Military Technical College
Kobry El-kobbah,
Cairo, Egypt

5th International Conference
on Electrical Engineering
ICEENG 2006

DETERMINATION OF THE GOODNESS OF FIT OF A DISTRIBUTION TO A SET OF EXPERIMENTAL DATA

Ashraf Mamdouh A. Aziz*

Abstract

Many systems of interest involve phenomena that exhibit unpredictable variation and randomness. For example, communication systems must provide continuous and error free communication over channels that are subject to random noise. Probability models are one of the tools that enable the designer to successfully build systems that are efficient and reliable. Processing of random signals postulates a probability model that is defined by the probability density function of the random signal. In this paper, we propose a method to determine the goodness of fit of a distribution to a set of experimental data. The proposed method depends on the chi-square test. It is applied to different examples of different probability density functions. The proposed method is proved to be efficient.

Keywords Probability density function – Fit of a distribution to data – Probability models – Histogram of random variables.

1. Introduction

A model is an approximate representation of a physical situation. It attempts to explain observed behavior using a set of simple and understandable rules. These rules can be used to predict the outcome of experiments involving the given physical situation. There are two main types of models; deterministic models and probability models [1]-[5]. In deterministic models the conditions under which an experiment is carried out determine the exact outcome of the experiment. In deterministic mathematical models, the solution of a set of mathematical equations specifies the exact outcome of the experiment. Circuit theory is an example of a deterministic mathematical model. Circuit theory models the interconnection of electronic devices by ideal circuits and predicts the observations. In practice there will be some variation in the observations due to measurement errors. Nevertheless, this deterministic model will be adequate as long as the deviation about the predicted values remains small.

* Associate Professor, Electrical Eng. Dept., Military Technical College, Cairo, Egypt

Many systems of interest involve phenomena that exhibit unpredictable variation and randomness. We will define a random experiment to be an experiment in which the outcome varies in an unpredictable fashion when the experiment is repeated under the same conditions. Deterministic models are not appropriate for random experiments since they predict the same outcome for each repetition of an experiment. The probability models are intended for random experiments.

Communication systems must provide error-free communication over channels that are subject to random interference and random noise [10, 11]. Many communication systems operate in the following way. Every T seconds, the transmitter accepts a binary input, namely a 0 or a 1, and transmits a corresponding signal. At the end of the T seconds, the receiver makes a decision as to what the input was; based on the signal it has received. Most communications systems are unreliable in the sense that the decision of the receiver is not always the same as the transmitter input. Thus, transmission errors occur randomly with probability ε. The optimum receiver structure at the receiver depends on a postulated probability model, which in many cases defined by the probability density function of the observed random signal. If the observed random signals and the postulated probability density functions are in good agreement, then we have a good fit; otherwise the receiver will be far away from its ideal performance. So an important question arises: How well does the model (the postulated probability density functions) fit the data (the received random signals)?.

2. Chi-Square Probability Density Function

The cumulative distribution function (cdf) of a random variable X is defined as the probability of the event $\{X \le x\}$ [1, 6, 9]:

$$F_X(x) = P[X \le x], \qquad \text{for } -\infty < x < +\infty, \tag{1}$$

that is, it is the probability that the random variable X takes on a value in the set $(-\infty, x)$. The probability density function of X (pdf) is defined as the derivative of $F_X(x)$:

$$f_X(x) = \frac{dF_X(x)}{dx} . \tag{2}$$

The pdf represents the density of probability at the point x in the following sense: The probability that X is in a small interval in the vicinity of x, that is, $\{x < X \le x+h\}$ is

$$P[x < X \le x+h] = F_X(x+h) - F_X(x) = \frac{F_X(x+h) - F_X(x)}{h}\, h . \tag{3}$$

If the cdf has a derivative at x, then as h becomes very small,

$$P[x < X \le x+h] \cong f_X(x)\, h . \tag{4}$$

The probability of an interval [a, b] is

$$P[a \le X \le b] = \int_a^b f_X(x)\, dx . \tag{5}$$

The probability density function of the gamma random variable has two parameters $\alpha > 0$ and $\lambda > 0$, and is given by

$$f_X(x) = \frac{\lambda(\lambda x)^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} , \qquad 0 < x < \infty, \tag{6}$$

where $\Gamma(z)$ is the gamma function, which is defined by the integral [7, 8, 12]:

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx, \qquad z > 0. \tag{7}$$

The pdf of the gamma random variable can assume a variety of shapes. By varying the parameters α and λ it is possible to fit the gamma pdf to many types of experimental data. By letting $\lambda = 1/2$ and $\alpha = \nu/2$, where ν is a positive integer, we obtain the chi-square random variable with ν degrees of freedom, which appears in certain statistical problems. The pdf of the chi-square random variable with ν degrees of freedom is then given by [2, 5]:

$$f_X(x) = \frac{x^{(\nu-2)/2}\, e^{-x/2}}{2^{\nu/2}\, \Gamma(\nu/2)}, \qquad x > 0. \tag{8}$$

3. Chi-Square Test

Suppose you have postulated a probability model for some random experiment, and you are now interested in determining how well the model fits the experimental data. In other word: How well does the model fit the data? In general, the mean and variance of a random variable do not provide enough information to determine the cdf/pdf.

