STEVENS INSTITUTE of TECHNOLOGY
School of Business

BIA-652

Multivariate Data Analytics

Bivariate Random Variables

Prof. Feng Mai
School of Business

Joint, Marginal, and Conditional Distributions

BIA 652, Multivariate Data Analytics

Jointly Distributed Random Variables

Usually some kind of association between the variables. E.g., two different financial assets

Joint CDF for two random variables

$F(x, y) = \text{Prob}(X \leq x, Y \leq y)$

Two Discrete Random Variables

- The probability mass function (pmf) of a single discrete rv *X* specifies how much probability mass is placed on each possible *X* value.

  The joint pmf of two discrete rv's *X* and *Y* describes how much probability mass is placed on each possible pair of values (*x*, *y*).

  Definition
  Let *X* and *Y* be two discrete rv's defined on the sample space of an experiment. The **joint probability mass function** $p(x, y)$ is defined for each pair of numbers (x, y) by

$$p(x, y) = P(X = x \text{ and } Y = y)$$

$f(x, y)$

It must be the case that $p(x, y) \geq 0$ and $\sum_x \sum_y \underbrace{p(x, y)} = 1$.

Definition

The **marginal probability mass function of *X*,** denoted by $p_X(x)$, is given by

$$p_X(x) = \sum_{y:\, p(x,\, y)>0} p(x, y) \text{ for each possible value } x$$

Similarly, the **marginal probability mass function of *Y*** is

$$p_Y(y) = \sum_{x:\, p(x,\, y)>0} p(x, y) \text{ for each possible value } y.$$

Two Continuous Random Variables

- The probability that the observed value of a continuous rv X lies in a one-dimensional set A (such as an interval) is obtained by integrating the pdf $f(x)$ over the set A.

  Similarly, the probability that the pair (X, Y) of continuous rv's falls in a two-dimensional set A (such as a rectangle) is obtained by integrating a function called the *joint density function*.

Two Continuous Random Variables

- **Definition**
- Let *X* and *Y* be continuous rv's. **A joint probability density function** *f* (*x*, *y*) for these two variables is a function satisfying *f* (*x*, *y*) ≥ 0 and

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

- Then for any two-dimensional set *A*

$$P[(X, Y) \in A] = \iint_A f(x, y)\, dx\, dy$$

Two Continuous Random Variables

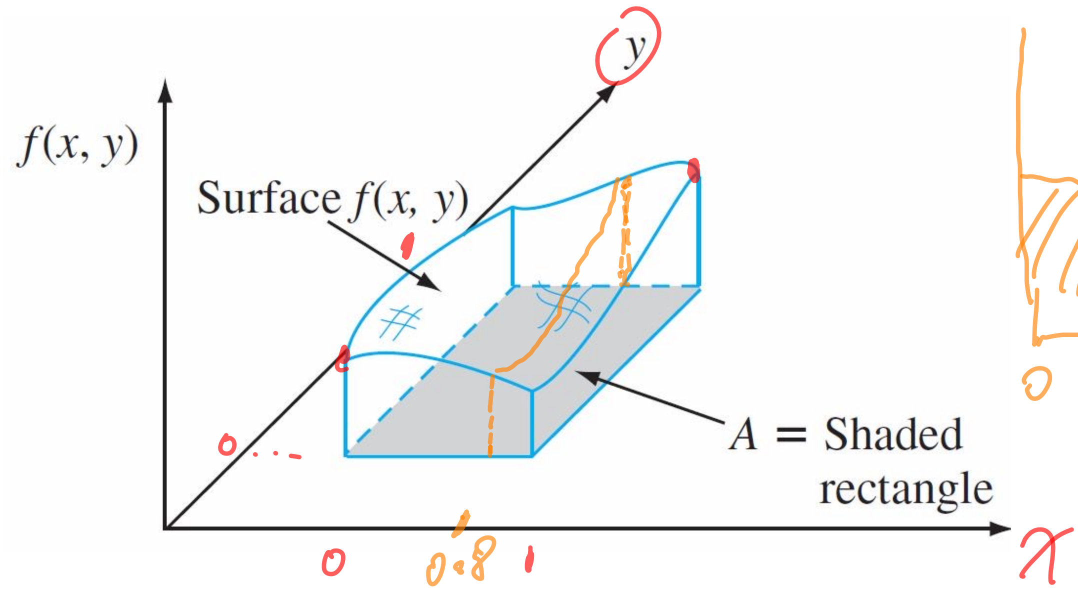
- In particular, if A is the two-dimensional rectangle $\{(x, y)\colon a \le x \le b,\ c \le y \le d\}$, then

