MATH 5010 –Foundations of Statistical Theory and Probability

❖ Joint and Conditional Probability

Instructor: He Wang
Department of Mathematics
Northeastern University

❖ Outline:

1. Expected Values (Review)
2. Joint and Marginal Distribution
3. Conditional Distributions and independence
4. Total probability and Bayes theorem
5. Conditional independence

❖ Expected Value (Review)

Expected value is a generalization of the concept "average".

The **expected value** or **mean** of a random variable X is defined to be

$$E[X] = \sum_{all\ k} k\, p_X(k)$$

Discrete random variable X

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)$$

Continuous random variable X

Example: Let X is the outcome of rolling a die.

k	$k = 0$	$k = 1$
$p_X(k)$	$1 - p$	p

$$E[X] = \sum_{all\ k} k\, p_X(k) = 0(1 - p) + 1(p) = p$$

The operational meaning is that $E[X]$ is the long-run average value of repeated measurements of the random variable X.

That is, suppose that we measure the random variable X in n independent trials, and record the results as $X_1, X_2, \ldots, X_n$. Then the long-run average value is

$$\overline{X_n} = \frac{1}{n}(X_1 + \cdots + X_n)$$

We will shortly see the Law of Large Numbers which implies that

$$\lim_{n \to \infty} \overline{X_n} = E(X)$$

Property: $E(aX + b) = aE(X) + b$

➢ Variance and Standard deviation

The **variance** of a random variable X is

$$Var(X) \coloneqq E\left[\left(X - E(X)\right)\right]^2$$

Variance is expected squared distance from the mean.

It measures the spread of the data.

The standard deviation is defined as the square root of the variance:

is $STD(X) \coloneqq \sqrt{Var(X)}$

Calculation formula: $\mathrm{Var}(X) = E(X^2) - \left(E(X)\right)^2$

Property: $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$

❖ Joint Distribution

For two random variables X, Y , the **joint CDF** is

$$P_{X,Y}(x, y) = F(x, y) = P(X \leq x, Y \leq y)$$

- When both variables are absolute ***continuous***, the corresponding joint pdf is

$$f_{X,Y}(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

- When both variables are ***discrete***, the joint **pdf** is the list of probabilities for all possible pairs of values:

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$

for all $x \in Range(X)$ and $y \in Range(Y)$

➢ Marginals

We can recover the individual pdf of X from the joint pdf of (X, Y) by summing over all values of Y. Similar for pdf of Y.

The **marginal pdfs** of random variables X and Y are defined by

$$p_X(x) = P(X = x) = \sum_{all\ y} p_{X,Y}(x, y) \quad and \quad p_Y(y) = P(Y = y) = \sum_{all\ x} p_{X,Y}(x, y)$$

The **marginal pdfs** of random variables X and Y are defined by

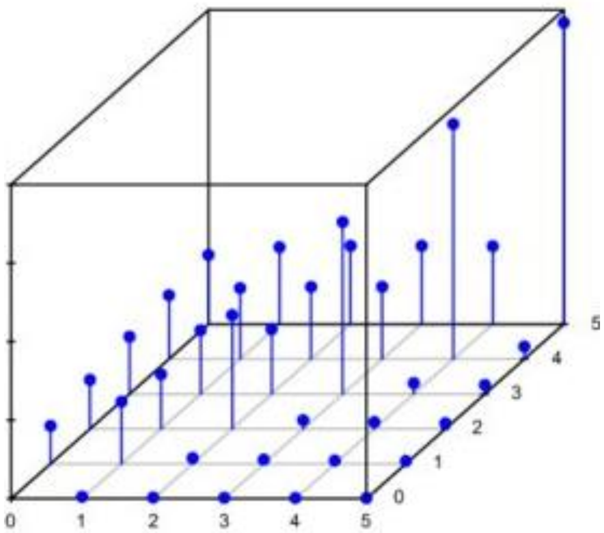
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx \quad and\ f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx$$

➢ Discrete Joint pdf

For two **discrete** random variables X and Y, the joint pmf $p_{X,Y}(x,y)$ satisfies

$$p_{X,Y}(x,y) \geq 0$$

$$\sum_{all\ x} \sum_{all\ y} p_{X,Y}(x,y) = 1$$



Example: Roll two dice, let X be **difference** of the two values, and let Y be the **maximum** of the two values.