In this section we use the chi-square test to determine the goodness of fit of a distribution to a set of experimental data. The chi-square test contain the following steps:

Assume that the random variable X has the postulated pdf .
Partition the sample space S_x into the union of K disjoint intervals.
Compute the probability b_k that an outcome falls in the k^{th} interval under the assumption that X has the postulated pdf.
The expected number of outcomes that fall in the k^{th} interval in n repetitions of the experiment is then:

$$m_k = n b_k . \tag{9}$$

The chi-square statistic is defined as the weighted difference between the observed number of outcomes, N_k, that fall in the k^{th} interval, and the expected number m_k:

$$D^2 = \sum_{k=1}^{K} \frac{(N_k - m_k)^2}{m_k} . \tag{10}$$

The chi-square test is based on the fact that for large n, the random variable D^2 has a pdf that is approximately a chi-square pdf with ν degrees of freedom.
(ν =K-1)
6- If the fit is good, then D^2 will be small.
7- The hypothesis is rejected if D^2 is two large, i.e. , if

$$D^2 \geq t_\alpha , \tag{11}$$

where t_α is a threshold determined by the significance level of the test. The threshold t_α can be computed by finding the point at which

$$P[X \geq t_\alpha] = \alpha , \tag{12}$$

where X is a chi-square random variable with ν degrees of freedom ($\chi^2_{\alpha,\nu}$).The thresholds for different values of significance levels and various degrees of freedom are found in Tables (see [2, 5, 9] and Table 1 for examples). For example, from Table 1, it is clear that the threshold for a 0.99 significance level and degree of freedom =30 is 14.954.

Table 1: Critical Values $\chi^2_{\alpha,\nu}$ (D^2) for the Chi-Squared Distribution

α						
ν	0.995	0.990	0.975	0.10	0.05	0.005
2	0.010	0.020	0.051	4.605	5.992	10.597
5	0.412	0.554	0.831	9.236	11.070	16.748
10	2.156	2.558	3.247	15.987	18.307	25.188
20	7.434	8.260	9.591	28.412	31.410	39.997
30	13.787	14.954	16.791	40.256	43.773	53.672
40	20.706	22.164	24.433	51.805	55.758	66.766

4. Applications of the Chi-Square Test to Different Probability Density Functions

In this section we simulate different probability density functions (experimental data) and compare them with the corresponding theoretical probability density functions. In each example, the histogram is obtained by generating 1000 samples of the random variables of the desired pdf. There are many ways of selecting the intervals in the partition, and that these can yield different results. The following rules of thumb are recommended. First, to the extent possible the intervals should be selected so that they are equiprobable. Second, the intervals should be selected so that the expected number of outcomes in each interval is five or more. This improves the accuracy of approximating the pdf of D^2 by a chi-square pdf. In our examples, the histogram is obtained by dividing the real line into 30 intervals of equal lengths. We can also divide the real line into 30 intervals of equal probability.

The discussion so far has assumed that the postulated distribution is completely specified. In the typical case, however, one or two parameters of the distribution, namely the mean and variance, are estimated from the data. It is often recommended that if r of the parameters of a pdf are estimated from the data, then D^2 is better approximated by a chi-square distribution with K – r – 1 degrees of freedom. In effect, each estimated parameter decreases the degrees of freedom by 1.

A narrow-band noise $n(t)$ can be represented in terms of its envelope and phase components as follows:

$$n(t)=r(t)\cos[2\pi f_c t+\phi(t)] \ , \qquad (13)$$

where the function $r(t)$ is called the envelope of $n(t)$, the function $\phi(t)$ is called the phase of $n(t)$, and f_c is the nominal carrier frequency. By letting R the random processes represented by the envelope $r(t)$, the probability density function of the random variable R will be [12]:

$$f_R(r)=\begin{cases}\dfrac{r}{\sigma^2}\exp(-\dfrac{r^2}{2\sigma^2}), & r\geq 0\\ 0, & \text{elsewhere}\end{cases} \quad , \tag{14}$$

where σ^2 is the variance of the original narrow-band noise $n(t)$. A random variable having the probability density function of Eq. (14) is said to be Rayleigh-distributed random variable. If we suppose that we add the sinusoidal wave $A\cos(2\pi f_c t)$ to the narrow-band noise $n(t)$, where A and f_c are both constant, a sample function of the sinusoidal wave plus noise is then expressed by:

$$x(t)= A\cos(2\pi f_c t)+n(t) \quad . \tag{15}$$

We assume that the frequency of the sinusoidal wave is the same as the nominal carrier frequency for the noise. In this case, the probability density function of the random variable R will be [12]:

$$f_R(r)=\frac{r}{\sigma^2}\exp(-\frac{r^2+A^2}{2\sigma^2})\ I_0(\frac{Ar}{\sigma^2}) \ , \tag{16}$$

where $I_0(x)$ is the modified Bessel function of the first kind of zero order and is defined as:

$$I_0(x)=\frac{1}{2\pi}\int_{2\pi}^{0}\exp(x\cos\phi)\ d\phi \quad . \tag{17}$$

The form of the probability density function of Eq.(16) is called the Rician distribution .

Figure 1 shows the plot of 1000 random variables of Rayleigh distribution (experimental samples). It is assumed that both the theoretical and experimental probability density functions are Rayleigh distributed random variables. Figure 2 shows the histogram of the Rayleigh random variables of the experimental samples. Figure 3 shows the theoretical probability density function. Figure 4 shows the experimental probability density function and compares it with the theoretical probability density function. It is clear that the theoretical probability density function is highly fitted the experimental probability density function. In our example, D^2 is 0.371 using a 1000 samples of the random variables and 30 equally disjoint intervals. Since the number of intervals is 30, then it is recommended to choose the degree of freedom to be 29 (ν = K-1 = 30 – 1). By comparing the value of D^2 with the table of the Chi-squared distribution [1,12] for ν=29, we conclude that the data is consistent with more than 99.5 % significance level.

