# Digital Signal Processing
# Statistics, Probability and Noise

Moslem Amiri, Václav Přenosil
Masaryk University

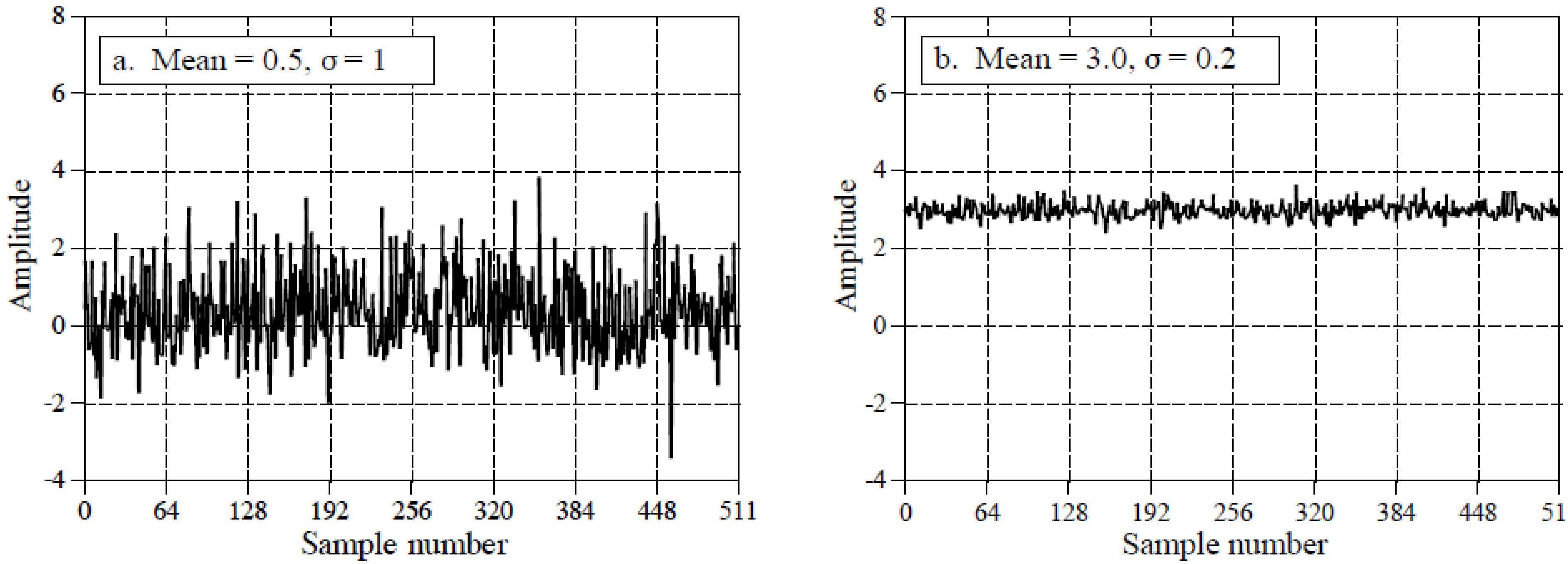
Resource: “The Scientist and Engineer's Guide to Digital Signal Processing”
(www.dspguide.com)
By Steven W. Smith

# Signal and Graph Terminology

- Signal
  - A description of how one parameter is related to another one
  - Continuous signal
    - Both parameters can assume a continuous range of values
  - Discrete signal or digitized signal
    - Signals formed from quantized parameters

# Signal and Graph Terminology

- Two discrete signals




- Domain
  - Type of parameter on horizontal axis

# Mean and Standard Deviation

- Mean (μ)
  - Average value of a signal (N samples)

