

## Statistical Background

In following notes, concepts regarding random variables, distribution functions, expectations and variances are discussed. This note concludes with a short discussion about the normal distribution.

### Random Variables, pdf and cdf

A variable is called a *random* (or stochastic) if the possible values of the variable have different probabilities. Therefore a random variable always has a probability density function (pdf) and a probability distribution. The relationship between distributions and probabilities can be defined as

$$\text{Prob}(a \leq x \leq b) = \int_a^b f(x) \partial x,$$

where function  $f(x)$  is the probability density function. In words, integration of function  $f(x)$  over the interval  $a$  and  $b$  indicates probability of occurrence of  $x$ . A cumulative probability distribution function (cdf) for continuous  $x$  is defined as

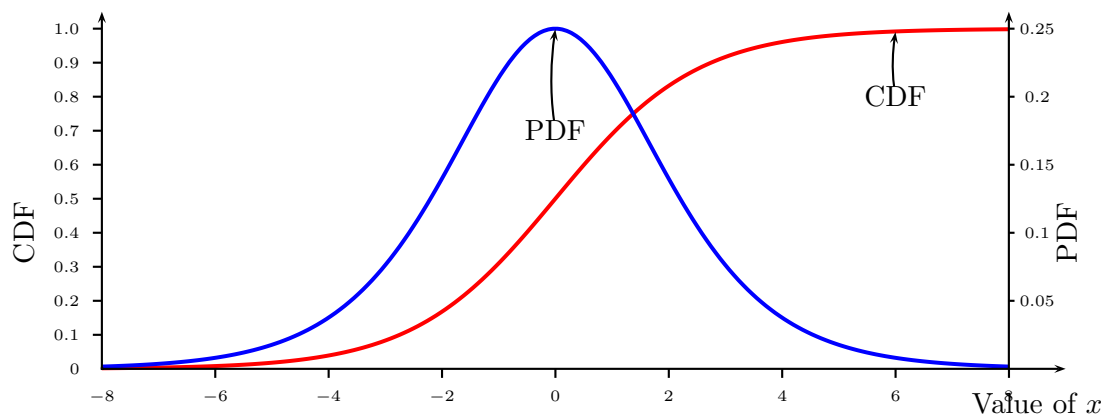
$$F(a) = \int_{-\infty}^a f(x) \partial x,$$

and for discrete  $x$  as

$$F(a) = \sum_{x \leq a} \text{Prob}(x).$$

There are three important properties of cumulative probability distribution functions. These are

1.  $0 \leq F(x) \leq 1$ . That is, cdf is bounded between 0 and 1.
2. If  $a < b$  then  $F(a) \leq F(b)$ . That is, cdf is non-decreasing function.
3.  $F(-\infty) = 0$  and  $F(\infty) = 1$ . This property is re-stating first one.



An Example of Cumulative and Probability Distribution Function

**Expectation**

The mean or expected value of a continuous variable as

$$\mathcal{E}(x) = \int_{-\infty}^{\infty} xf(x)\partial x,$$

or for a discrete random variable is defined as

$$\mathcal{E}(x) = \sum_{i=1}^n x_i \text{Prob}_i.$$

**Example:** Suppose interest is about price increases faced by individuals in the industrialized countries. Following table contains data collected from *Economist* webpage<sup>1</sup> as of January 5, 2002 on consumer price indices for 15 countries. Estimates for number of consumers in each country was obtained from the United Nations webpage.