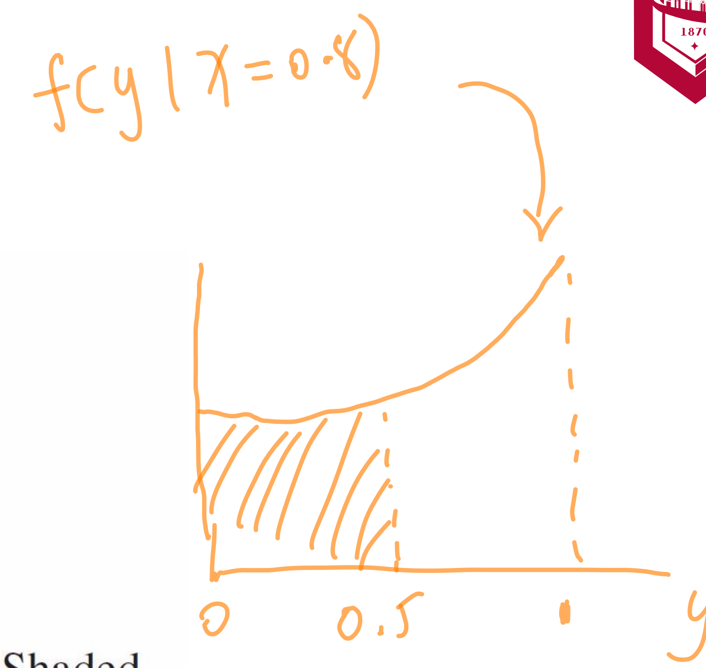
$$P[(X,Y) \in A] = P(a \le X \le b, c \le Y \le d) = \int_a^b \int_c^d f(x, y)\, dy\, dx$$

- We can think of $f(x, y)$ as specifying a surface at height $f(x, y)$ above the point (x, y) in a three-dimensional coordinate system.

  Then $P[(X, Y) \in A]$ is the volume underneath this surface and above the region A, analogous to the area under a curve in the case of a single rv.

Two Continuous Random Variables





$P[(X, Y) \in A]$ = volume under density surface above A

BIA 652, Multivariate Data Analytics

Two Continuous Random Variables

- **Definition**
- The **marginal probability density functions** of X and Y, denoted by $f_X(x)$ and $f_Y(y)$, respectively, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \qquad \text{for } -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx \qquad \text{for } -\infty < y < \infty$$

Independent Random Variables

- In many situations, information about the observed value of one of the two variables X and Y gives information about the value of the other variable.

  In Example 1, the marginal probability of X at $x = 250$
  was .5, as was the probability that $X = 100$. If, however, we are told that the selected individual had $Y = 0$, then $X = 100$ is four times as likely as $X = 250$.

  Thus there is a dependence between the two variables. Earlier, we pointed out that one way of defining independence of two events is via the condition
  $P(A \cap B) = P(A) \cdot P(B)$.

Independent Random Variables

Definition

Two random variables *X* and *Y* are said to be **independent** if for every pair of *x* and *y* values

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

If the above is not satisfied for all (*x*, *y*), then *X* and *Y* are said to be **dependent**

Events A and B are independent if

$$P(A \cap B) = P(A)P(B).$$

Random variables X and Y are independent if

$$F(x, y) = F_X(x)F_Y(y).$$

Discrete random variables X and Y are independent if

$$p(x_i, y_j) = p_X(x_i)p_Y(y_j).$$

Continuous random variables X and Y are independent if

$$f(x, y) = f_X(x)f_Y(y).$$

Independent Random Variables

- The definition says that two variables are independent if their joint pmf or pdf is the product of the two marginal pmf's or pdf's.

  Intuitively, independence says that knowing the value of one of the variables does not provide additional information about what the value of the other variable might be.

Conditional Distributions

Definition

The conditional probability density function of *Y* given that *X* = ***x*** is

$$f_{Y|X}(y\,|\,x) = \frac{f(x, y)}{f_X(x)} \qquad -\infty < y < \infty$$

$$f(x|y = f(x,y)/f(y)$$

If *X* and *Y* are discrete, replacing pdf's by pmf's in this definition gives the **conditional probability mass function of *Y* when *X*** = ***x*.**

Conditional Distributions

- Notice that the definition of $f_{Y\,|\,X}(y \mid x)$ parallels that of $P(B \mid A)$, the conditional probability that B will occur, given that A has occurred.

