Understanding Histograms, Probability and Probability Density Using MATLAB

Gordana Jovanovic Dolecek [1, 2] and fred harris[2]
[1]Department of Electronics Institute INAOE, Puebla, Mexico
E-mail: gordana@inaoep.mx
[2]Department of Electrical Engineering, SDSU San Diego,USA
E-mail: fred.harris@sdsu.edu

Abstract

This paper presents an attractive way to introduce the fundamental terms used to describe a random variable using a MATLAB environment. The Uniform and the Gaussian random variables are considered. The demo programs include histograms, probability, probability density and distribution functions. The results of the evaluation of the program are also presented.

1. Introduction

The way engineering is being taught has changed in recent years with the introduction of commercial and educational software that enable and facilitate a better understanding of the subject matter and increased teaching efficiency. Students learn better, remember longer and are better able to identify the appropriate concepts to solve new problems when they learn by addressing concrete problems and actively participate in exploration and pursuit of knowledge[1].

It is known that random variable is generally considered one of the most abstract and conceptually difficult areas in engineering education and the teaching of random variables is one of the subjects that requires more time for its understanding[2,3]. The use of computers gives students the visual and intuitive representation of the random variables which had traditionally been stated in terms of abstract mathematical description. To this end we present a demo program to aid in, and improve of, understanding the different terms used to describe a random variable.

The programs are written in MATLAB in form of *m* files. We choose MATLAB because MATLAB along with the accompanying toolboxes is the tool of choice for most educational and research purposes[4-7]. It provides powerful computation and advanced visualization tools and is also available on a number of hardware platforms.

At each step, the program provides the user with all necessary instructions, including what to do in the next step. Not only passive, but also an active role of the user is required during interactive dialogues prompted through the program. The additional advantage of our approach is that the student does not need MATLAB or any other programming language experience.

The programs can be used as a complement to theoretical classes or alone as a self-study tool. The next section provides a brief description of the probability density and distribution functions, mean value and variance. Third Section presents the demo HISTOGRAMS. The relation of probability with the density and distribution functions is demonstrated in the Fourth Section. Last Section relates the mean value and the variance to the shape of the Gaussian density function.

2. Theoretical background

The Cumulative distribution function (CDF) of a random variable *X* is defined as the probability that the variable is less or equal to any value of *x*

$$F_X(x) = P\{X \le x\}, \quad -\infty < \mathrm{x} < \infty . \tag{1}$$

The axioms of probability and their corollaries imply that the distribution has the following properties [8, 9],

$$\begin{aligned} &1.\ 0 \le F_X(x) \le 1 .\\ &2. F_X(-\infty) = 0; F_X(\infty) = 1 .\\ &3.\ F_X(x_1) \le F_X(x_2); x_1 \le x_2 .\\ &4.\ P\{x_1 < X \le x_2\} = F_X(x_2) - F_X(x_1) \end{aligned} \tag{2}$$

The probability density function (PDF) of *X*, is defined as the derivative of $F_X(x)$

$$f_X(x) = \frac{dF_X(x)}{dx} . \tag{3}$$

Some properties of the PDF are[8, 9] ,

$$\begin{aligned} &1. f_X(x) \ge 0\\ &2.\ \int_{-\infty}^{\infty} f_X(x)dx = 1\\ &3. F_X(x) = \int_{-\infty}^{x} f_X(x)dx \end{aligned} \tag{4}$$

When we generate a random variable in a computer we have *N* values of random variable. In order to estimate the PDF of the random variable *X* we divide the range of the variable into *M* equidistant cells of width Δx .

Let N_i be the number of values of random variable *X* that belong to the *i*-th cell. Then the probability that the random variable belongs to the *i*-th cell is approximated by the quantity N_i/N called the relative frequency[7],

$$P\{(i-1)\Delta x < X \le i\Delta x\} \approx N_i / N , \tag{5}$$

where *i*=1,…,*M*.
Using (3) and (2) we estimate PDF in i-th cell as

$$f_X(i\Delta x) \approx \frac{P\{(i-1)\Delta x < X \le i\Delta x\}}{\Delta x} = \frac{N_i}{N}\frac{1}{\Delta x} \tag{6}$$

