Probabilistic Graphical Models in Bioinformatics

Tutorial2:Introduction to random variable & normal distribution

Fahimeh Palizban
Department of Bioinformatics, IBB, University of Tehran

13 Esfand ,1397
(4 March ,2019)

Contents

Contents

Random variables

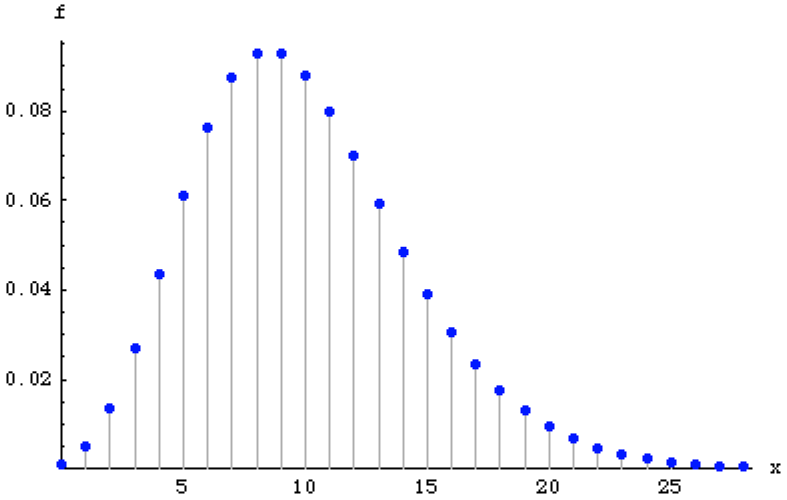
Definition. Given a random experiment with sample space *S*, a **random variable *X*** is a set function that assigns one and only one real number to each element *s* that belongs in the sample space *S*.
The set of all possible values of the random variable *X*, denoted *x*, is called the **support**, or **space**, of *X*.

Discrete random variable

Discrete RV

Definition. A random variable *X* is a **discrete random variable** if:

- there are a finite number of possible outcomes of *X*, or
- there are a countably infinite number of possible outcomes of *X*.



Probability mass function

The probability that a discrete random variable X takes on a particular value x, that is, $P(X = x)$, is frequently denoted $f(x)$.

Definition. The **probability mass function**, $P(X = x) = f(x)$, of a discrete random variable X is a function that satisfies the following properties:

(1) $P(X = x) = f(x) > 0$ if $x \in$ the support S

(2) $\sum_{x \in S} f(x) = 1$

(3) $P(X \in A) = \sum_{x \in A} f(x)$

Cumulative distribution function

The cumulative distribution function (CDF) of the random variable X has the following definition:

$$FX(t)=P(X\leq t)$$

The cdf of random variable X has the following properties:

1. $FX(t)$ is a nondecreasing function of t, for $-\infty<t<\infty$.
2. The cdf, $FX(t)$, ranges from 0 to 1. This makes sense since $FX(t)$ is a probability.
3. If X is a discrete random variable whose minimum value is a, then $FX(a)=P(X\leq a)=P(X=a)=fX(a)$. If c is less than a, then $FX(c)=0$.
4. If the maximum value of X is b, then $FX(b)=1$.
5. Also called the *distribution function*.
6. All probabilities concerning X can be stated in terms of F.

Mathematical expectations

Definition. If *f*(*x*) is the p.m.f. of the discrete random variable *X* with support S, and if the summation:

$$\sum_{x\in S} u(x)f(x)$$

exists (that is, it is less than ∞), then the resulting sum is called the **mathematical expectation**, or the **expected value** of the function *u*(*X*). The expectation is denoted *E*[*u*(*X*)]. That is:

$$E[u(X)]=\sum_{x\in S} u(x)f(x)$$

Example:

x	0	1	2	3
$f(x)$	0.2	0.1	0.4	0.3

Suppose the p.m.f. of the discrete random variable *X* is:

What is *E*(2)? What is *E*(*X*)? And, what is *E*(2*X*)?