Fig.1 Plotting the Rayleigh Random Variables

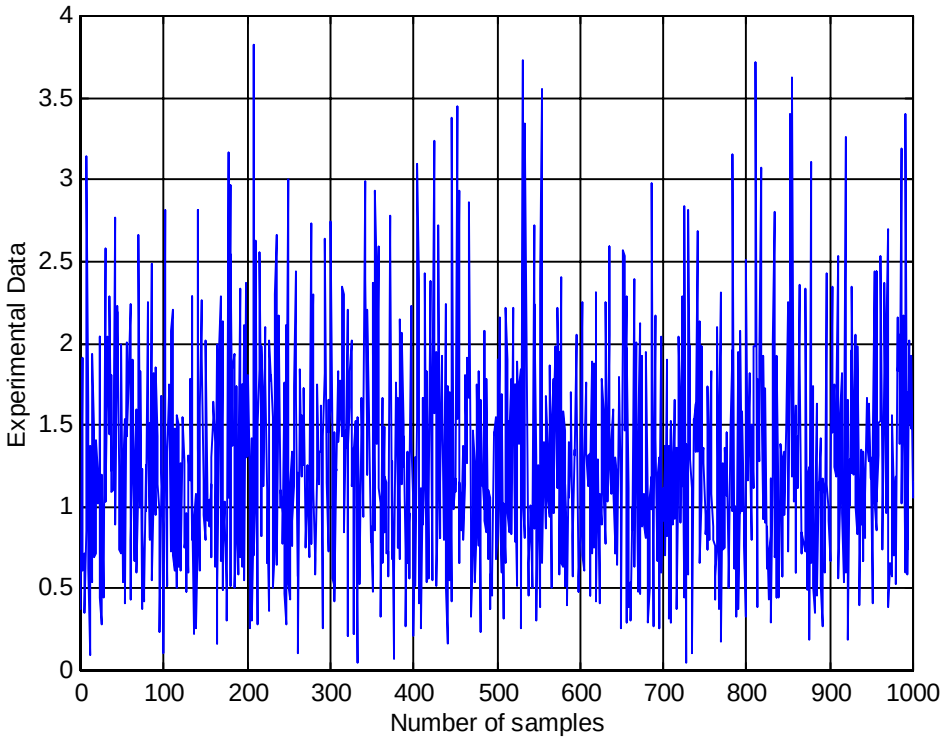


Fig.2 Histogram of the Rayleigh R. V.

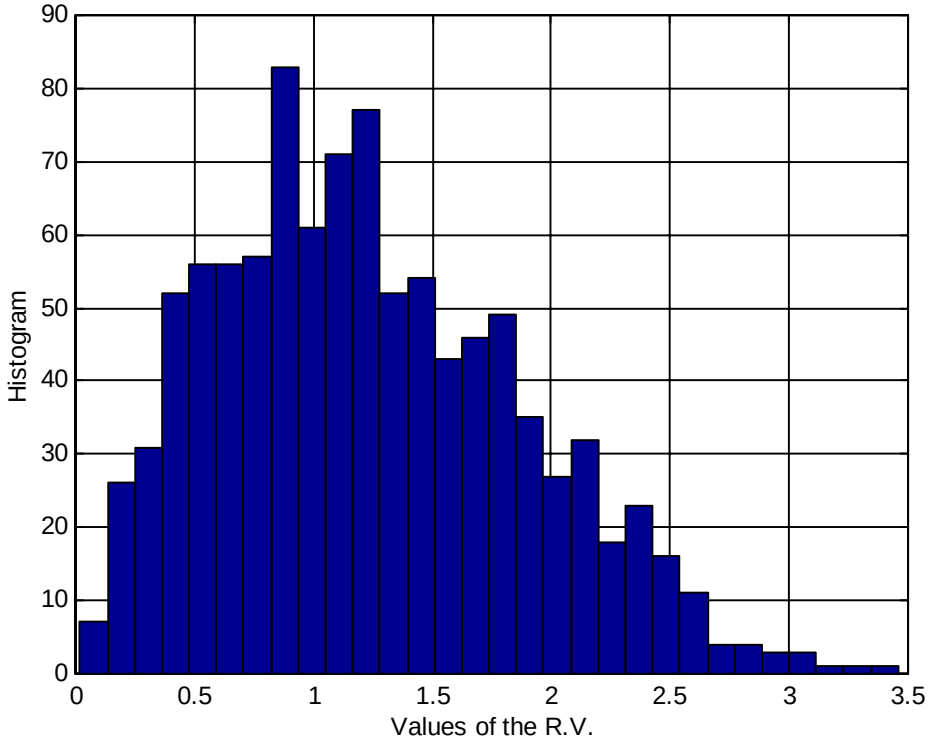


Fig.3 Theoretical pdf (Rayleigh)