The mean value *m* of the random variable *X* is defined as

$$m = E\{X\} = \int_{-\infty}^{\infty} x f_X(x)dx \,. \tag{7}$$

The mean value provides us with very limited information about X. We are interested not only in the mean of a random variable, but also in variation of random variable about its mean[8]. The variance, σ^2_x, of the random variable X is defined as the expected average or mean of the squared deviation of X about its mean value. The parameter σ_x, the square-root of σ^2_x, is called the standard deviation or Root Mean Square (RMS) vale of the random variable.

$$\sigma^2_X = E\left\{(X - E\{X\})^2\right\} = \int_{-\infty}^{\infty} (x - E\{X\})^2 f_X(x)dx \,. \tag{8}$$

From (6) it follows

$$\sigma^2_X = E\{X^2] - m^2 \;, \tag{9}$$

where

$$E\{X^2\} = \int_{-\infty}^{\infty} x^2 f_X(x)dx \,. \tag{10}$$

The Gaussian random variable has the density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-m)^2}{2\sigma^2}} \;, \; -\infty < x < \infty \tag{11}$$

where m is the mean value, σ is the standard deviation, and σ^2 is the variance. The corresponding CDF is

$$F_X(x) = P\{X \le x\} = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x} e^{-\frac{(x-m)^2}{2\sigma^2}} dx \;\; . \tag{12}$$

The Uniform random variable in the interval $[R_1, R_2]$ has the density function

$$f_X(x) = \begin{cases} \dfrac{1}{R_2 - R_1} & R_1 \le x \le R_2 \\ 0 & otherwise \end{cases} \,. \tag{13}$$

3. Histograms

In this demo program we relate the histogram and the probability density function. We consider the uniform random variable in the interval $[R_1, R_2]$. The user is asked to choose the number of values N and the values R_1 and R_2. Figure 1 shows the plot of the variable for the first 1000 samples of N=100000, and R_1=1, R_2=2.

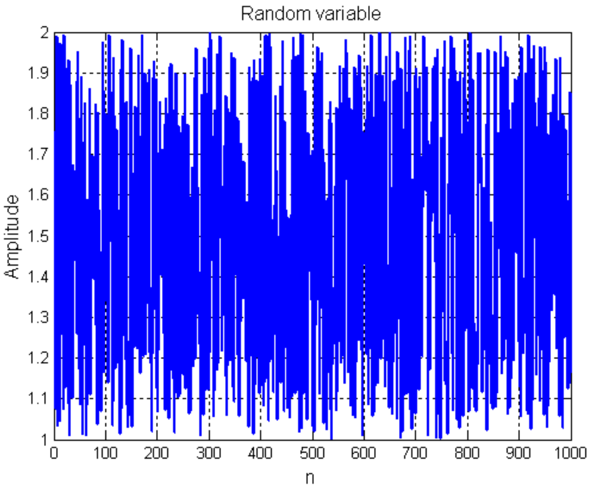


Fig.1. Uniform random variable.

The range of the variable (in this example [1, 2]) is divided into M equal cells. The user chooses the number of cells M.

HISTOGRAM *hist*(x,M) shows the values N_i, i=1,.., M, i.e. how many values of the random variable x are in each cell, where M is the number of cells. Figure 2 shows the histogram of the uniform random variable for M=50.

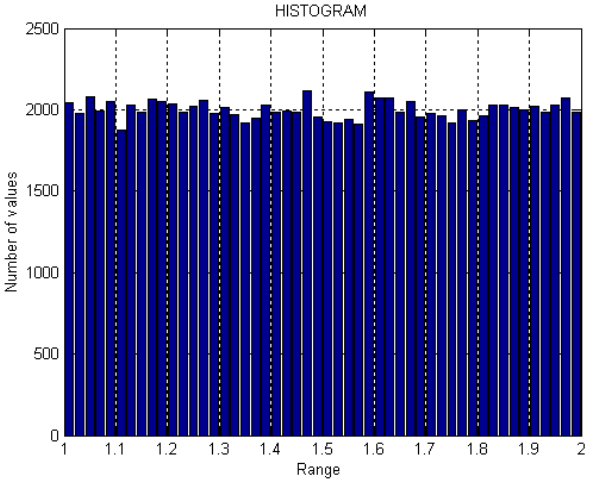


Fig.2. Histogram.

According to (5) we have.

$$P\{X \text{ belongs to the cell}\} = N_i \,/\, N = hist(x,\, M)/N\,. \tag{14}$$

The plot of the probabilities for all cells is shown in Fig.3.