The sample space S has 36 sample points given by

$$S = \left\{ \begin{array}{llllll} (1,1), & (1,2), & (1,3), & (1,4), & (1,5), & (1,6) \\ (2,1), & (2,2), & (2,3), & (2,4), & (2,5), & (2,6) \\ (3,1), & (3,2), & (3,3), & (3,4), & (3,5), & (3,6) \\ (4,1), & (4,2), & (4,3), & (4,4), & (4,5), & (4,6) \\ (5,1), & (5,2), & (5,3), & (5,4), & (5,5), & (5,6) \\ (6,1), & (6,2), & (6,3), & (6,4), & (6,5), & (6,6) \end{array} \right\}$$

The range of X is $Range(X) = \{0,1,2,3,4,5\}$ and the pdf of X is

$X = x$	0	1	2	3	4	5
$p_X(x)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

The range of Y is $Range(Y) = \{1,2,3,4,5,6\}$ and the pdf of Y is

$Y = y$	1	2	3	4	5	6
$p_Y(y)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

If $Ran(X)$ and $Ran(Y)$ are finite, then the joint pdf is conveniently presented as a table of values.

$p_Y(y)$

$Y \backslash X$	0	1	2	3	4	5
1	$1/36$	0	0	0	0	0
2	$1/36$	2/36	0	0	0	0
3	$1/36$	2/36	2/36	0	0	0
4	$1/36$	2/36	2/36	2/36	0	0
5	$1/36$	2/36	2/36	2/36	2/36	0
6	$1/36$	2/36	2/36	2/36	2/36	2/36

$p_X(x)$

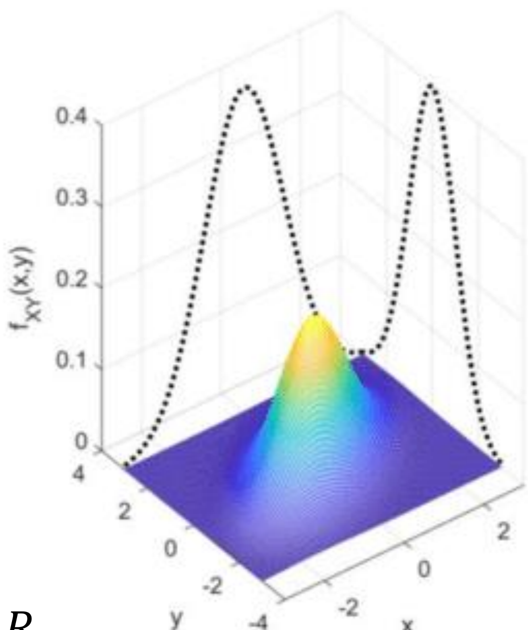
Example: Same marginals, different joint pmf

➢ Continuous Joint pdf

If X and Y are **continuous** random variables. the **joint pdf** $f_{X,Y}\ (x;\ y)$ of X and Y is a piecewise continuous multi-variable function satisfying

$$f_{X,Y}(x,y) \geq 0$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dxdy = 1$$



The **probability** for the pair $(X, Y\ )$ to be in some region R in the xy-plane is the double integral

$$P\big((X,Y) \in R\big) = \iint_R f_{X,Y}(x,y)\ dxdy$$

Example:

A joint pdf is defined by

$$f(x, y) \coloneqq 6xy^2 \text{ for } 0 < x < 1 \text{ and } 0 < y < 1.$$

1) Check that it is well-defined.
2) Calculate $P(X + Y \geq 1)$
3) Calculate Marginal pdf of X, $f_X(x)$.
4) Calculate probability $P\left(\frac{1}{2} < X < \frac{3}{4}\right)$