Mean of DRV

Definition. When the function $u(X) = X$, the expectation of $u(X)$, when it exists:

$$E[u(X)]=E(X)=\sum_{x \in S} x f(x)$$

is called the **expected value of *X***, and is denoted $E(X)$. Or, it is called the **mean of *X***, and is denoted as μ . That is, $\mu = E(X)$. The expected value of X can also be called the **first moment about the origin**.

Variance of DRV

Definition. When $u(X) = (X - \mu)^2$, the expectation of $u(X)$:

$$E[u(X)]=E[(X-\mu)2]=\sum x\in S(x-\mu)2f(x)$$

is called the **variance of *X***, and is denoted as ***Var*(*X*)** or ***σ*²** ("sigma-squared"). The variance of *X* can also be called the **second moment of *X* about the mean *μ***.
The positive square root of the variance is called the **standard deviation of *X***, and is denoted ***σ*** ("sigma"). That is:

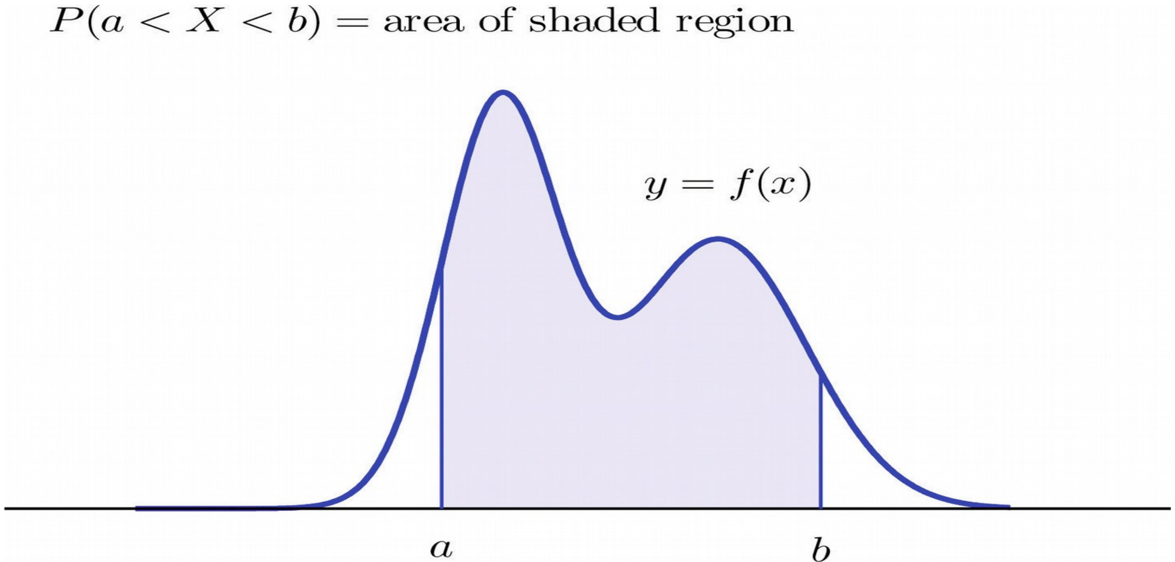
$$\sigma=\sqrt{}Var(X)=\sqrt{}\sigma 2$$

Continuous random variable

Continuous random variable

It takes on an uncountably infinite number of possible outcomes

Finding the probability that X falls in some interval, that is finding $P(a < X < b)$, where a and b are some constants. We'll do this by using $f(x)$, the probability density function ("p.d.f.") of X, and $F(x)$, the cumulative distribution function ("c.d.f.") of X.



Probability density function

Definition. The **probability density function** ("**p.d.f.**") of a continuous random variable X with support S is an integrable function $f(x)$ satisfying the following:

(1) $f(x)$ is positive everywhere in the support S, that is, $f(x) > 0$, for all x in S

(2) The area under the curve $f(x)$ in the support S is 1, that is:

$$\int_S f(x)dx=1$$

(3) If $f(x)$ is the p.d.f. of x, then the probability that x belongs to A, where A is some interval, is given by the integral of $f(x)$ over that interval, that is:

$$P(X \in A)=\int_A f(x)dx$$

Example

Let X be a continuous random variable whose probability density function is:

$$f(x) = 3x^2$$

Cumulative distribution function

Definition. The **cumulative distribution function** ("**c.d.f.**") of a continuous random variable X is defined as:

$$F(x)=\int x f(t)dt$$

for $-\infty < x < \infty$.