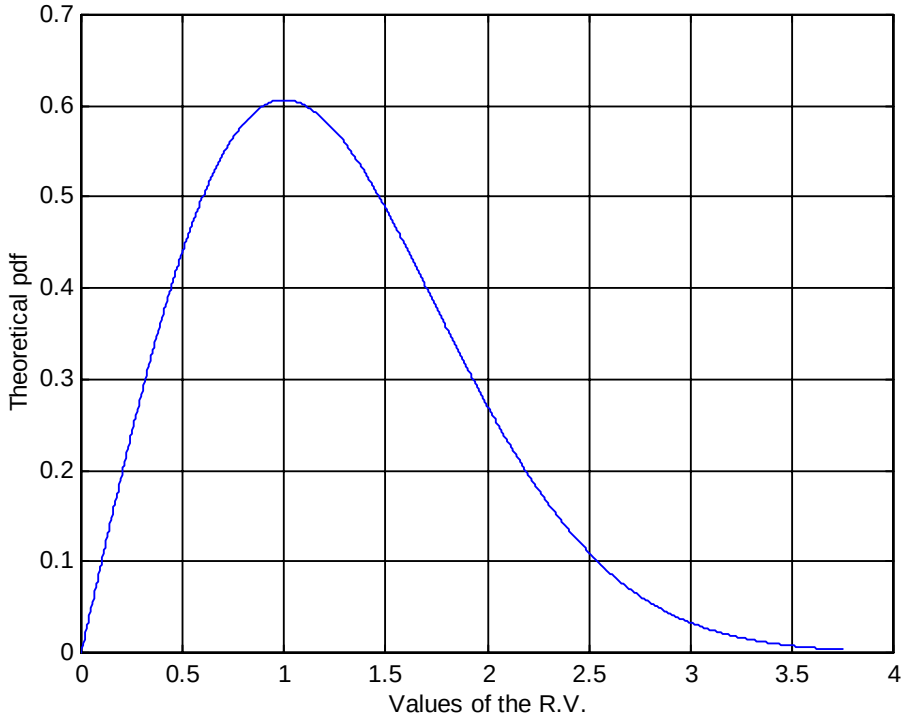
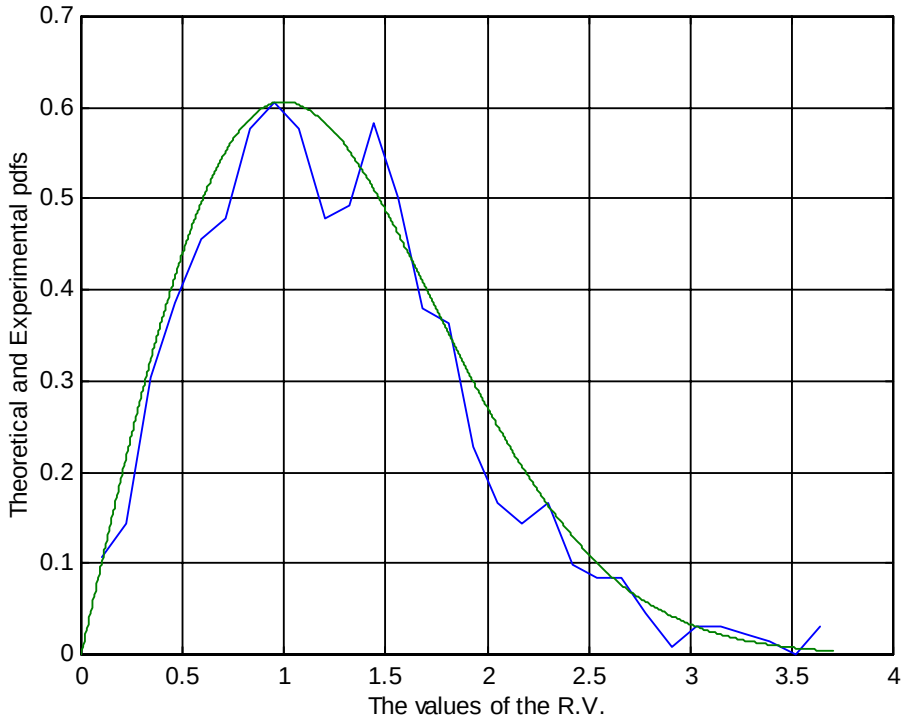


Fig.4 Comparison of The Theoretical(-) and Exp.(--) PDFs (RAYLEIGH)



Similar results are obtained in case of Chi-square, Rician, non-central Chi-square, uniform and Gaussian probability distribution functions. Figures 5 - 8 show the results in case of Rician probability density functions. In these figures, It is assumed that both the theoretical and experimental probability density functions are Rician distributed random variables. In this case, the value of D^2 for 29 degrees of freedom is 7.8833. By comparing the value of D^2 with the table of the Chi-squared distribution [1,11] for K=29, we conclude that the data is consistent with more than 99.9 % significance level.

Fig.5 Plotting the Rician Random Variables

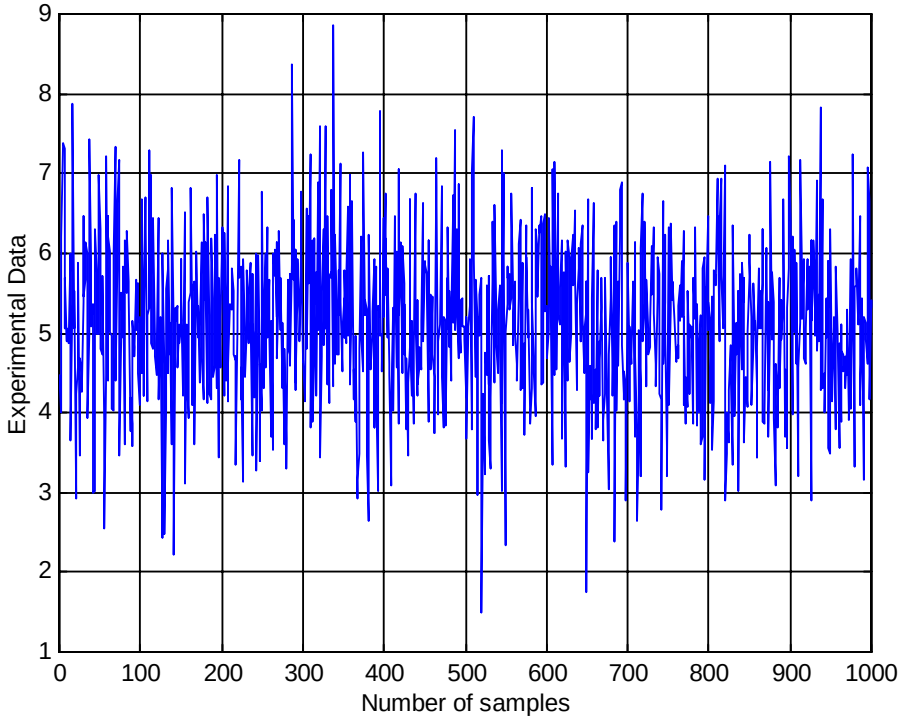


Fig.6 Histogram of the Rician R. V.