(1)

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_0^1\int_0^1 6xy^2\,dx\,dy = \int_0^1 3x^2y^2\Big|_0^1\,dy$$
$$= \int_0^1 3y^2\,dy = y^3\Big|_0^1 = 1.$$

(2)

$$P(X+Y\geq 1) = \int_A\int f(x,y)\,dx\,dy = \int_0^1\int_{1-y}^1 6xy^2\,dx\,dy = \frac{9}{10}.$$

(3)

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^1 6xy^2\,dy = 2xy^3\Big|_0^1 = 2x.$$

(4)

$$P\left(\frac{1}{2} < X < \frac{3}{4}\right) = \int_{\frac{1}{2}}^{\frac{3}{4}} 2x\,dx = \frac{5}{16}.$$

➢ **Multivariate normal distribution.**

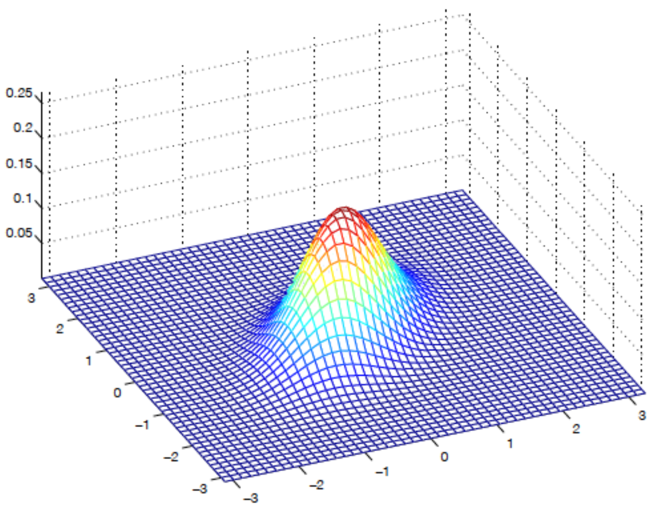
Vector random variable $\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix} \sim \text{Normal}(\vec{\mu}, \Sigma)$

Here $\vec{\mu} \in \mathbb{R}^d$ and Σ is an $d \times d$ symmetric, positive definite matrix.

- The **joint** probability density function (**pdf**) for $\vec{X}$ is

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}) \right)$$

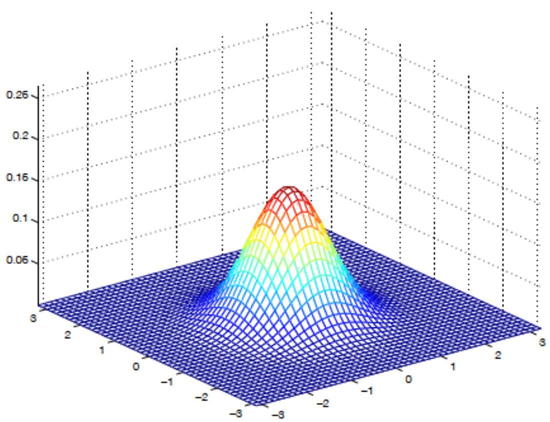
$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\vec{X}}(\vec{x}) \; d\vec{x} = 1$$



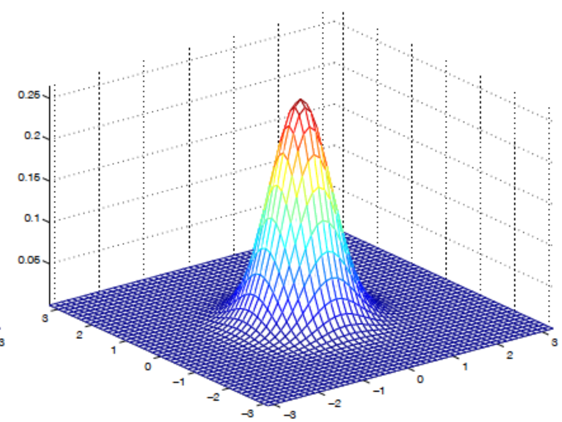
- The **mean vector** of $\vec{X}$ is $E(\vec{X}) = \vec{\mu}$