$$\mu = \frac{1}{N}\sum_{i=0}^{N-1} x_i$$

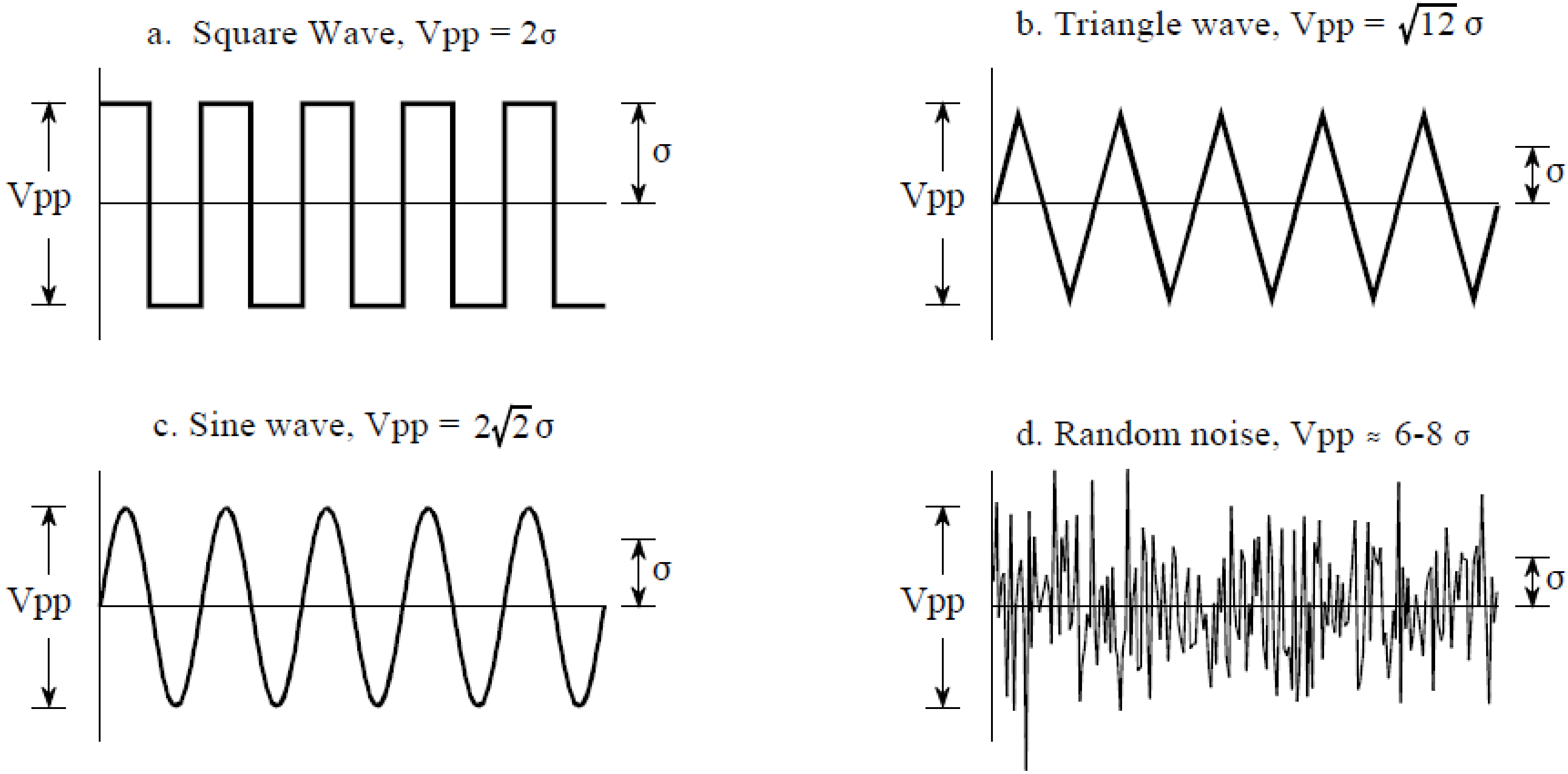
- Standard deviation (σ)
  - A measure of how far the signal fluctuates from mean
  - Only measures AC portion of a signal
  - Variance ($\sigma^2$): power of this fluctuation

$$\sigma^2 = \frac{1}{N-1}\sum_{i=0}^{N-1} (x_i - \mu)^2$$

- rms (root-mean-square)
  - Measures both AC and DC components of a signal
  - A signal with no DC → rms = σ

# Mean and Standard Deviation

- Relationship between σ and peak-to-peak value of several waveforms



# Mean and Standard Deviation

- Two limitations of calculation using

$$\mu = \frac{1}{N}\sum_{i=0}^{N-1} x_i \qquad \sigma^2 = \frac{1}{N-1}\sum_{i=0}^{N-1}(x_i - \mu)^2$$

  - If μ >> σ → subtracting two numbers that are very close in value → excessive round off error
  - Running statistics → requires all samples be involved in each new calculation
- Solution

$$\sigma^2 = \frac{1}{N-1}[\sum_{i=0}^{N-1} x_i{}^2 - \frac{1}{N}(\sum_{i=0}^{N-1} x_i)^2]$$

# Mean and Standard Deviation

- In some situations
  - Mean ($\mu$) = what is being measured
  - Standard deviation ($\sigma$) = noise and other interference
  - Signal-to-noise ratio (SNR) = $\mu/\sigma$
  - Coefficient of variation (CV) = $(\sigma/\mu)*100\%$