Once the conditional pdf or pmf has been determined, questions of the type posed at the outset of this subsection can be answered by integrating or summing over an appropriate set of Y values.

Covariance

$\Sigma x f(x)$

$cov(x,y) = cov(y,\cancel{x})$

- Random variables X,Y with joint discrete distribution p(X,Y) or continuous density f(x,y).
- Covariance = E({X – E[X]}{Y-E[Y]})

$$= \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x, y) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)f(x, y)\, dx\, dy \end{cases}$$

$\int x f(x)\, dx$

3D

= E[XY] – E[X] E[Y].

- Note, Covariance[X,X] = Var[X].

2D

$Var(x) \quad cov(x\,y) \quad cov(x\,z)$

$cov(xy) \quad Var(y) \quad cov(y,z)$

$cov(xz) \quad cov(y\,z) \quad Var(z)$

$cov(X,X) \quad cov(x,y)$

$cov(X,Y) \quad cov(y,y)$

$\rho_{x,z} = \frac{cov(x,z)}{\sqrt{Var(x)}\sqrt{Var(z)}}$

BIA 652, Multivariate Data Analytics

Correlation Coefficient

- **Definition**
- The **correlation coefficient** of *X* and *Y*, denoted by Corr(*X*, *Y*), $\rho_{X,Y}$, or just ρ, is defined by

[-1, 1]

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

$\sqrt{Var(X)}$ $\sqrt{Var(Y)}$

The Bivariate Normal Distribution

The Bivariate Normal Distribution

(x, y) distributed as bivariate normal if

(1) Infinite range $-\infty < x,y < +\infty$

(2) Normal marginal distributions:

$\mu_x \;\; \sigma_x$ $\qquad$ $\mu_y \;\; \sigma_y$

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\exp\left[\frac{-1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right], f(y) = \frac{1}{\sigma_y\sqrt{2\pi}}\exp\left[\frac{-1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]$$

(3) Joint Density

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left[\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right\}\right]$$

(6) New parameter ρ is the correlation between x and y.

$\mu_x, \mu_y, \underbrace{\sigma_x, \sigma_y, \rho}$ $\leftarrow \begin{pmatrix} var(X) & cov(X,Y) \\ cov(X,Y) & var(Y) \end{pmatrix}$