- The **(co)variance matrix** is Cov($\vec{X}$)$= \Sigma$

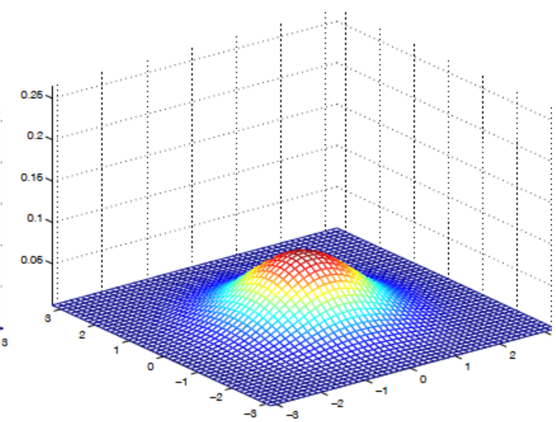
Standard normal

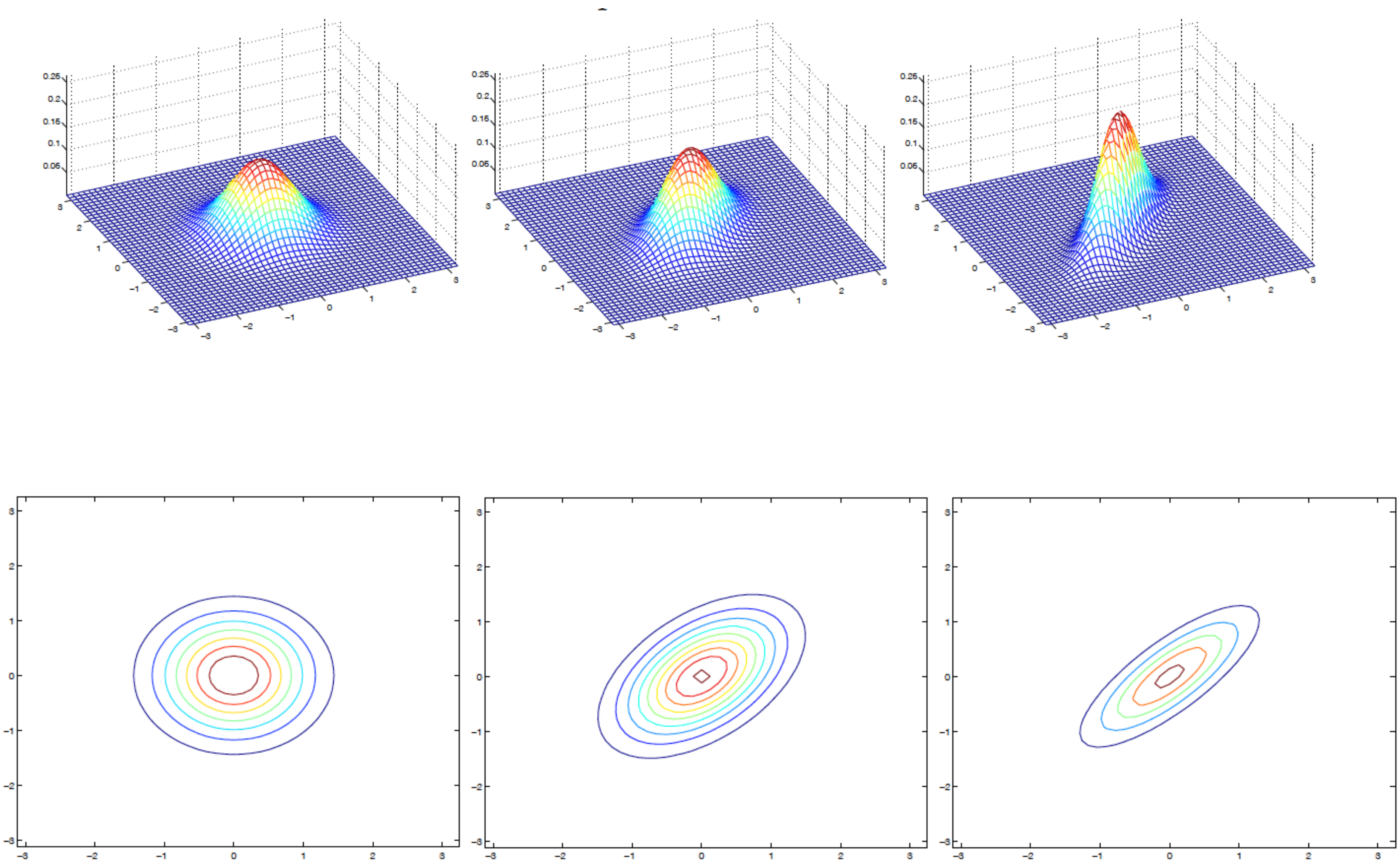


Compressed



Spread-out




0.25
0.2
0.15
0.1
0.05
−3 −2 −1 0 1 2 3

Second order statistics

In addition to the means $E[X]$ and $E[Y]$ and the variances, there is another statistic which measures the relation between X and Y : the **Covariance** is

$$COV[X,Y] = E\big[(X - E[X])\big(Y - E[Y]\big]\big) = E[XY] - E[X]E[Y]$$

The units depend on X and Y , so convenient to define the dimensionless (Pearson's) **correlation** coefficient:

$$CORR[X,Y] = \frac{COV[X,Y]}{STD(X)STD(Y)}$$

Correlation satisfies: $-1 \leq CORR[X,Y] \leq 1$.

❖ Conditional Probability

Definition. Probability that event A occurs **given** that event B already occurs, denoted by **P(A|B),** is a **conditional probability of A given B**, defined by