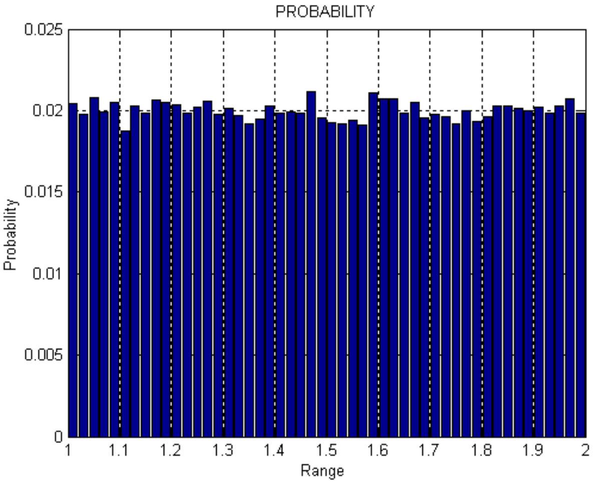


Fig.3. Plot of probabilities.

According to (6) the estimation of the probability density function in the given cell is obtained dividing the probability that the random variable belongs to the cell with the length of cell.

$$PDF= P\{X \text{ belongs to the cell}\}/\Delta=hist(x,\mathrm{M})/(N\Delta), \qquad (15)$$

where Δ is the length of the cell. The PDF estimation is given in Fig.4.

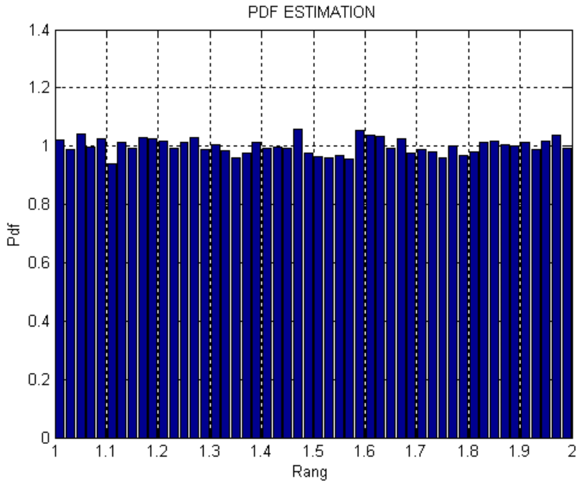


Fig.4. PDF estimation.

4. Probability, PDF and CDF

First we demonstrate how the probability that the random variable X is less then chosen value A is related with its density and distribution functions. To this end we generate the Gaussian random variable with the parameters m and σ^2. Figure 5 shows the Gaussian variable, PDF, and CDF for $m = 0$ and $\sigma^2 = 4$.

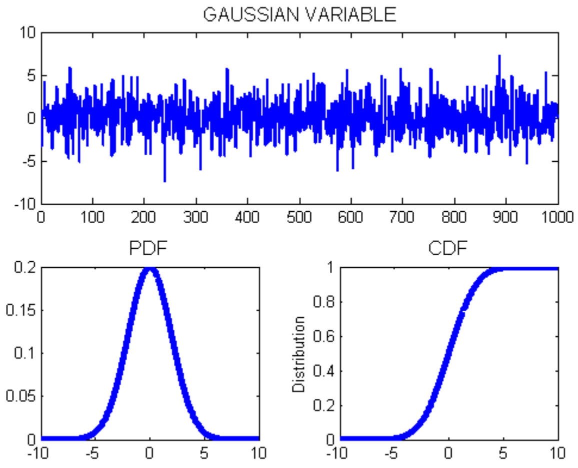


Fig.5. Gaussian variable.

The user chooses the value *A* to find the probability that the random variable *X* is less than *A*, as illustrated in Fig.6 for *A*=2. The corresponding probability is equal to 0.8413 and corresponds to the shaded surface under the PDF. The same probability presents the point in CDF at *n*=2.

Next, the user is asked to chose the interval [*B*1, *B*2] and find the probability that the random variable belongs to this interval, as shown in Fig.7 for the interval [-2, 3].