Independent Normals

(x, y) distributed as bivariate normal

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right\}\right]$$

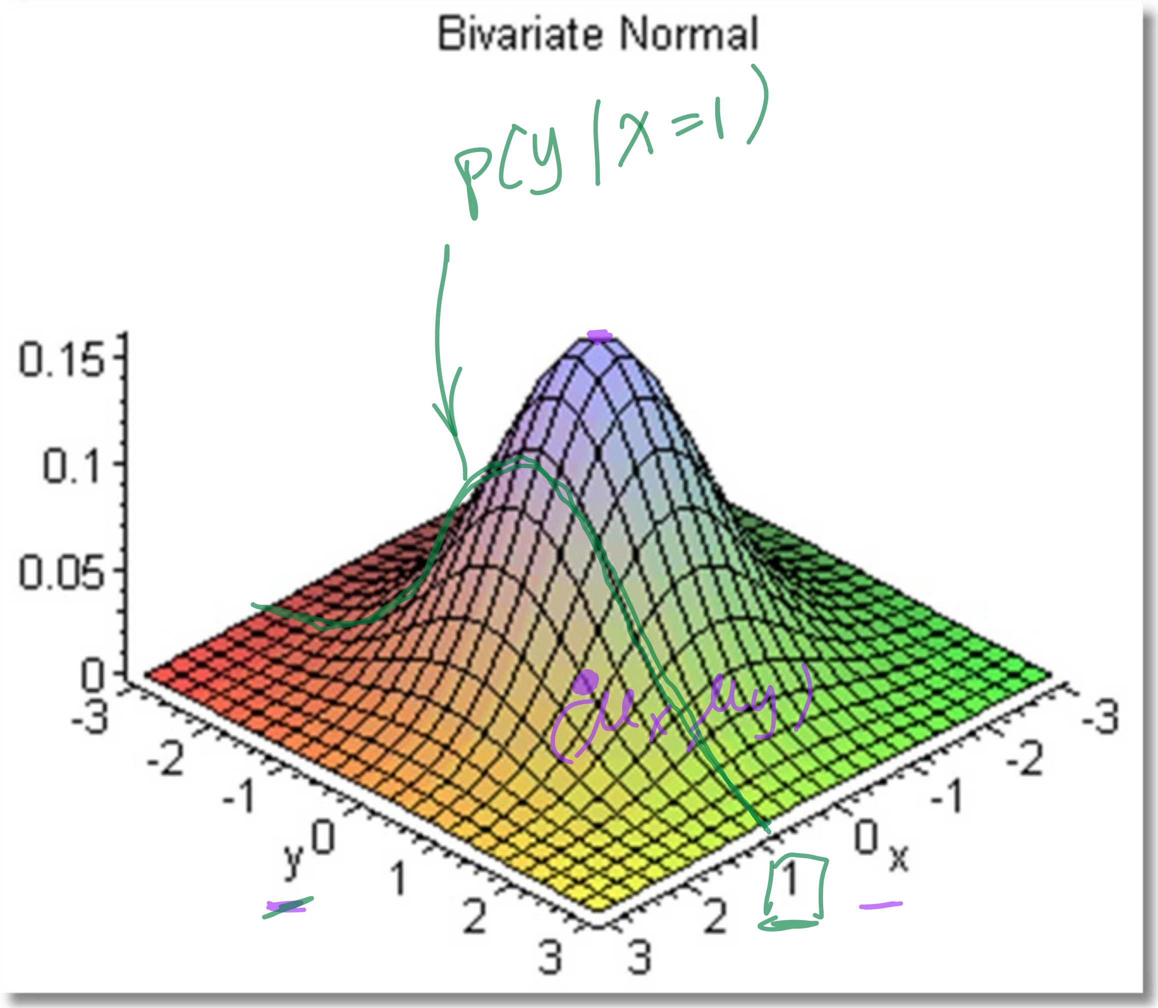
(2) Uncorrelated; $\rho = 0$

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[\frac{-1}{2}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\}\right] = f(x)\cdot f(y)$$

$$= \frac{1}{\sigma_x\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right] \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]$$

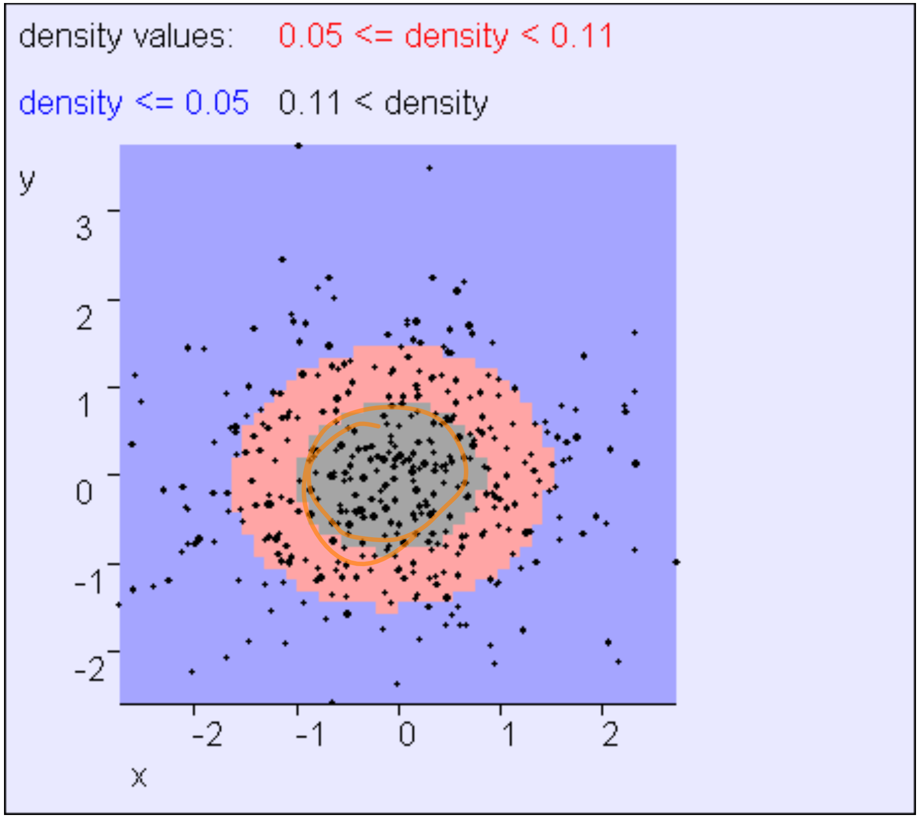
In the bivariate normal case (not generally), correlation = 0 implies indepencence

