Prelim. vocabulary for 'Statistical methods of HEP analysis'

1월 20일 저녁 강의 'Statistical methods of HEP analysis'에 꼭 필요한 기초 내용입니다. 강의 전까지 숙지하고 오시면 좋겠습니다. – 권영준

1 Random variables, PDF, CDF

- A **random variable** is a numerical characteristic assigned to an element of the sample space. It can be discrete or continuous.

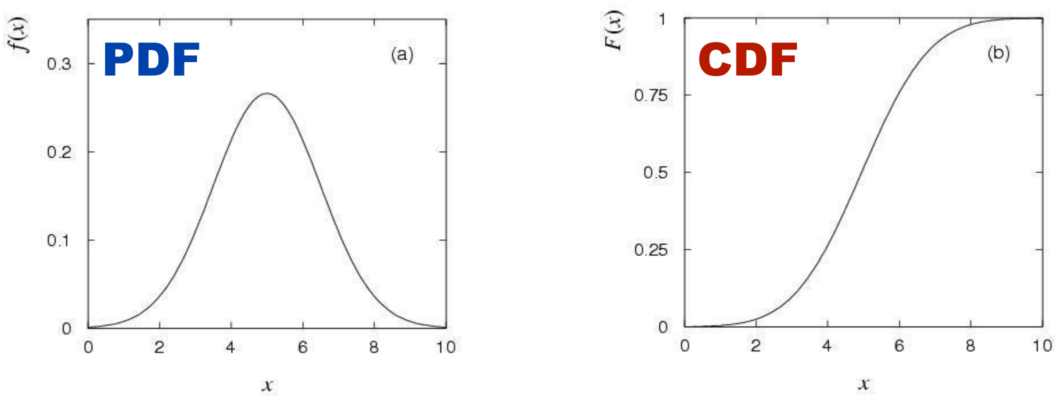
- Suppose the probability $P(x \in [x, x+dx])$ of a random variable x to be found within the region $[x, x+dx]$ is $f(x)dx$. Then we call $f(x)$ the **probability density function** (PDF). The PDF must be properly normalized:

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \; .$$

  (Q) How will it appear if x is a discrete random variable?

- The probability $F(x)$ to have an outcome less than or equal to x is called the **cumulative distribution function** (CDF).

$$\int_{-\infty}^{x} f(x')dx' \equiv F(x) \; .$$



2 Expectation value, mean, variance, covariance

- **Expectation value** of a function $g(x)$

$$E[g] \equiv \int_{\Omega} dx f(x)g(x) \; ,$$

  where Ω is the random variable space and $x \in \Omega$.

  For discrete random variable x,

$$E[g] \equiv \sum_{\Omega} P(x)g(x) \; .$$

- Expectation value is a linear operation:

$$E[\alpha g(x) + \beta h(x)] = \alpha E[g(x)] + \beta E[h(x)]$$

- **mean** = expectation value for the random variable x

$$\mu = \overline{x} = \langle x \rangle = \int_\Omega dx\ f(x)x = E[x]$$

- **variance** $V(x) = \sigma^2$

  The square root of the variance is often called the standard deviation, σ.

$$\begin{aligned} V(x) = \sigma^2 &= E[(x-\mu)^2] \\ &= E[x^2] - (E[x])^2 \\ &= \int_\Omega dx f(x)\,(x-\mu)^2 \end{aligned}$$

- sample mean & sample variance

  Since we don't a priori know the true mean and the true variance[1], we often use the measured sample to estimate the mean and variance. Suppose we have n measurements $\{x_i\}$ where x_i follows $N(\mu, \sigma)$ which is a normal ("Gaussian") distribution with mean μ and variance σ^2.

  - sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\ \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

    With more measurements, the estimation of the mean will become more accurate.

  - sample variance

$$V(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}\left(\overline{x^2} - \overline{x}^2\right)$$

    Sample variance approaches σ^2 for large n.
    (Q) Why is the denominator $n-1$ rather than n?

- For a multiple-dimensional random variable space,

$$E[g(x,y)] = \int\int_\Omega dx\ dy f(x,y) g(x,y)\ ,$$

  where $f(x,y)$ is the 2-dimensional PDF.

  We also have, for the mean and variance,

$$\mu_x = E[x] = \int\int_\Omega dx\ dy f(x,y)x$$

$$\sigma_x^2 = E[(x-\mu_x)^2] = \int\int_\Omega dx\ dy f(x,y)\,(x-\mu_x)^2$$

[1]In most problem, these are the variables we want to find out.