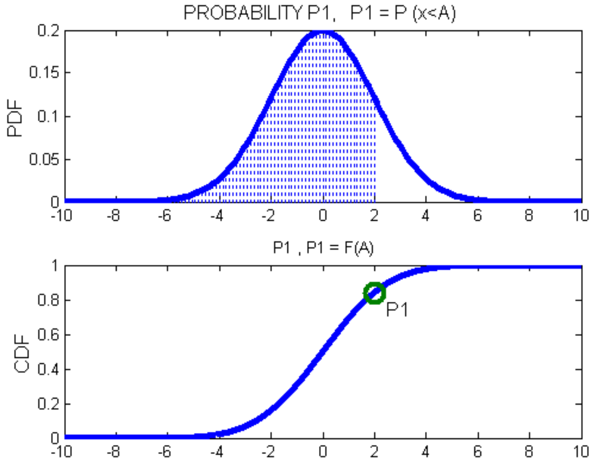


Fig.6. Probability that the random variable *X* is less then 2.

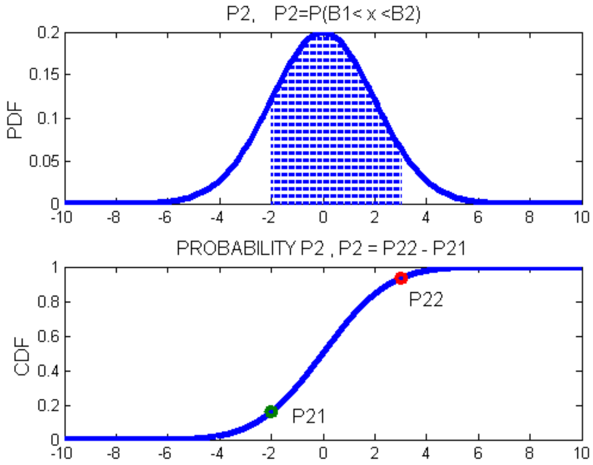


Fig. 7. Probability that the random variable belongs to the interval [-2, 3].

The probability corresponds to the shaded area under the PDF and to the difference of the values of CDF at points 2 and 3: $P_2 = P_{22} - P_{11} = 0.9332 - 0.1587 = 0.7745$.

5. Mean value, Variance and PDF

In this demo program we consider how the mean value and variance affects the shape of the density function. To this end we generate the Gaussian variable with given variance and 4 different mean values, as shown in Fig.8 for $\sigma^2 = 4$ and $m = 0, 2, 4,$ and 6.

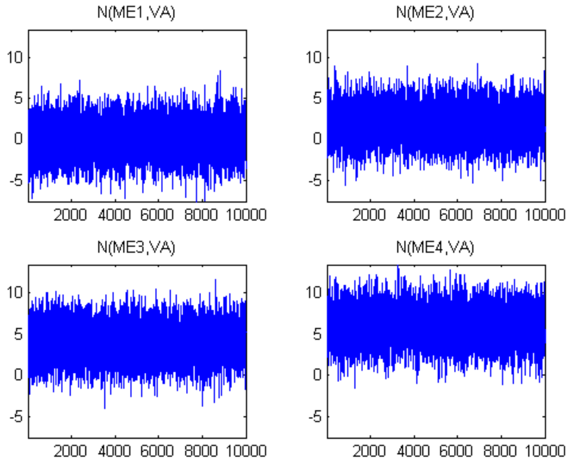
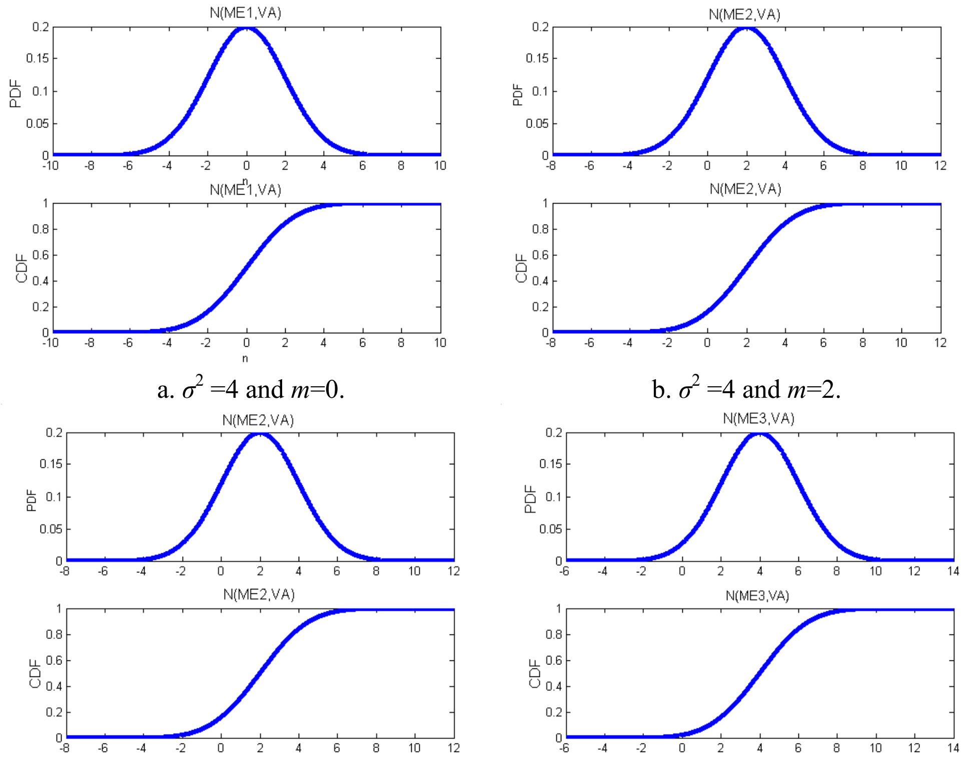