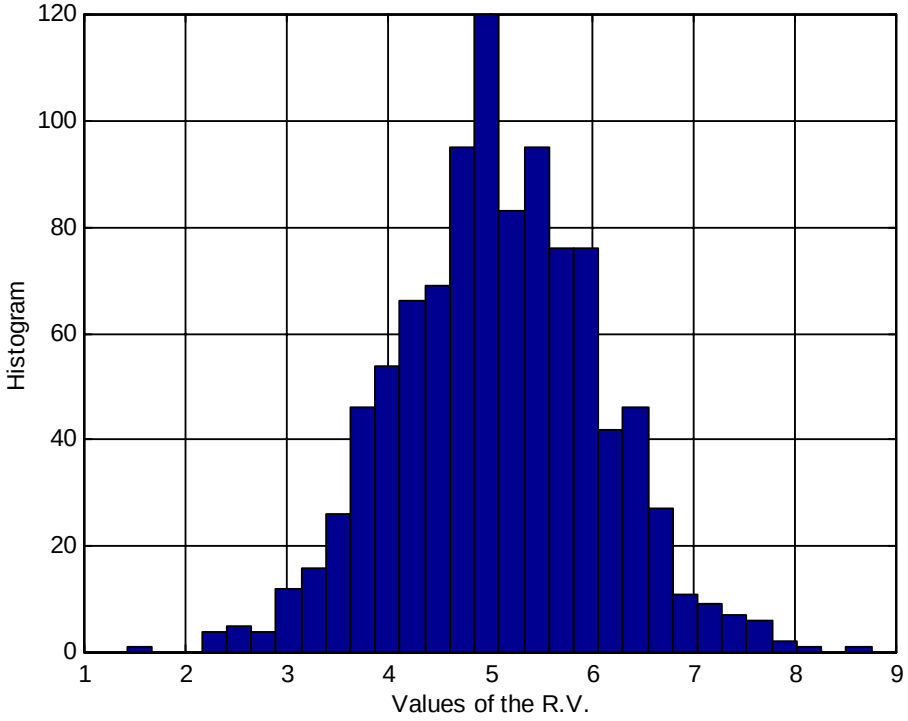


Fig.7 Theoretical pdf (Rician)

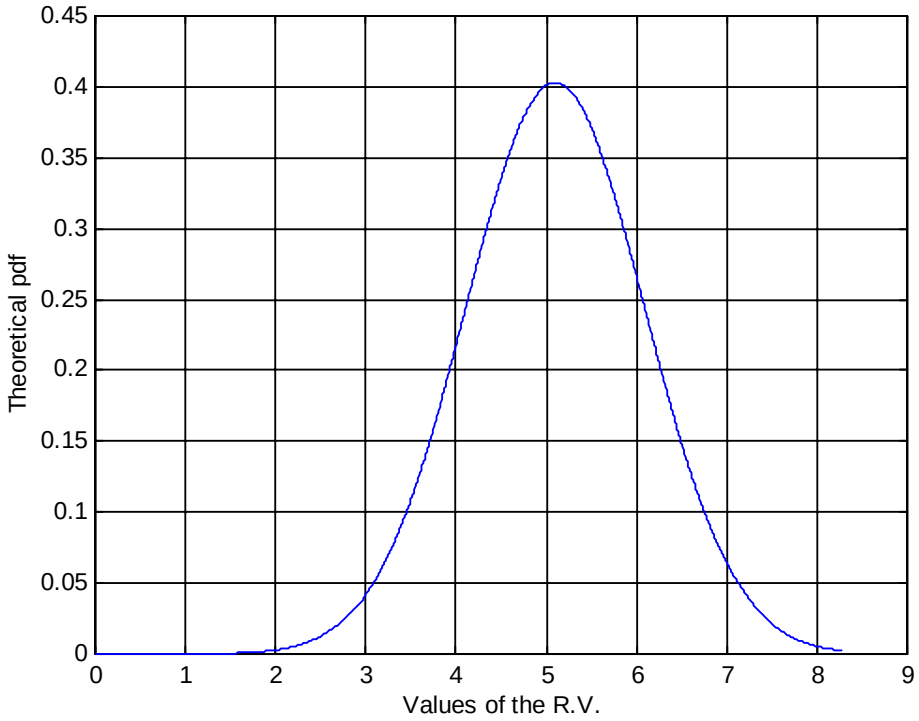


Fig.8 Comparison of The Theoretical(-) and Exp.(--) PDFs (Rician)