# Signal vs. Underlying Process

- Statistics
  - Interpreting numerical data such as acquired signals
  - Statistical noise
    - Statistics of acquired signal change each time experiment is repeated
- Probability
  - Used to understand processes that generate signals
  - Probabilities of underlying process are constant
- Typical error in calculating mean of an underlying process due to statistical noise

$$Error = \frac{\sigma}{N^{\frac{1}{2}}}$$

# Signal vs. Underlying Process

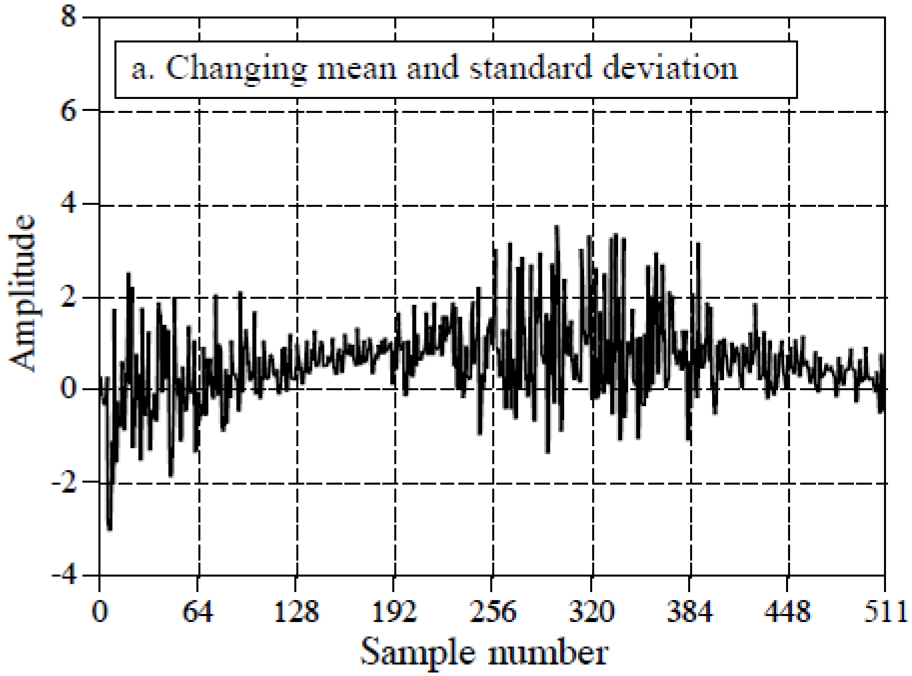
- Error in mean
  - Reduces value of σ
  - To compensate for this

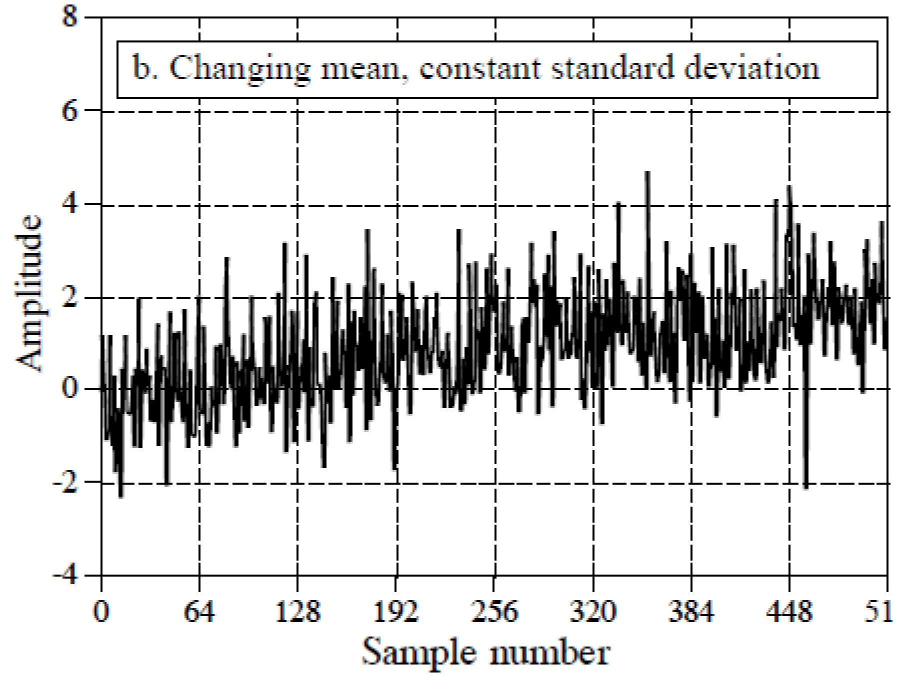
$$\sigma^2 = \frac{1}{N}\sum_{i=0}^{N-1}(x_i - \mu)^2 \longrightarrow \sigma^2 = \frac{1}{N-1}\sum_{i=0}^{N-1}(x_i - \mu)^2$$