Fig.8. Gaussian variables with different mean values.

Note that the signals show the same behavior but they are displaced around the y axis. The corresponding densities and distributions have the same shape and are translated around the *x* axis as demonstrated in Fig.9.



a. σ^2 =4 and m=0.

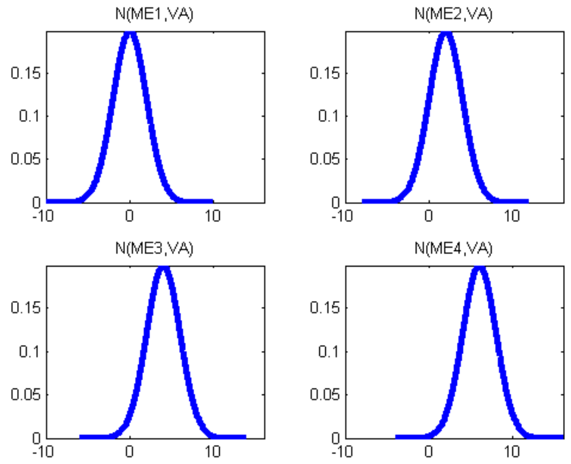
b. σ^2 =4 and m=2.

c. σ^2 =4 and m=4.

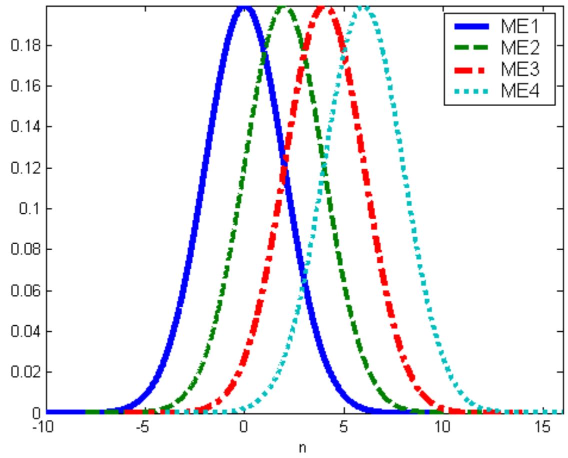
d. σ^2 =4 and m=6.

Fig.9. Gaussian densities and distributions.

In order to compare the densities from Fig.9 are again plotted in Fig.10.



a. Gaussian densities.



b. Gaussian densities.

Fig.10. Comparison of Gaussian densities having the same variance and different mean values.

Next we generate the Gaussian variables with the same mean value and the different values of variances. Figure 11 presents the signals for m=0 and σ^2=1, 4, 9, and 16.