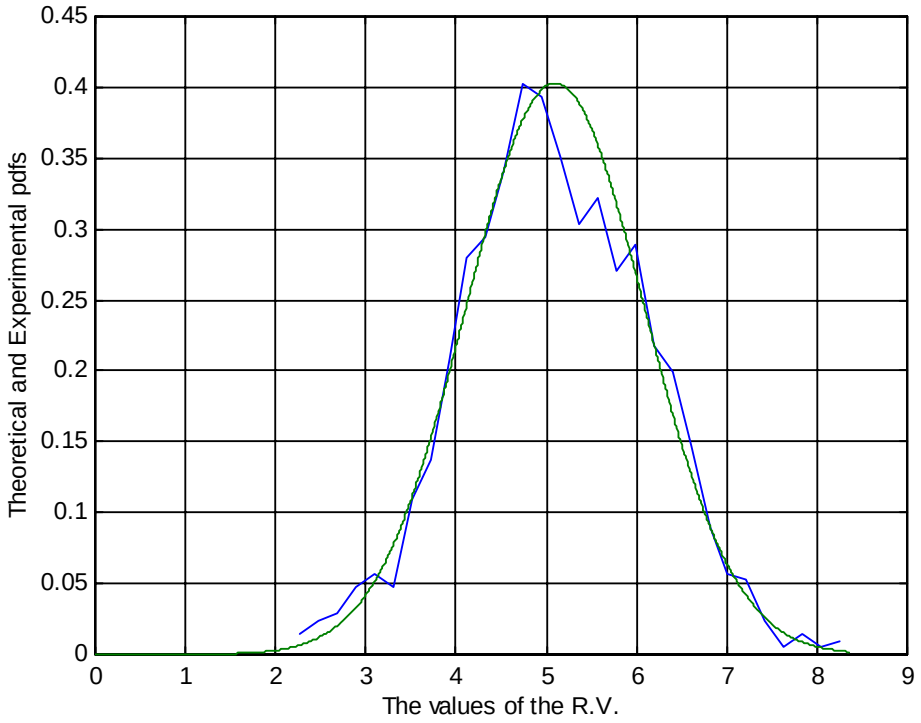
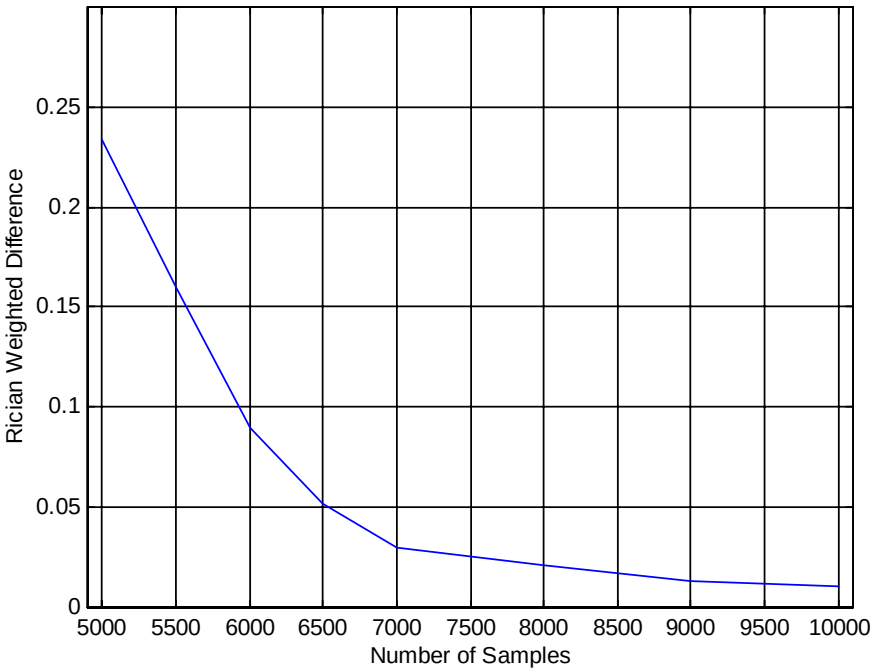


Table 2 compares the values of the first two moments ($E\{X\}$, $E\{(XE\{X\})^2\}$) of the theoretical and experimental probability density functions. The results show that the data is in good agreement with the Rayleigh and Rician distributions. The dependence of the weighted difference (D^2) on the number of simulated random variables (samplers) is shown in Figure 9. As shown in Figure 9, The value of the Chi-square weighted difference decreases as the number of simulated samples increases, i.e. the accuracy of Chi-square goodness of fit increases as the number of experimental data increases (expected result if the postulated pdf were correct).

Table 2: Comparison of the moments in case of Rayleigh and Rician distributions

	Rayleigh		Rician	
	Theoretical	Experimental	Theoretical	Experimental
Mean (m) = $E\{X\}$	1.2533	1.2320	5.1003	5.0818
$Var\{X\}= E\{(X-m)^2\}$	0.4292	0.4275	0.9801	0.9142

Fig.9 The Weighted Difference for Various Number of Samples



Also we considered an example when the postulated probability density function does not fit the experimental data. It is assumed that the theoretical probability density functions is Rayleigh distributed random variables with σ^2=1, while the experimental probability density functions is Exponential distributed random variables with σ^2=1 and. Figure 10 shows the experimental probability density function and compares it with the theoretical probability density function. It is clear that the theoretical probability density function (Rayleigh distributed random variables) does not fit the experimental probability density function (Exponential distributed random variables). In our example, D^2 is 132.8 using a 1000 samples of the random variables and 31 equally disjoint intervals. By comparing the value of D^2 with the tables of the Chi-squared distribution [1,12] and Table 1, for ν=30, we conclude that the data is not consistent at all. In this example, the consistence level is almost zero.

Fig.11 and 12 show the performance of a tracking system that tracks two crossing targets in case of incorrect (when the postulated probability density function does not fit the experimental data) and correct (when the postulated probability density function fits the experimental data) modelings respectively. It is clear that the performance of the tracker is very poor in case of incorrect modeling. Fig.13 compares the performance of a PCM (pulse code modulation) communication system in case of correct and incorrect modelings for different signal to noise ratios (SNR). It is clear that the probability of error (Pe) in case of correct modeling is much smaller than the Pe in case of incorrect modeling for all values of the signal to noise ratios. Figures 11-13 show the importance of the proposed test.