$$P(A|B) := \frac{P(A \cap B)}{P(B)} \qquad \text{if } P(B) \neq 0$$

When B is fixed, the function $P(\cdot\,|B) : S \to \mathbb{R}$ is another probability measure.

Conditional Distribution

For two random variables X, Y, the **conditional pdf/pmf of $\boldsymbol{Y}$ given $\boldsymbol{X = x}$** is

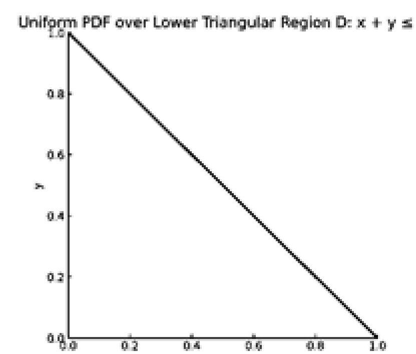
$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

where $p_{X,Y}(x,y)$ is the joint pdf and $p_X(x)$ is the marginal density function.

Example (Triangle Uniform).

Triangle Uniform PDF over the region $D = \{(x,y)\,|x \geq 0, y \geq 0, x + y \leq 1\}$.

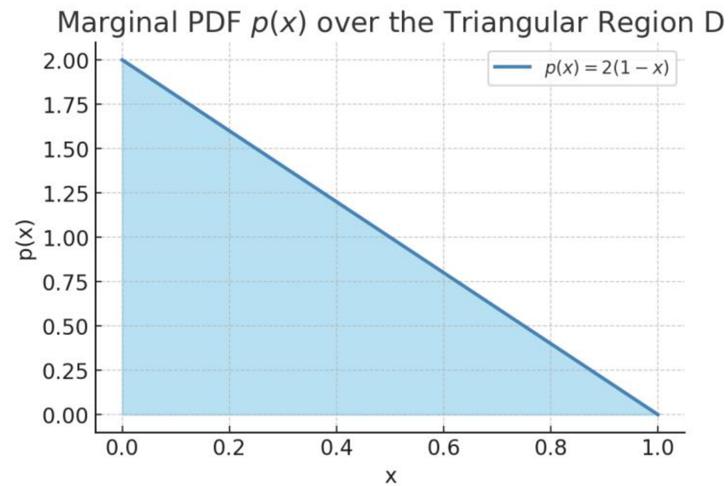
$$p(x,y) = \begin{cases} 2 & \text{if } (x,y) \in D \\ 0 & \text{otherwise} \end{cases}$$