- **covariance**, $V_{x,y}$

$$
\begin{aligned}
V_{x,y} &\equiv E[(x-\mu_x)(y-\mu_y)] \\
&= E[xy] - E[x]\ E[y]
\end{aligned}
$$

Often, the **correlation coeffiient** is used to show the correlation between two random variables (here, x and y):

$$\rho(x,y) \equiv \frac{V_{x,y}}{\sigma_x\ \sigma_y}$$

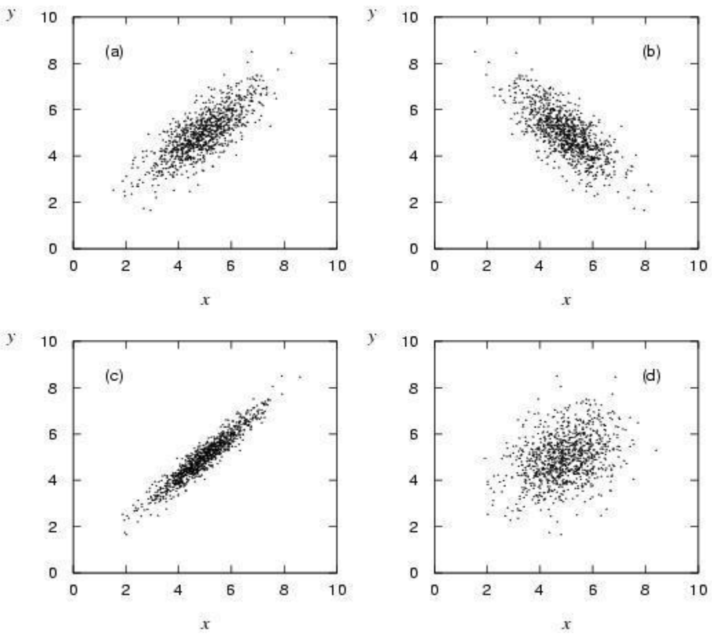
(HW) Show the following:

- * $-1 \le \rho(x,y) \le +1$
- * For *independent* variables x and y, $\rho(x,y) = 0$.
- * But the reverse is not true.
  For example, consider $y = x^2$ for $-1 \le x \le +1$.

Some examples of 2D correlated distributions:

$\rho = 0.75$ $\rho = -0.75$

$\rho = 0.95$ $\rho = 0.25$



3 Error propagation

Suppose we have a known function $f(x,y)$ having 2D random variables x and y as its arguments. Assume that we have the 2D covariance matrix for (x,y). Then the error (uncertainty) in $f(x,y)$ is obtained by:

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) \begin{pmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{pmatrix} \begin{pmatrix} \partial f/\partial x \\ \partial f/\partial y \end{pmatrix}$$

(Q) What happens if x and y are independent?

4 Some common PDF's

Table 35.1. Some common probability density functions, with corresponding characteristic functions and means and variances. In the Table, $\Gamma(k)$ is the gamma function, equal to $(k-1)!$ when k is an integer; ${}_1F_1$ is the confluent hypergeometric function of the 1st kind [11].

Distribution	Probability density function f (variable; parameters)	Characteristic function $\phi(u)$	Mean	Variance σ^2
Uniform	$f(x;a,b) = \begin{cases} 1/(b-a) & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$	$\dfrac{e^{ibu} - e^{iau}}{(b-a)iu}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Binomial	$f(r;N,p) = \dfrac{N!}{r!(N-r)!}\, p^r q^{N-r}$ $r = 0,1,2,\ldots,N\ ;\quad 0 \le p \le 1\ ;\quad q = 1-p$	$(q + pe^{iu})^N$	Np	Npq
Poisson	$f(n;\nu) = \dfrac{\nu^n e^{-\nu}}{n!}\ ;\quad n = 0,1,2,\ldots\ ;\quad \nu > 0$	$\exp[\nu(e^{iu}-1)]$	ν	ν
Normal (Gaussian)	$f(x;\mu,\sigma^2) = \dfrac{1}{\sigma\sqrt{2\pi}}\ \exp(-(x-\mu)^2/2\sigma^2)$ $-\infty < x < \infty\ ;\quad -\infty < \mu < \infty\ ;\quad \sigma > 0$	$\exp(i\mu u - \frac{1}{2}\sigma^2 u^2)$	μ	σ^2
Multivariate Gaussian	$f(\boldsymbol{x};\boldsymbol{\mu},V) = \dfrac{1}{(2\pi)^{n/2}\sqrt{\lvert V\rvert}}$ $\times \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$ $-\infty < x_j < \infty;\quad -\infty < \mu_j < \infty;\quad \lvert V\rvert > 0$	$\exp\left[i\boldsymbol{\mu}\cdot\boldsymbol{u} - \frac{1}{2}\boldsymbol{u}^T V \boldsymbol{u}\right]$	$\boldsymbol{\mu}$	V_{jk}
χ^2	$f(z;n) = \dfrac{z^{n/2-1}e^{-z/2}}{2^{n/2}\Gamma(n/2)}\ ;\quad z \ge 0$	$(1-2iu)^{-n/2}$	n	$2n$
Student's t	$f(t;n) = \dfrac{1}{\sqrt{n\pi}}\ \dfrac{\Gamma[(n+1)/2]}{\Gamma(n/2)}\left(1 + \dfrac{t^2}{n}\right)^{-(n+1)/2}$ $-\infty < t < \infty\ ;\quad n$ not required to be integer	—	0 for $n > 1$	$n/(n-2)$ for $n > 2$
Gamma	$f(x;\lambda,k) = \dfrac{x^{k-1}\lambda^k e^{-\lambda x}}{\Gamma(k)}\ ;\quad 0 \le x < \infty\ ;$ k not required to be integer	$(1 - iu/\lambda)^{-k}$	k/λ	k/λ^2
Beta	$f(x;\alpha,\beta) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ $0 \le x \le 1$	${}_1F_1(\alpha;\alpha+\beta;iu)$	$\dfrac{\alpha}{\alpha+\beta}$	$\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$