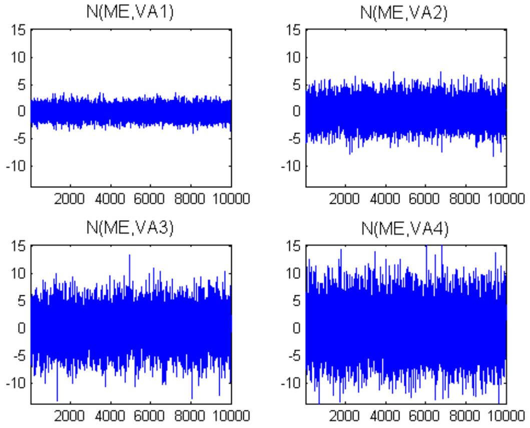
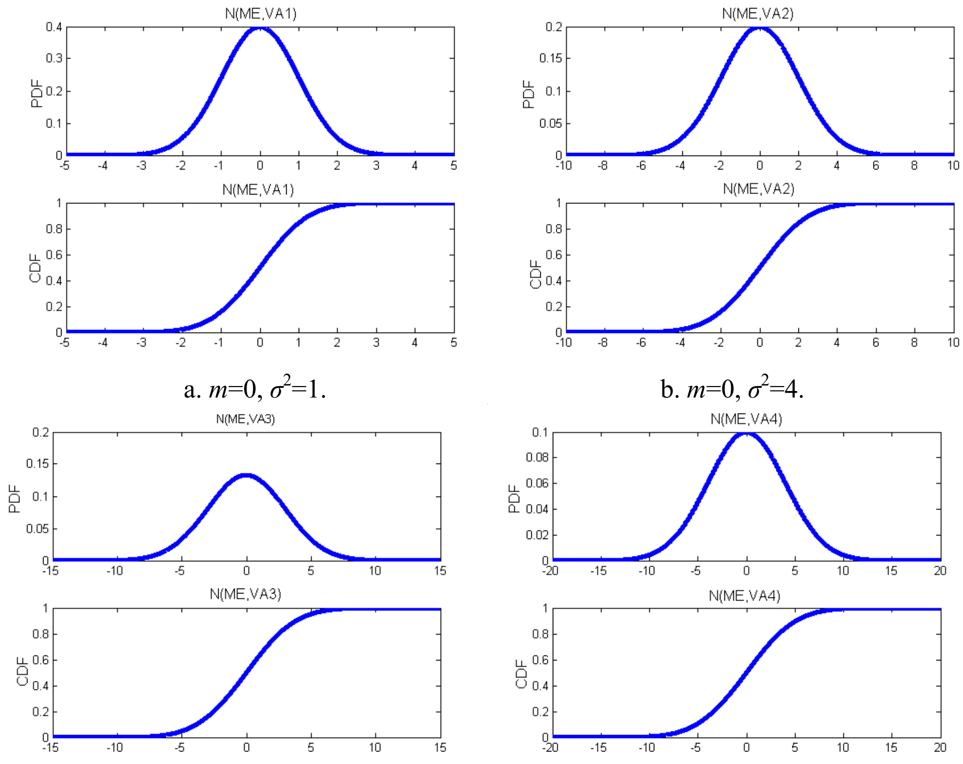


Fig.11.Gaussian variables.

Note that that the Gaussian signals with a high value of variance exhibit more dissipation of its values around its mean value, and vice versa. The corresponding densities and distributions are shown in Fig.12. The densities are compared in Fig.13. The PDF becomes narrower with the decreasing of its variance. Similarly the PDF peak values are decreasing with the increasing of the values of variances. As a consequence, the densities become more spread about its mean value.



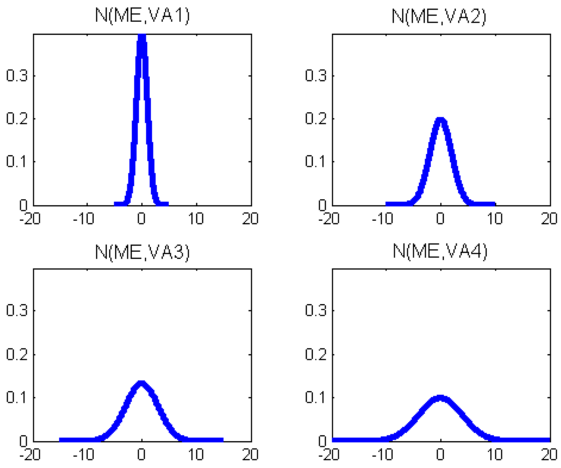
a. m=0, σ^2=1.

b. m=0, σ^2=4.