For continuous random variables, $F(x)$ is a non-decreasing *continuous* function.

Expected value of a continuous random variable

Definition:Let X be a continuous random variable with range [a,b] and probability density function f(x). The expected value of X is defined by:

$$E(X) = \int x f(x) dx$$

-Example 1.Let X∼uniform(0,1). FindE(X).

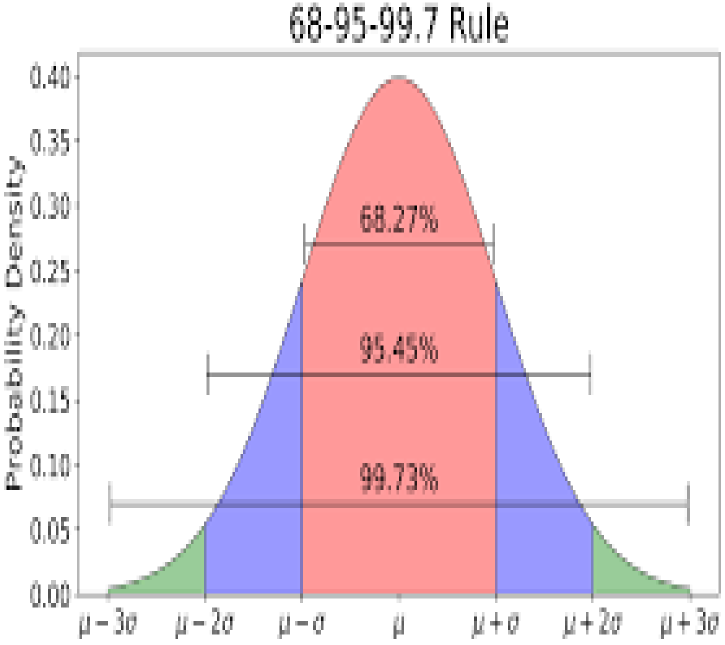
Definition:Let X be a continuous random variable with meanμ. The variance of X is:

$$Var(X) = E((X-\mu)^2) \qquad var(X)=\sigma_X^2=E[(X-\mu_X)^2]=\int (x-\mu_X)^2 f_X(x)dx.$$

-Example 2.Let X∼uniform(0,1). Find Var(X) and σ_X.

Normal distribution

Normal distribution (univariate)



Definition. The continuous random variable *X* follows a **normal distribution** if its probability density function is defined as:

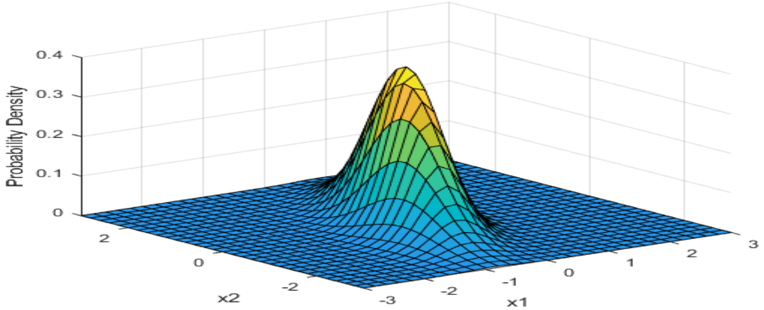
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, and $0 < \sigma < \infty$. The **mean** of *X* is $\boldsymbol{\mu}$ and the **variance** of *X is* $\boldsymbol{\sigma^2}$. We say $\boldsymbol{X \sim N(\mu, \sigma^2)}$.