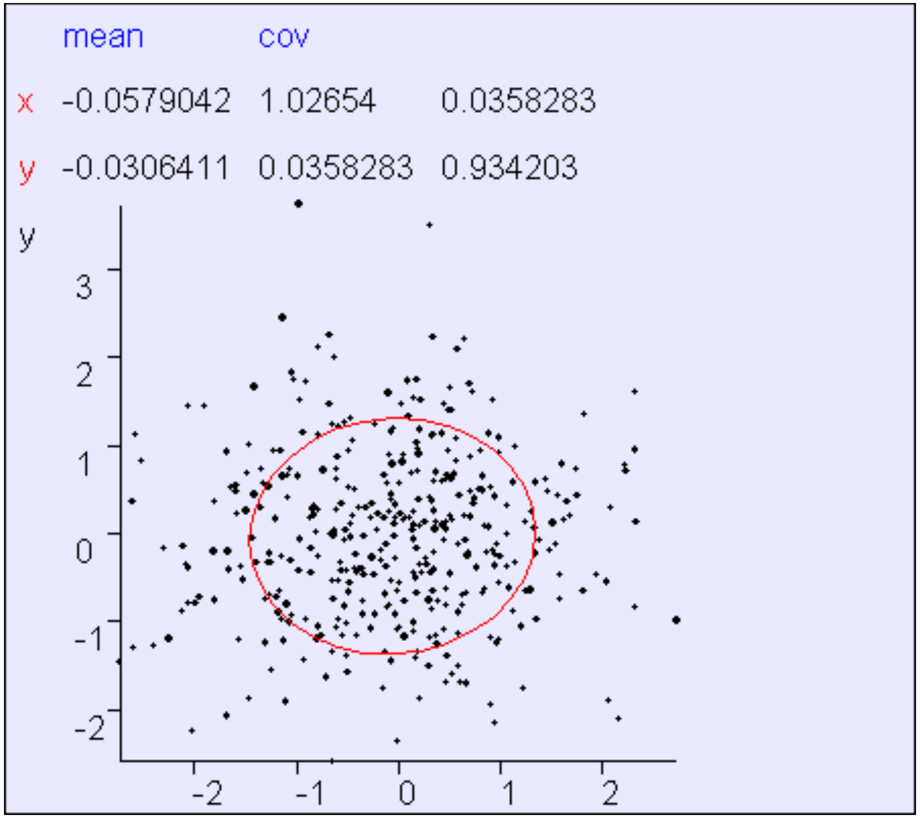
Independent Normals



$f(y) = \int f(x,y)\,dx$

Independent Normals



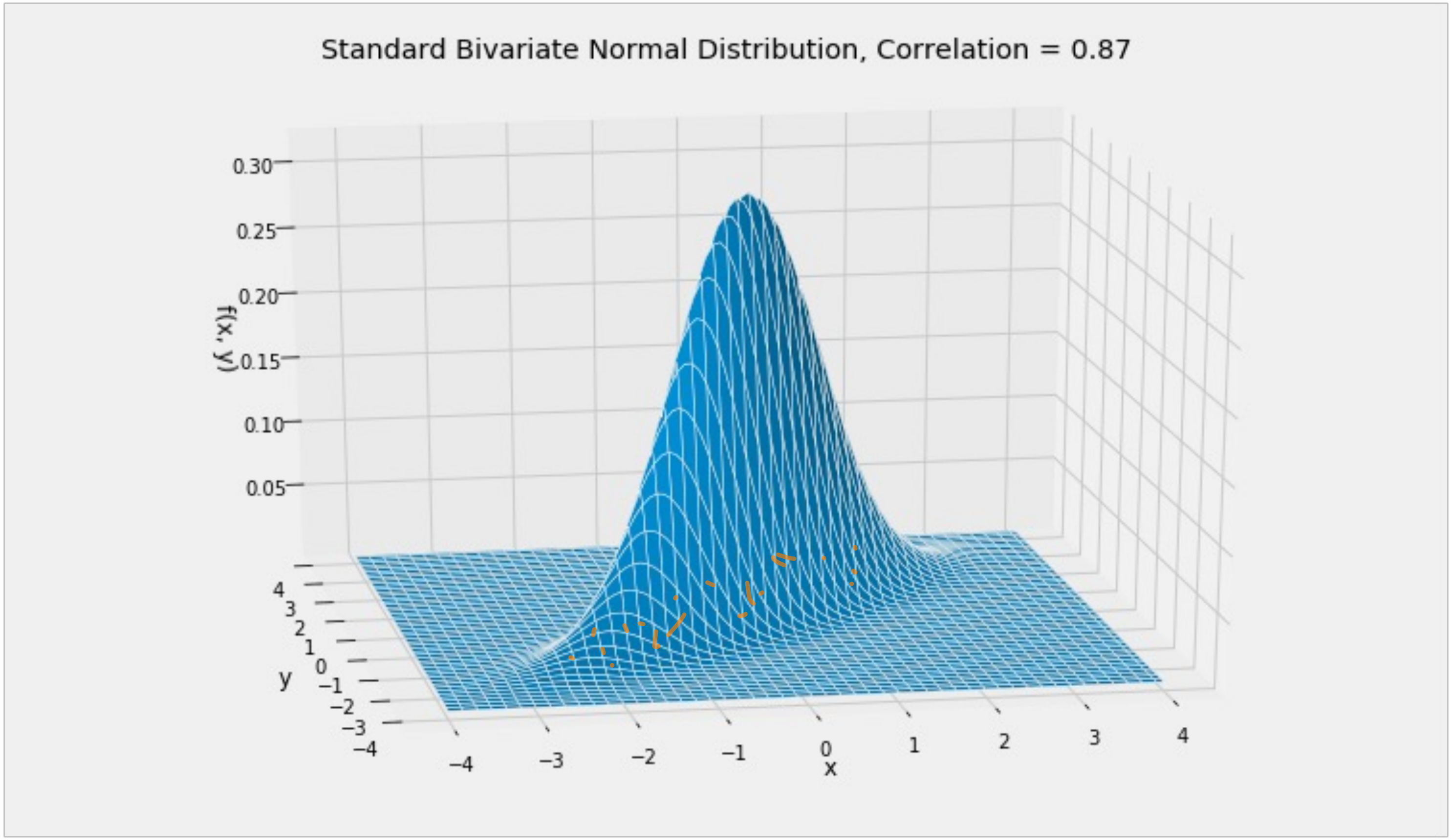


In this example, x and y are almost independent

BIA 652, Multivariate Data Analytics

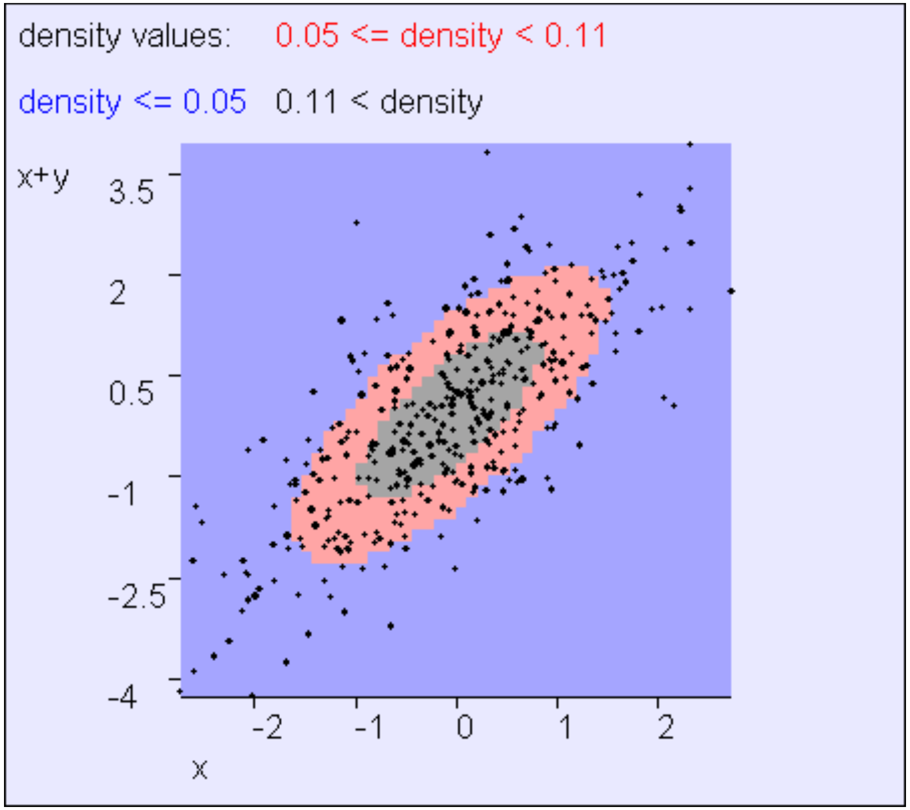
Correlated

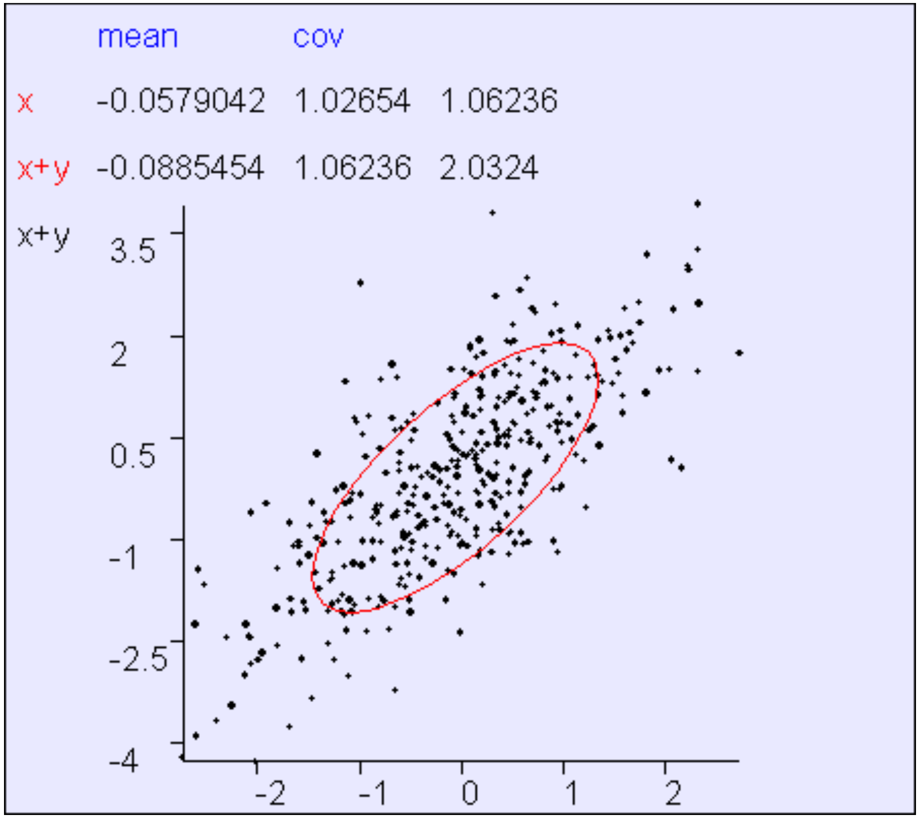
$\rho \neq 0$



BIA 652, Multivariate Data Analytics

Correlated





In this example, x and “x+y” are clearly not independent

Conditional Distribution for Bivariate Normal

$$f(y|x) = \underbrace{\frac{f(y,x)}{f(x)}}$$