The **marginal** pdf $p(x)$ is computed as:

$$p(x) = \int_{y=0}^{y=1-x} 2\,dy$$

$$p(x) = 2(1-x) \text{ for } 0 \leq x \leq 1$$



Since the joint PDF $p(x, y)$ is constant within the triangular region and zero elsewhere, the conditional PDF $p(y \mid x)$ will also be uniform for a fixed x, but restricted to the range of y that satisfies $x + y \leq 1$.

The **conditional pdf of $\boldsymbol{Y}$ given $\boldsymbol{X = x}$ is**

$$p(y|x) = \frac{1}{(1-x)} \text{ for } 0 \leq y \leq 1 - x.$$

Example: Beta-Bernoulli

The Beta-Bernoulli distribution is a common model in Bayesian statistics.

A Bernoulli random variable X takes values 0 or 1 with probability Y:

$$P(X = x \mid Y) = Y^x(1 - Y)^{1-x}$$

The Beta distribution Y has two parameters α and β:

$$p(y) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1 - y)^{\beta-1}$$

The joint Beta-Bernoulli distribution is

$$p(x, y) \ = \ p(x|y)p(y) \ = y^x(1 - y)^{1-x} \, \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1 - y)^{\beta-1}$$

$$= \frac{1}{B(\alpha, \beta)} y^{\alpha+x-1} \, (1 - y)^{\beta-x}$$

Independence

Definition: The **events** $A\ and\ B$ are called **independent** if

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

If $A\ and\ B$ are not empty set, A and B are **independent** if and only if

$P(A|B)\ =\ P(A)$ if and only if $P(B|A)\ =\ P(B)$

- **Independent Random Variables**

The **discrete** random variables X and Y are **independent** if the events $\{X = x_i\}$ and $\{Y = y_j\}$ are independent for every pair of possible values (x_i, y_j). Equivalently, the joint pdf is the product of the marginal pmf's:

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \text{ for all } x, y$$

The **continuous** random variables $X \; and \; Y$ are **independent** if the joint pdf is the product of the marginals, that is if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \text{ for all } x, y$$

Similarly, we can define the (mutually) independent r.v.'s $X_1, X_2, \ldots, X_n$.

Example: Coin tosses are *independent*. Suppose we toss a coin twice and

$$X = \begin{cases} 1 & \text{first toss is Heads} \\ 0 & \text{first toss is Tails} \end{cases}$$

$$Y = \begin{cases} 1 & second\ toss\ is\ heads \\ 0 & second\ toss\ is\ tails \end{cases}$$

Suppose the coin is biased and p is the probability of Heads, so $1 - p$ is the probability of Tails. Then for example

$$P(X = 1, Y = 0) = P(X = 1)P(Y = 0) = p(1 - p)$$

Example (Independence and information).

Suppose we observe $X_1, \dots, X_n \in \{0, 1\}$ that are **independent** and all from the same Bernoulli distribution with an unknown parameter

$$\theta_0 = P(X_i = 1)$$

Under a probabilistic model, any parameter θ would implies a joint pmf

$$p(x_1, \dots, x_n; \theta) = p(x_1; \theta)\, p(x_2; \theta) \cdots p(x_n; \theta)$$

The famous maximal likelihood estimator (MLE) finds an estimated value of θ by maximizing the log-likelihood value

$$\log p(x_1, \dots, x_n; \theta) = \log p(x_1; \theta) + \cdots + \log p(x_n; \theta)$$

$$l(\theta | x_1, \dots, x_n) = l(\theta | x_1) + \cdots + l(\theta | x_n)$$

the total information = sum of all individual information

Checking independence

Example: Roll two dice, let X be **difference** of the two values, and let Y be the **maximum** of the two values.