  - Left equation → σ of acquired signal
  - Right equation → an estimate of σ of underlying process

# Signal vs. Underlying Process

- Nonstationary signals
  - Are not a result of statistical noise
  - Generated from underlying process changing

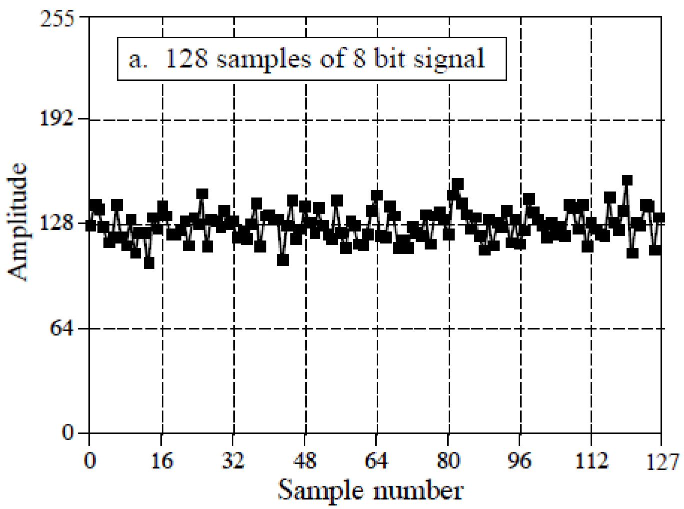


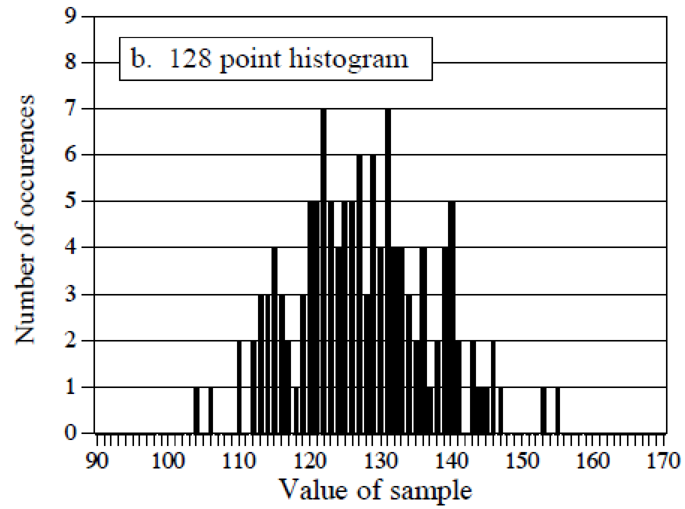




  - Problem: slowly changing μ interferes with calculating σ
  - Solution: breaking signal into short sections – calculating statistics for each section – averaging σs

# The Histogram, Pmf and Pdf

- Histogram
  - Displays number of samples having each possible value



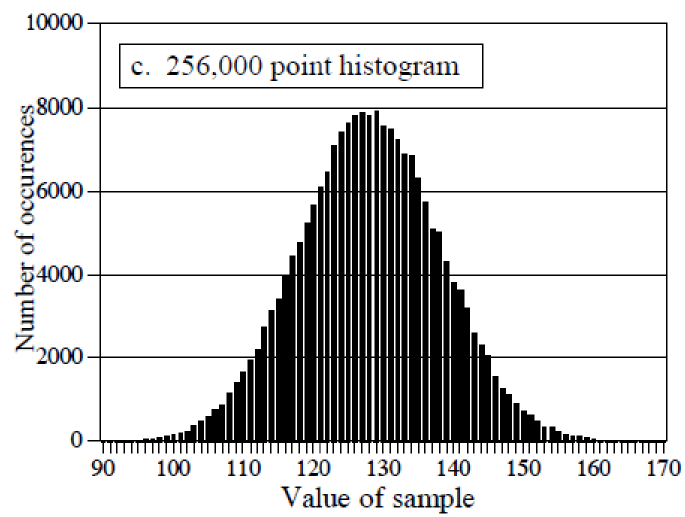




FIGURE 2-4
Examples of histograms. Figure (a) shows 128 samples from a very long signal, with each sample being an integer between 0 and 255. Figures (b) and (c) show histograms using 128 and 256,000 samples from the signal, respectively. As shown, the histogram is smoother when more samples are used.