(x, y) distributed as bivariate normal

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left[\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right\}\right] \leftarrow \text{joint}$$

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\exp\left[\frac{-1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right]$$

conditional

$\mu_{y|x}$

$$\underbrace{f(y\,|\,x)} = \frac{1}{\sigma_y\sqrt{2\pi}\sqrt{1-\rho^2}}\exp\left[\frac{-1}{2}\left\{\left(\frac{y-[\mu_y+\rho(\sigma_y/\sigma_x)(x-\mu_x)]}{\sigma_y\sqrt{1-\rho^2}}\right)^2\right\}\right] = \frac{1}{\underbrace{\sigma_{y|x}}\sqrt{2\pi}}\exp\left[\frac{-1}{2}\left\{\left(\frac{y-\mu_{y|x}}{\sigma_{y|x}}\right)^2\right\}\right]$$

$$= \underbrace{\sigma_y\sqrt{1-\rho^2}}$$

0.9

-0.9

BIA 652, Multivariate Data Analytics

Bivariate Normal

- Joint distribution is bivariate normal
- Marginal distributions are normal
- Conditional distributions are normal

Model Building

$y \quad x_1 \quad x_2 \quad x_3$

- Do we care more about joint f(x,y), conditional f(y|x), or marginal f(y), f(x)?

