

$$f_X(x) = \int f_{XY}(x,y) \, dy$$
$$f_Y(y) = \int f_{XY}(x,y) \, dx$$

$$f_X(x) = \sum_y f_{XY}(x,y)$$
$$f_Y(y) = \sum_x f_{XY}(x,y)$$
Conditional Probability Density Function Defined

Given continuous random variables $X$ and $Y$ with joint probability density function $f_{XY}(x, y)$, the conditional probability density function of $Y$ given $X=x$ is

$$f_{Y|X}(y) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{f_{XY}(x, y)}{\int_y f_{XY}(x, y) \, dy} \quad \text{if} \quad f_X(x) > 0 \quad (5-4)$$

which satisfies the following properties:

1. $f_{Y|X}(y) \geq 0$

2. $\int f_{Y|X}(y) \, dy = 1$

3. $P(Y \in B | X = x) = \int_B f_{Y|X}(y) \, dy$ for any set $B$ in the range of $Y$

Compare to discrete: $P(Y=y|X=x)=f_{XY}(x,y)/f_X(x)$

Sec 5-1.3 Conditional Probability Distributions
Conditional Probability Distributions

• Conditional probability distributions can be developed for multiple random variables by extension of the ideas used for two random variables.

• Suppose \( p = 5 \) and we wish to find the distribution of \( X_1, X_2 \) and \( X_3 \) conditional on \( X_4=x_4 \) and \( X_5=x_5 \).

\[
f_{x_1x_2x_3|x_4x_5}(x_1, x_2, x_3) = \frac{f_{x_1x_2x_3x_4x_5}(x_1, x_2, x_3, x_4, x_5)}{f_{x_4x_5}(x_4, x_5)}
\]
for \( f_{x_4x_5}(x_4, x_5) > 0 \).
Independence for Continuous Random Variables

For random variables $X$ and $Y$, if any one of the following properties is true, the others are also true. Then $X$ and $Y$ are independent.

1. $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$
2. $f_{Y|X}(y) = f_Y(y)$ for all $x$ and $y$ with $f_X(x) > 0$
3. $f_{X|Y}(y) = f_X(x)$ for all $x$ and $y$ with $f_Y(y) > 0$
4. $P(X \subset A, Y \subset B) = P(X \subset A) \cdot P(Y \subset B)$ for any sets $A$ and $B$ in the range of $X$ and $Y$, respectively.

$P(Y=y|X=x)=P(Y=y)$ for any $x$ or
$P(X=x|Y=y)=P(X=x)$ for any $y$ or
$P(X=x, Y=y)=P(X=x)\cdot P(Y=y)$ for any $x$ and $y$
Example 1:
Uniform distribution in the square
$-1 < x < 1, \ -1 < y < 1$

\[
\begin{cases}
  f_{X,Y}(x,y) = c \quad \text{if} \quad -1 < x < 1 \quad \text{and} \quad -1 < y < 1 \\
  0 \quad \text{otherwise}
\end{cases}
\]

\[1 = \int_{-1}^{1} \int_{-1}^{1} f_{X,Y}(x,y) \, dx \, dy = c \cdot \text{Area} = c \cdot 4 \implies c = \frac{1}{4}\]
Are $X$ and $Y$ independent? Yes they are

Let's test if $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) \, dy = \int_{-1}^{1} \frac{1}{4} \, dy = \frac{1}{2} \text{ if } -1 < x < 1$$

Same for $f_Y(y) = \frac{1}{2} \text{ if } -1 < y < 1$

$$\frac{1}{4} = f_{XY}(x, y) = \frac{1}{2} \cdot \frac{1}{2} = f_X(x) \cdot f_Y(y)$$

0 otherwise if both $x$ and $y$ are in $[-1, 1]$
X and Y are uniformly distributed in the disc $x^2+y^2 \leq 1$

Are they independent?

A. yes
B. no
C. I could not figure it out

Get your i-clickers
Joint PDF \( f_{xy}(x, y) = \frac{1}{\text{area}} = \frac{1}{\pi} \) if \( x, y \) in the disc

Marginal distributions:

\[
f_x(x) = \int_{-\infty}^{\infty} f_{xy}(x, y) \, dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{dy}{\pi} = \frac{2\sqrt{1-x^2}}{\pi}
\]

Same for \( f_y(y) = \frac{2\sqrt{1-y^2}}{\pi} \)

\[
\frac{1}{\pi} = f_{xy}(x, y) \neq \frac{2}{\pi} \sqrt{1-x^2} \cdot \frac{2}{\pi} \sqrt{1-y^2} = f_x(x) \cdot f_y(y)
\]

Variables are **not** independent