Fig.10 Theoretical:Rayleigh(-)and Experimental:Exponential(--) pdfs

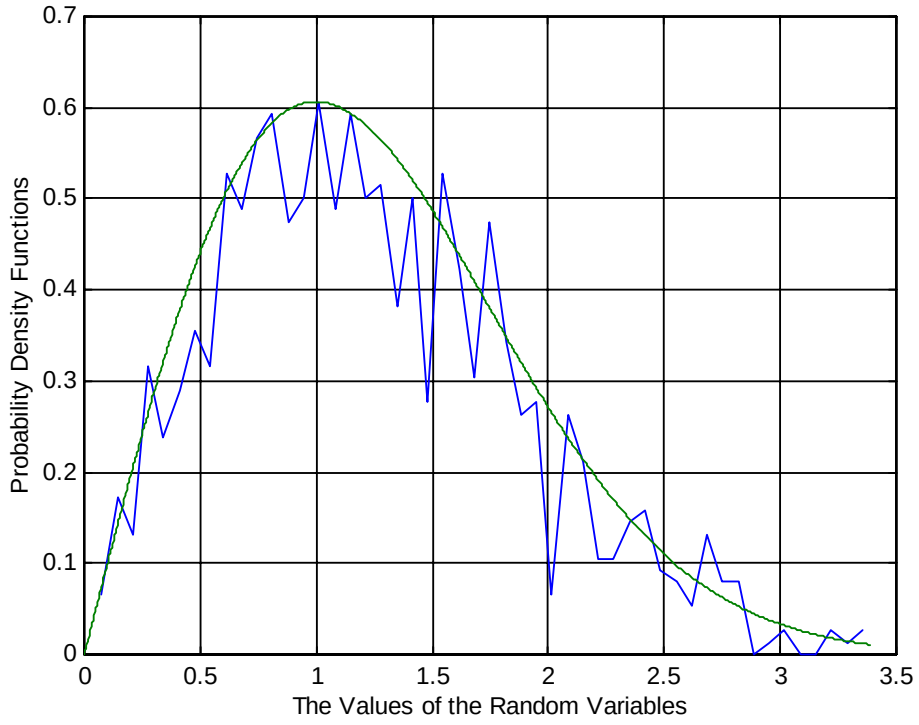


Fig.11 Tracking in Case of Incorrct Modeling

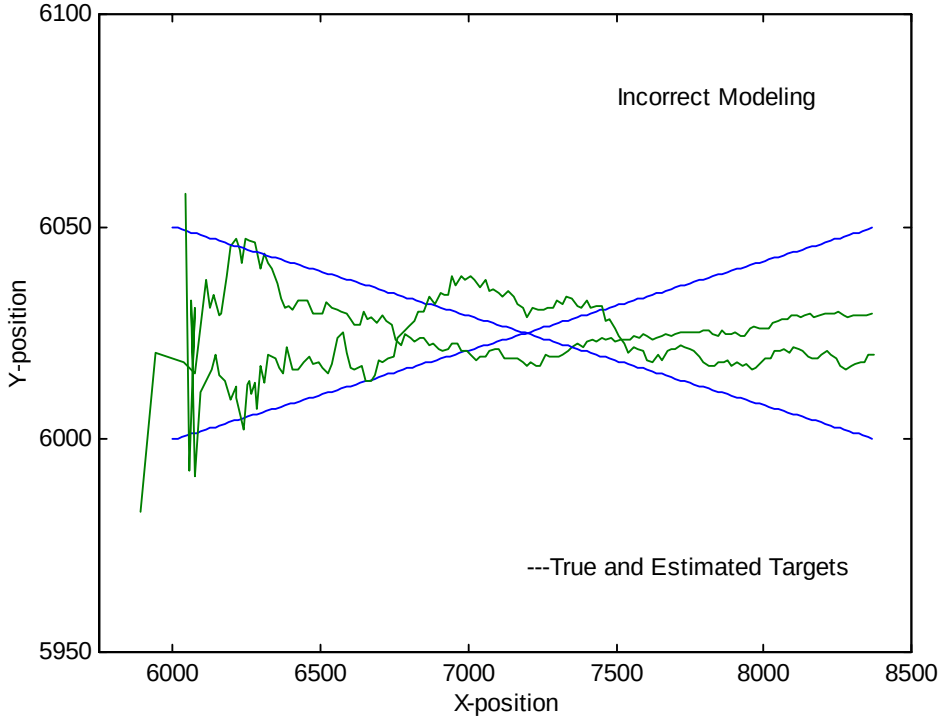


Fig.12 Tracking in Case of Corrct Modeling

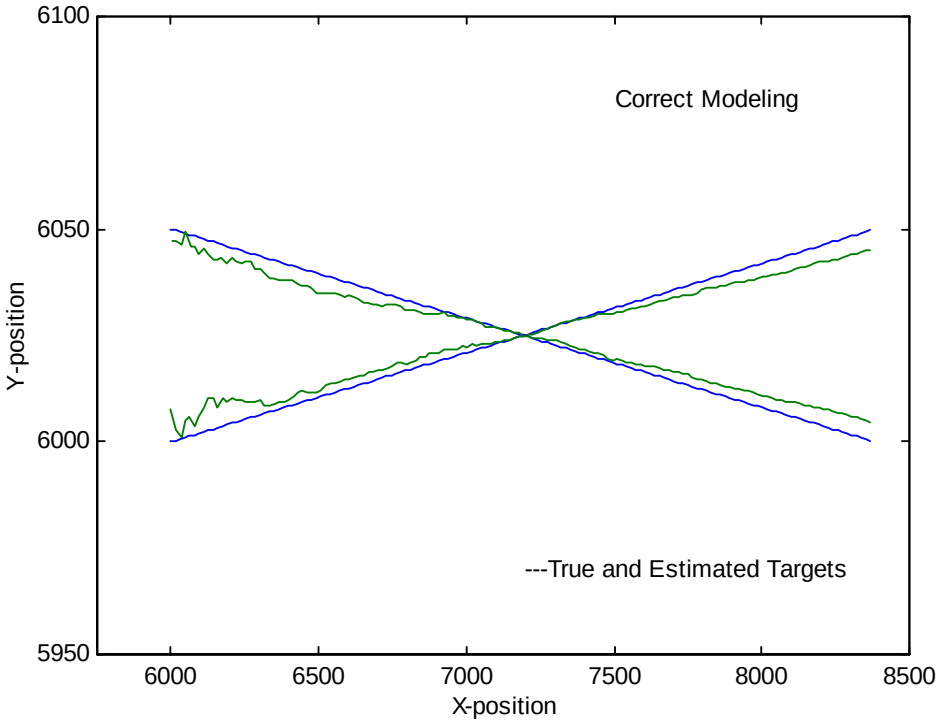
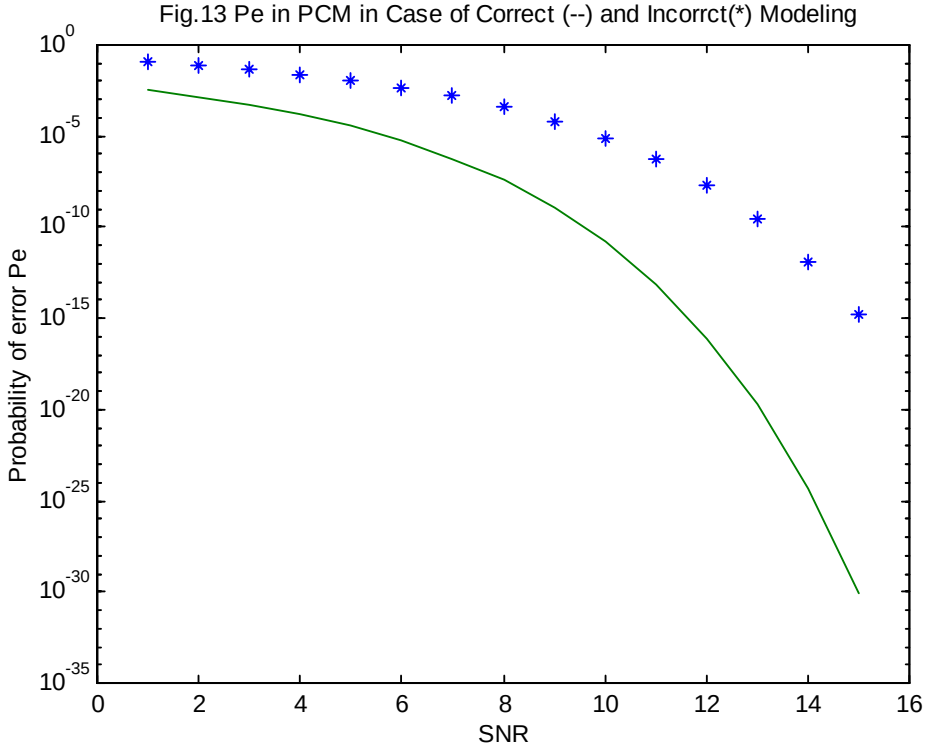


Fig.13 Pe in PCM in Case of Correct (--) and Incorrct(*) Modeling



Conclusion

We are interested in determining how well a postulated probability model fits an experimental probability model .In this paper, a method to determine the goodness of fit of a distribution to a set of experimental data has been proposed. The proposed method depends on the Chi-square test to carry out the comparison between the postulated probability density function and the experimental probability density function. The chi-square statistic is defined as the weighted difference between the observed number of outcomes and the expected number. The weighted difference is compared with a threshold determined by the significance level of the test. If the fit is good, then the weighted difference will be small. Therefore the hypothesis is accepted if the weighted difference is too small and vice versa. The proposed method is applied to different examples of postulated and experimental probability models and is proved to be efficient and accurate. The importance of the proposed test is investigated by considering two practical examples. Comparison of the systems performances in case of correct and incorrect modelings has been done. The results show that the proposed test is essential.