Country	CPI (%)	Population (in millions)	Percent of total
Australia	2.5	18.91	0.0235
Austria	2.0	8.18	0.0102
Belgium	2.2	10.15	0.0126
Britain	0.9	58.74	0.0730
Canada	0.7	30.49	0.0379
Denmark	1.9	5.31	0.0066
France	1.2	58.89	0.0732
Germany	1.7	82.18	0.1022
Italy	2.4	57.34	0.0713
Japan	-1.0	126.51	0.1573
Netherlands	4.2	15.85	0.0197
Spain	2.7	39.42	0.0490
Sweden	2.7	8.86	0.0110
Switzerland	0.3	7.14	0.0089
United States	1.9	276.22	0.3435
Simple Average	1.75	804.19	
Weighted Average	1.39		

Note that the simple average is unweighted by the population size while weighted average is based on size of population in each country. Explain different  $f(x)$  used in computing these two expected or mean value calculations and which one is more reflective on an average price increases faced by consumers? Why?

<sup>1</sup><http://www.economist.com>

**Example:** Above example relates to continuous variable, the same approach also applies to ordered or interval scaled variables. Suppose on a 5-point satisfaction (5  $\equiv$  completely satisfied to 1  $\equiv$  completely dissatisfied) scale, following percentage or probabilities were observed.

Scale item ( $x_i$ )	1	2	3	4	5
Probabilities, Prob.	0.1	0.2	0.3	0.2	0.2

Then it follows that

$$\mathcal{E}(x) = 1 \times 0.1 + 2 \times 0.2 + 3 \times 0.2 + 4 \times 0.2 + 5 \times 0.2 = 3.2.$$

In this example, one would conclude that respondents are neither satisfied nor dissatisfied. If  $x$  and  $y$  are random variables and  $a$  and  $b$  are real constants then following expectation properties can be demonstrated.

1.  $\mathcal{E}(a + bx) = a + b\mathcal{E}(x)$ .
2.  $\mathcal{E}(x + y) = \mathcal{E}(x) + \mathcal{E}(y)$ .
3.  $\mathcal{E}(x \times y) = \mathcal{E}(x) \times \mathcal{E}(y)$ .
4.  $\mathcal{E}[g(x) + h(x)] = \mathcal{E}[g(x)] + \mathcal{E}[h(x)]$ .

Why might these properties be interesting? Can you demonstrate above properties are valid with examples?

### Variance, Covariance and Correlation

The variance of a random variable is an indication for the spread of a variable. Mathematically,

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

If  $x$  and  $y$  are random variables and  $a$  and  $b$  are real constants then following variance properties can be demonstrated.

1.  $\mathcal{V}(a + bx) = b^2\mathcal{V}(x)$ .
2.  $\mathcal{V}(x + y) = \mathcal{V}(x) + \mathcal{V}(y) + 2\mathcal{C}(x, y)$ .
3.  $\mathcal{V}(x - y) = \mathcal{V}(x) + \mathcal{V}(y) - 2\mathcal{C}(x, y)$ .

where  $\mathcal{C}(x, y) = \mathcal{E}[(x - \mu_x)(y - \mu_y)] = \mathcal{E}(xy) - \mathcal{E}(x)\mathcal{E}(y)$  or covariance between variable  $x$  and  $y$ . Note that  $\mu_y$  and  $\mu_x$  are expected values associated with variables  $y$  and  $x$  respectively. Note also that in general  $\mathcal{V}(x \times y) \neq \mathcal{V}(x) \times \mathcal{E}(y)$ .

Further, value of covariance is unbounded where as the correlation coefficient between two variables is always bounded from  $-1$  to  $1$ . The correlation ( $\rho_{xy}$ ) between random variable  $x$  and  $y$  is

$$\rho_{xy} = \frac{\mathcal{C}(x, y)}{\sqrt{\mathcal{V}(x)\mathcal{V}(y)}}$$

**Example:** Suppose another variable is added to consumer price index (CPI) data that indicates percent wage increase. Such observations are included below. What would be procedure to compute correlation between two indicators of inflation that reflect population size?