Checking independence

The verification that X and Y are independent by direct use of definition require the knowledge of $f_X(x)$ and $f_Y(y)$.

Theorem. Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$. Then X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(x)$ such that, for every x, y

$$f(x, y) = g(x)h(y)$$

Example: $$f(x, y) = \frac{1}{384} x^2 y^4 e^{-y-(x/2)},\ x > 0 \text{ and } y > 0$$

$$g(x) = x^2 e^{-x/2} \text{ for } x > 0$$

$$h(y) = \frac{1}{384} y^4 e^{-y} \text{ for } y > 0$$

Theorem: Let $X\ and\ Y$ be independent random variables.

- Then the expect value:

$$E\big(g(X)h(Y)\big) = E\big(g(X)\big)E\big(h(Y)\big)$$

- Then the moment generating function of the random variable $Z = X + Y$ is given by

$$M_Z(t) = M_X(t)M_Y(t)$$

Theorem:

Suppose $X \sim Normal(\mu_1, \sigma_1^2)$ and $Y \sim Normal(\mu_2, \sigma_2^2)$ are independent

Then $Z = X + Y \sim Normal(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Proof:

$$M_Z(t) = M_X(t)M_Y(t)$$

$$= \exp\left(\mu_1 t + \frac{\sigma_1^2 t^2}{2}\right)\exp\left(\mu_2 t + \frac{\sigma_2^2 t^2}{2}\right)$$

$$= \exp\left((\mu_1+\mu_2)t + \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}\right)$$

Covariance and independence

Suppose $X\ and\ Y$ are **any** random variables on the same sample space.

- $E(aX + bY) = aE(X) + bE(Y)$
- $Var(aX + bY) = a^2\, Var(X) + b^2\, Var(Y) + 2ab\, Cov(X,Y)$

$Cov(X,Y)$ is the **covariance** of X and Y defined as

$$Cov(X,Y) := E(XY) - E(X)E(Y)$$

If X and Y are **independent**, then

- $Var(aX + bY) = a^2\, Var(X) + b^2\, Var(Y)$
- $Cov(X,Y) = 0$
- $E(XY) = E(X)E(Y)$

The converse is not true.

Theorem. Law of Total Probability for *Random Variables*

$$p_Y(y) = \sum_{x'} p_{Y|X}(y|x')p_X(x')$$

Discrete X

$$p_Y(y) = \int p_{Y|X}(y|x')p_X(x')dx'$$

Continuous Discrete X

Example (Poisson-Binomial)

Let $$X \sim Poisson(\lambda)$$
$$Y\,|X = x \sim Binomial\,(x, p).$$

Find the marginal distribution $P(Y = y)$.

$$p_Y(y) = P(Y = y) = \sum_x P(Y = y, X = x) = \sum_x p_{Y|X}(y|x) p_X(x)$$

$$= \sum_{x \geq y} P(Y = y, X = x) P(X = x)$$

$$= \sum_{x \geq y} \binom{x}{y} p^y (1-p)^{\mathrm{x}-y} \frac{\lambda^x e^{-\lambda}}{x!}$$

$$= \sum_{x \geq y} \frac{x!}{(x-y)!\, y!} p^y (1-p)^{\mathrm{x}-y} \frac{\lambda^x e^{-\lambda}}{x!}$$

Set $k := x - y$

$$= \frac{p^y e^{-\lambda}}{y!} \sum_{k=0}^{\infty} \frac{1}{k!} (1-p)^{\mathrm{k}} \lambda^{y+k}$$

$$= \frac{(\lambda p)^y e^{-\lambda p}}{y!} \sum_{k=0}^{\infty} \frac{1}{k!} (1-p)^{\mathrm{k}} \lambda^k \, e^{-\lambda(1-p)}$$