$P(y \mid x_1, x_2, x_3)$

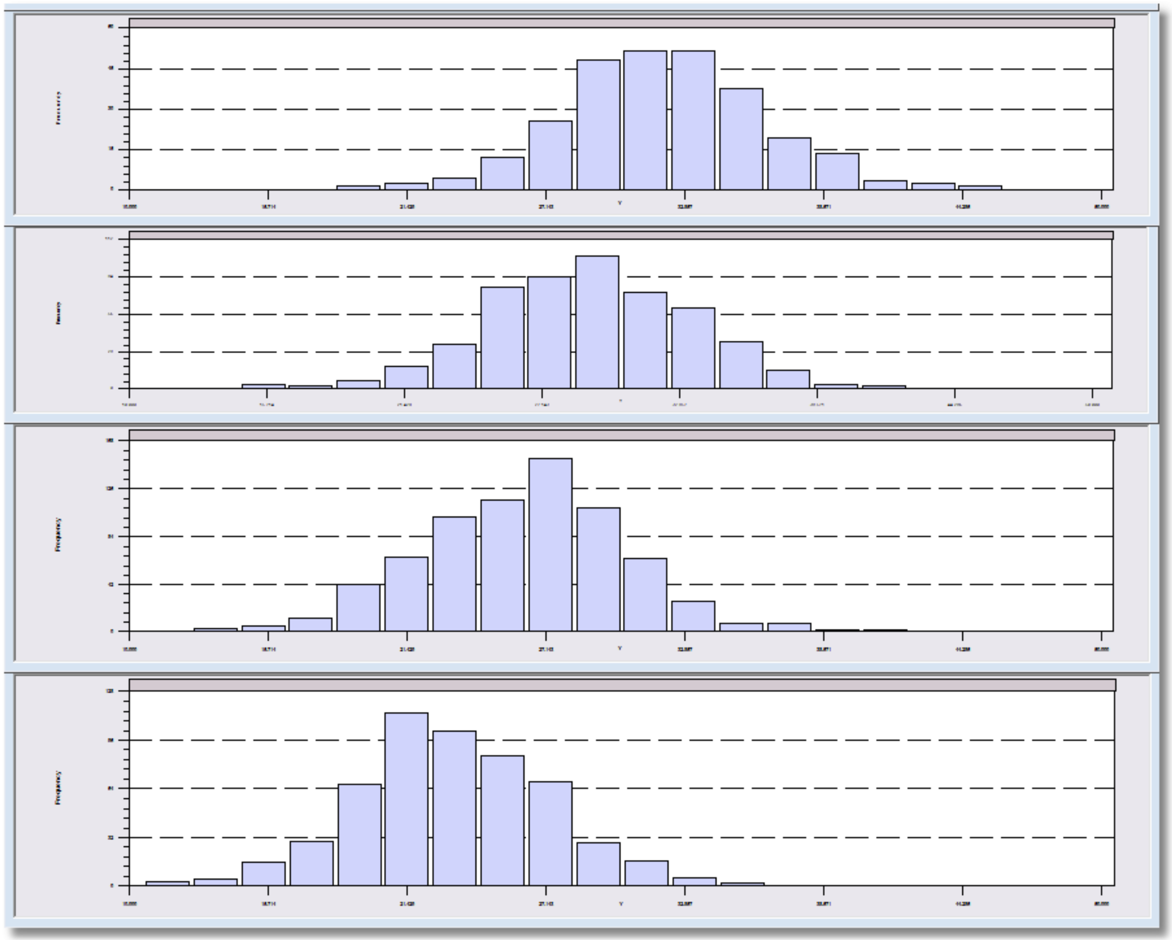
- x is generated by a separate process f(x)

$\mu_{y \mid x_1, x_2 \ldots}$

- Joint distribution is f(y,x)=f(y|x)f(x)

$\mathrm{Var}\ y \mid x_1, x_2 \ldots$

- Ex: demographic
  y = log(household income|family size)
  x = family size
  - $y|x \sim \text{Normal}(\mu_{y|x}, \sigma_{y|x})$
  - $x \sim \text{Poisson}(\lambda)$



$y|x \sim \text{Normal}[\,20 + 3x,\ 4^2\,],\ x = 1,2,3,4;$ Poisson

← 2

BIA 652, Multivariate Data Analytics

Thank you!

Prof. Feng Mai
School of Business