Country	CPI (%)	Wage increase (%)	Population (in millions)	Percent of total
Australia	2.5	4.0	18.91	0.0235
Austria	2.0	2.7	8.18	0.0102
Belgium	2.2	3.1	10.15	0.0126
Britain	0.9	4.4	58.74	0.0730
Canada	0.7	2.4	30.49	0.0379
Denmark	1.9	4.3	5.31	0.0066
France	1.2	4.1	58.89	0.0732
Germany	1.7	2.2	82.18	0.1022
Italy	2.4	2.8	57.34	0.0713
Japan	-1.0	-0.8	126.51	0.1573
Netherlands	4.2	4.6	15.85	0.0197
Spain	2.7	3.6	39.42	0.0490
Sweden	2.7	3.3	8.86	0.0110
Switzerland	0.3	1.3	7.14	0.0089
United States	1.9	3.9	276.22	0.3435
Simple Average	1.75	3.06	804.19	
Weighted Average	1.39	2.85		
Unweighted Variance	1.49	2.01		
Weighted Variance	1.42	2.97		

Following table summarizes calculations of variance, covariance and correlation, both un-weighted and weighted.

	Covariance	Correlation
Unweighted	1.1655	0.6734
Weighted	1.6741	0.8146

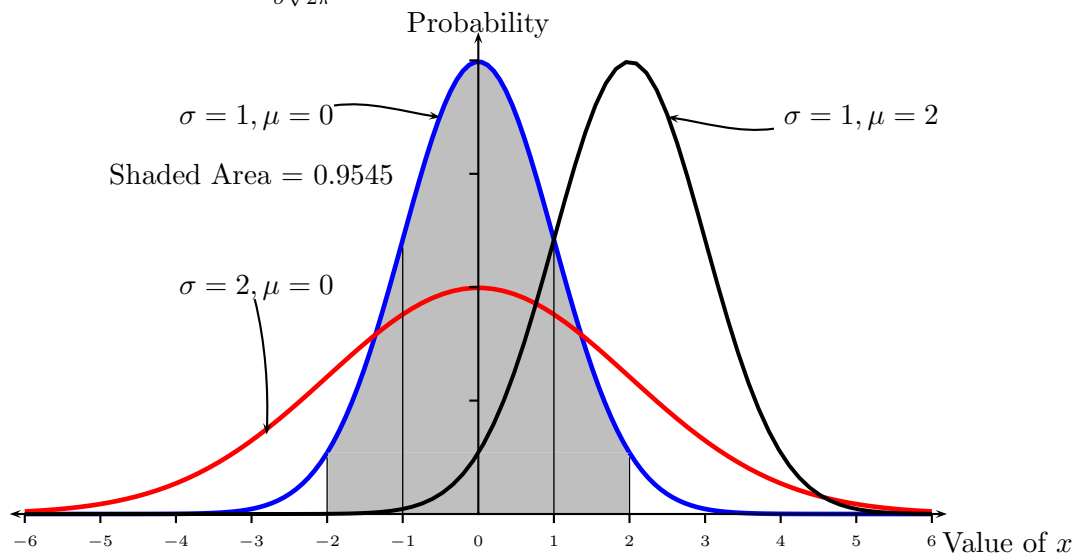
Comment on these estimates.

**Normal Distribution**

A random variable,  $x$ , is normally distributed if, and only if, its probability distribution function has the following form:

$$\text{Prob}(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{1}{2\sigma^2}(x - \mu)^2 \right], \quad -\infty < x < \infty.$$

This probability distribution function has two parameters: a location parameter  $\mu$ , with  $-\infty < \mu < \infty$ , and a scale parameter  $\sigma$ , with restrictions that  $0 < \sigma < \infty$ . It is often denoted by  $\mathcal{N}(\mu, \sigma^2)$ . Moreover, the function has a single mode or maximum frequency at  $x = \mu$  and such point  $\text{Prob}(x = \mu)$  is  $\frac{1}{\sigma\sqrt{2\pi}}$ . If  $\sigma$  is 1 (or the standard normal), then probability is about 0.4.



Several test procedures to determine whether an observed variable is normally distributed will be provided in Regression analysis.