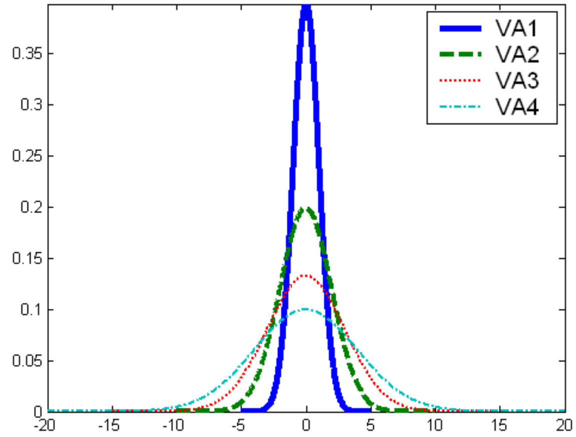
c. m=0, σ^2=9.

d. m=0, σ^2=16.

Fig.12. Gaussian densities and distributions.



a. Gaussian densities.



b. Gaussian densities.
Fig.13. Comparisons of Gaussian densities with different variances.

6. Evaluation

We consider that an important factor in application of educational software is to measure the usefulness of the software in the teaching-learning process. To this end we defined the evaluation form with set of questions attempted to test the usefulness of the software and the quality of its design features. All questions are rated with marks varying from 1 to 4; with the latter being the highest mark. The following set of questions has been asked:

1. Justification for the computer use in teaching random variables. (1=unjustified; 4=absolutely justified).
2. Contribution to study of random variables by demo program use. (1=irrelevant; 4= very effective).
3. Clarity of explanations and features of demo. (1=confusing; 4=absolutely clear).
4. Did this demo help you to understand better the Histograms & PDF? (1=NO; 4=Absolutely YES).
5. Did this demo help you to understand better the Probability, PDF, &CDF? (1=NO; 4=Absolutely YES).
6. Did this demo help you to understand better the Mean value & PDF? (1=NO; 4=Absolutely YES).
7. Did this demo help you to understand better the Variance & PDF? (1=NO; 4=Absolutely YES).
8. Did this demo help you to understand better the characteristics of Gaussian random variable? (1=NO; 4=Absolutely YES).
9. Special knowledge or programming skills required. (1=excessive; 4=null).
10. Ease of operation. (1=complex; 4= very easy).
11. General quality of presentation (figures, resolution, visibility, etc). (1=pure; 4=excellent)

The demo program is used as a complimentary tool to teaching basic course of Random signals and processes.Fig.14 presents the result of the evaluation in terms of the average marks for all questions.

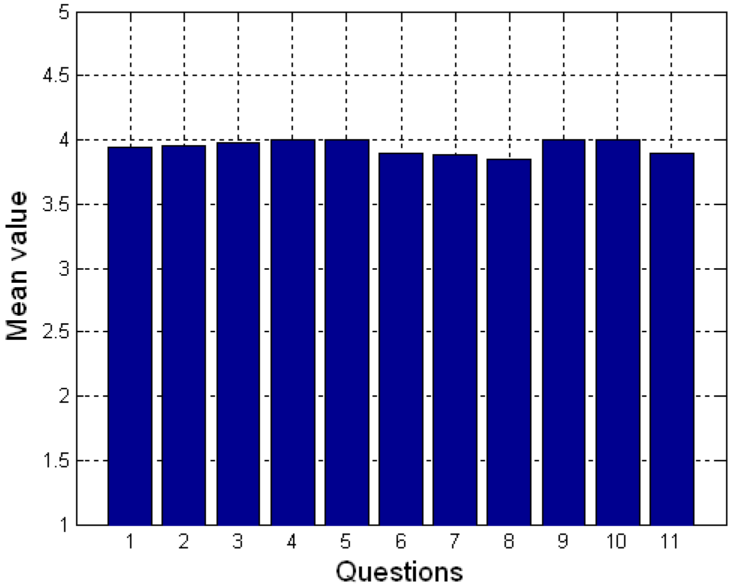


Fig.14. Rating scheme.

The result of the demo evaluation shows that students liked this way of teaching random variables. More specifically, they find this demo very useful for understanding histogram, PDF, probability and the characteristics of Gaussian random variable. They especially highly rated the features of the program; ease of operation, and no programming skills required to run the program.