Normal distribution (multivariate)

A vector-valued random variable $X = [X_1 \cdots X_n]^T$ is said to have a **multivariate normal (or Gaussian) distribution** with mean $\mu \in \mathbf{R}^n$ and covariance matrix $\Sigma \in \mathbf{S}^n_{++}$ [1] if its probability density function[2] is given by

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$



MND vs UND

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = 1.$$

$$\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)dx_1dx_2\cdots dx_n = 1.$$

Multivariate Conditional Distributions

Given variables $\mathbf{x} = (x_1, \ldots, x_p)'$ and $\mathbf{y} = (y_1, \ldots, y_q)'$, we have

$$f_{Y|X}(\mathbf{y}|X = \mathbf{x}) = \frac{f_{XY}(\mathbf{x}, \mathbf{y})}{f_X(\mathbf{x})}$$

where

- $f_{Y|X}(\mathbf{y}|X = \mathbf{x})$ is the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$
- $f_{XY}(\mathbf{x}, \mathbf{y})$ is the joint pdf of $\mathbf{x}$ and $\mathbf{y}$
- $f_X(\mathbf{x})$ is the marginal pdf of $\mathbf{x}$

Suppose that $\mathbf{z} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

- $\mathbf{z} = (\mathbf{x}', \mathbf{y}')' = (x_1, \ldots, x_p, y_1, \ldots, y_q)'$
- $\boldsymbol{\mu} = (\boldsymbol{\mu}_x', \boldsymbol{\mu}_y')' = (\mu_{1x}, \ldots, \mu_{px}, \mu_{1y}, \ldots, \mu_{qy})'$
  Note: $\boldsymbol{\mu}_x$ is mean vector of $\mathbf{x}$, and $\boldsymbol{\mu}_y$ is mean vector of $\mathbf{y}$
- $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy}' & \boldsymbol{\Sigma}_{yy} \end{pmatrix}$ where $(\boldsymbol{\Sigma}_{xx})_{p\times p}$, $(\boldsymbol{\Sigma}_{yy})_{q\times q}$, and $(\boldsymbol{\Sigma}_{xy})_{p\times q}$,
  Note: $\boldsymbol{\Sigma}_{xx}$ is covariance matrix of $\mathbf{x}$, $\boldsymbol{\Sigma}_{yy}$ is covariance matrix
  and $\boldsymbol{\Sigma}_{xy}$ is covariance matrix of $\mathbf{x}$ and $\mathbf{y}$

In the multivariate normal case, we have that

$$\mathbf{y}|\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$$

where $\boldsymbol{\mu}_* = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{xy}'\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x)$ and $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{xy}'\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$

$$\mathbf{y}|\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \equiv \mathrm{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{yy})$$

if and only if $\boldsymbol{\Sigma}_{xy} = \mathbf{0}_{p\times q}$ (a matrix of zeros).

Note that $\boldsymbol{\Sigma}_{xy} = \mathbf{0}_{p\times q}$ implies that the p elements of $\mathbf{x}$ are uncorrelated with the q elements of $\mathbf{y}$.

- For multivariate normal variables: uncorrelated $\rightarrow$ independent
- For non-normal variables: uncorrelated $\not\rightarrow$ independent

Example

Each Delicious Candy Company store makes 3 size candy bars: regular (X_1), fun size (X_2), and big size (X_3).

Assume the weight (in ounces) of the candy bars (X_1, X_2, X_3) follow a multivariate normal distribution with parameters:

- $\mu = \begin{pmatrix} 5 \\ 3 \\ 7 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & 2 \\ 0 & 2 & 9 \end{pmatrix}$

Suppose we select a store at random. What is the probability that...

(a) the weight of a regular candy bar is greater than 8 oz?

(b) the weight of a regular candy bar is greater than 8 oz, given that the fun size bar weighs 1 oz and the big size bar weighs 10 oz?

(c) $P(4X_1 - 3X_2 + 5X_3 < 63)$?

References

https://newonlinecourses.science.psu.edu/stat414/node/57/

http://users.stat.umn.edu/~helwig/notes/norm-Notes.pdf

http://bateni.persiangig.com/.JZM4xf8cC5/document/Koller,Friedman-Probabilistic%20Graphical%20Models_%20Principles%20and%20Techniques-The%20MIT%20Press%20(2009).pdf

Thanks.