References

[1] Alberto Leon Garcia, Probability and Random Processes for Electrical Engineering, Second Edition, Addison-Wesley Publishing Company, New York, 1995.
[2] Ronald E. Walpole, Raymond H. Myers, and Sharon L. Myers, Probability and statistics for Engineers and Scientists, Sixth Edition, Prentice-Hall Inc., 1998.
[3] L. Breiman, Probability and Stochastic Processes: With a view Toward Applications, Houghton Mifflin, Boston, 1969.
[4] A. Papoulis, Probability, Random Variables, and Stochastic Processes, McGraw-Hill, New York, 1965.
[5]A. B. Clarke and R. L. Disney, Probability and Random Processes: A First Course With Applications, Wiley, New York, 1985.
[6] C. W. Helstrom, Probability and Stochastic Processes for Engineers, Macmillan, New York, 1984.
[7] K. L. Chung, Elementary Probability Theory, Springer-Veralg, New York, 1974.
[8] A. M. Law and W. D. Kelton, Simulation Modeling and Analysis, McGraw-Hill, New York, 1982.
[9] P. L. Meyer, Introductory Probability and Statistical Applications, Addison-Wesley, Reading, Mass., 1970.
[10] A. V. Oppenheim and R. W. Schafer, Digital Signal Processing, Prentice-Hall, Englewood Cliffs, N. J., 1975.
[11] H. Cramer, Mathematical Methods of Statistics, Princeton University Press, Princeton, N. J., 1964.
[12] S. Haykin, Communication Systems, Wiley, New York, 1994.