So, $Y \sim Poisson(\lambda p)$

Theorem. Bayes' Theorem (for Random Variables)

For *random variables*, we also have the Bayes theorem:

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{\int p_{Y|X}(y|x')p_X(x')dx'}$$ Continuous X

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x'} p_{Y|X}(y|x')p_X(x')}$$ Discrete X

Theorem:

If random variables $X\ and\ Y$ are conditional independent given Z, then

(1). $p_{X,Y|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)$

(2). $p_{X|YZ}(x|y,z) = p_{X|Z}(x|z)$

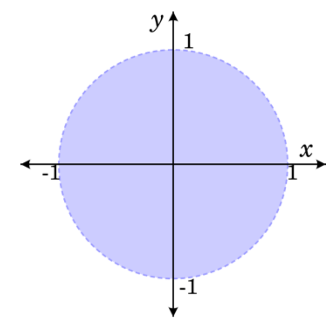
(3). $p_{X,Y,Z}(x,y,z) = \dfrac{p_{X,Z}(x,z)p_{Y,Z}(y,z)}{p_Z(z)}$

(4). $p_{X,Y,Z}(x,y,z) = g(x,z)h(y,z)$

Example:

Let (X, Y) be uniformly distributed over the **unit disk** $D = \{(x, y) \mid x^2 + y^2 \leq 1\}$

The joint PDF is $$f_{XY}(x,y) = \begin{cases} c & \text{if } (x,y) \in D \\ 0 & \text{otherwise} \end{cases}$$



a. Find the constant c

b. Find the marginal PDFs $f_X(x)$ and $f_Y(Y)$.

c. Find the conditional PDF $f_{X|Y}(x|y)$ for $-1 \leq y \leq 1$

d. Are X and Y independent?

a. Find the constant c

Since the joint PDF integrates to 1 over its support:

$$\iint_D f_{XY}(x,y)\,dx\,dy = 1$$

Because $f_{XY}(x,y) = c$ on D, and the area of the unit disk is π:

$$\iint_D c\,dx\,dy = c \cdot \text{Area}(D) = c \cdot \pi = 1 \quad \Rightarrow \quad c = \frac{1}{\pi}$$

b. Find the marginal PDFs $f_X(x)$ and $f_Y(Y)$.

To find the marginal PDF of X, integrate out y:

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} f_{XY}(x,y)\,dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}dy = \frac{2\sqrt{1-x^2}}{\pi}, \quad -1 \le x \le 1$$

By symmetry, the marginal PDF of Y is:

$$f_Y(y) = \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} f_{XY}(x,y)\,dx = \frac{2\sqrt{1-y^2}}{\pi}, \quad -1 \le y \le 1$$

c. Find the conditional PDF $f_{X|Y}(x|y)$ for $-1 \leq y \leq 1$

Using the definition of conditional PDF:

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)} = \begin{cases} \frac{\frac{1}{\pi}}{\frac{2\sqrt{1-y^2}}{\pi}} = \frac{1}{2\sqrt{1-y^2}} & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Support: $x \in [-\sqrt{1-y^2}, \sqrt{1-y^2}]$

d. Are X and Y independent?

No. If X and Y were independent, the joint PDF would be the product of the marginals:

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$$

But here:

- $f_{XY}(x, y) = \frac{1}{\pi}$ in the disk.
- $f_X(x) f_Y(y) = \frac{4\sqrt{(1-x^2)(1-y^2)}}{\pi^2}$, which is not constant over the disk.

Also, the conditional PDF $f_{X|Y}(x \mid y)$ depends on y, so X **and** Y **are not independent.**

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition (Section 4.1, 4.2)
- **Book 2. [W]: All of Statistics: Larry Wasserman**
- **Book 3. Introduction to Probability**. C.M. Grinstead and J.L. Snell. American Mathematical Society, 2012
- **Book 4. Introduction to Probability Models**, S. Ross, 12th edition (published by Academic Press).