7. Concluding remarks

This paper presents demo program for teaching the fundamental terms used to describe a random variable using a MATLAB environments. However the user does not need MATLAB or any other programming language experience. At each step, the program provides the user with all necessary instructions, including what to do in the next step. Additionally, the advantage of our demo program is that not only passive, but also an active role of the user is required during interactive dialogues prompted through the program. This program has been used to teaching basic course of Random Signals and Processes. Students have evaluated this program as very friendly and useful for better understanding of the basic terms used in random variables and processes.

Acknowledgement

This work was supported in part by the CONACYT under Grant No. 91013.

References

1. Li, S. G. and Lie, Q. (2004). Interactive Groundwater (IGW): An Innovative Digital Laboratory for Groundwater Education and Research, *Computer Applications in Engineering Education,* 11(4), 179‑203.
2. Jovanovic Dolecek, G. (1997). RANDEMO: Educational Software for Random Signal Analysis, *Computer Applications in Engineering Education*, 5 (2), 93-99.
3. Quere, R., Lalande, M., Boutin, J. N., and Valente, C. (1995). An Automatic Characterization of Gaussian Noise Source for Undergraduate Electronics Laboratory, *IEEE Transaction on Education*, vol.38, No2, 126-130.
4. Kim A. S., Park C., and Park, S. H. (2003). Development of web-based Engineering Numerical Software (WENS) Using MATLAB: Applications to Linear Algebra*, Computer Applications in Engineering Education,* 11 (2), 67‑75.
5. Orsak G. C., Etter D. M. (1995). Collaborative Signal Processing Education Using the Internet and MATLAB. *IEEE Signal Processing Magazine*, 12 (6), 23‑32.
6. Pires V. F. and Silva J. F. A. (2002). Teaching Nonlinear Modeling, Simulation and Control of Electronic Power Converters Using MATLAB/SIMULINK , *IEEE Trans. Education,* 45(3), 253‑261.
7. Prabhu G. S., and Shankar P. M. (2002) Simulation of Flat Fading Using MATLAB for Classroom Instructions*, IEEE Trans. Education,* 45(2), 19‑25.
8. Leon-Garcia, A. (2008). *Probability and Random Processes for Electrical Engineering*, New Jersey, 3th edition, Prentice Hall.
9. Jovanovic Dolecek, G. (1985). *Random Variables and Processes in Communications*, Sarajevo, Svjetlost.

Gordana Jovanovic Dolecek received a BS degree from the Department of Electrical Engineering, University of Sarajevo, an MSc degree from University of Belgrade, and a PhD degree from the Faculty of Electrical Engineering, University of Sarajevo. She was professor at the Faculty of Electrical Engineering, University of Sarajevo until 1993, and 1993-1995 she was with the Institute Mihailo Pupin, Belgrade. In 1995 she joined Institute INAOE, Department for Electronics, Puebla, Mexico, where she works as a professor and researcher. During 2001-2002 and 2006 she was with Department of Electrical & Computer Engineering, University of California, Santa Barbara, as visiting researcher. She is currently with San Diego State University as visiting researcher on a sabbatical leave. She is the author of three books, editor of one book, and author of more than 200 papers. Her research interests include digital signal processing and digital communications. She is a Senior member of IEEE, the member of Mexican Academy of Science, and the member of National Researcher System (SNI) Mexico.

fred harris is professor of Electrical and Computer Engineering at San Diego State University, where he holds the CUBIC Signal Processing Chair of the Communication Systems and Signal Processing Institute. He has extensive practical experience applying his skills to satellite and cable TV communication systems, wire-line and wireless modems, underwater acoustics, advanced radar and high performance laboratory instrumentation. He holds several patents on digital receiver and DSP technology, lectures on DSP worldwide, and consults for organizations requiring high performance DSP systems including the SPAWAR, Lockheed, Cubic, Hughes, Rockwell, Northrop Grumman, SAIC, GDE, and Motorola. He is has published over 160 papers and has contributed to a number of books on DSP. In 1990 and 1991 he was the Technical and then the General Chair of the Asilomar Conference on Signals, Systems, and Computers which meets annually in Pacific Grove, California. He is Editor of Signal Processing, Journal of Elsevier. He is the author of the Prentice-Hall textbook “Multirate Signal Processing for Communication Systems” and is a Life Fellow of the IEEE.