# The Histogram, Pmf and Pdf

- Histogram
  - Represented by $H_i$, where i is index for value of sample
  - $H_i$ is number of samples that have a value of i
  - If M is number of points in histogram and N number of points in signal

$$N = \sum_{i=0}^{M-1} H_i$$

$$\mu = \frac{1}{N} \sum_{i=0}^{M-1} i H_i$$

$$\sigma^2 = \frac{1}{N-1} \sum_{i=0}^{M-1} (i-\mu)^2 H_i$$

  - Histogram is formed from an acquired signal → noisy

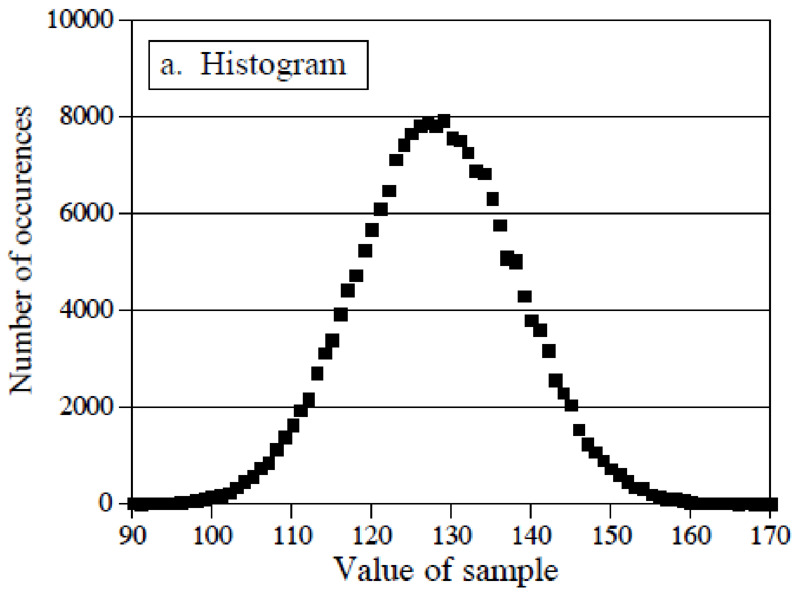
# The Histogram, Pmf and Pdf

- Probability mass function (pmf)
  - Curve for underlying process
  - What would be obtained with an infinite number of samples
  - can be estimated from histogram or by some mathematical technique
  - Vertical axis of pmf expressed on a fractional basis: each value in histogram divided by total number of samples
  - Sum of all values in pmf = 1

# The Histogram, Pmf and Pdf

- Probability density (or distribution) function (pdf)
  - Pdf is to continuous signals what pmf is to discrete ones
  - Indicates signal can take on a continuous range of values
  - Vertical axis of pdf is in units of probability density
  - Total area under pdf curve = $\int_{-\infty}^{+\infty} curve$ = 1

# The Histogram, Pmf and Pdf



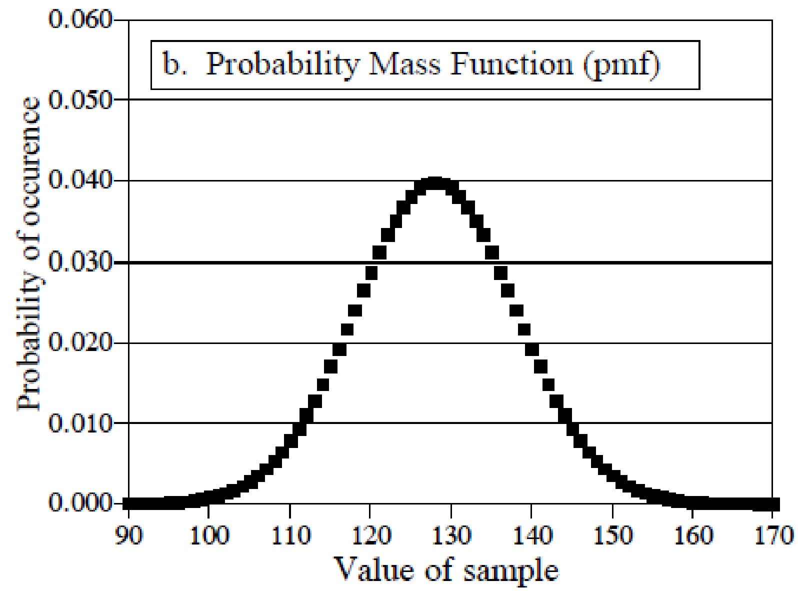
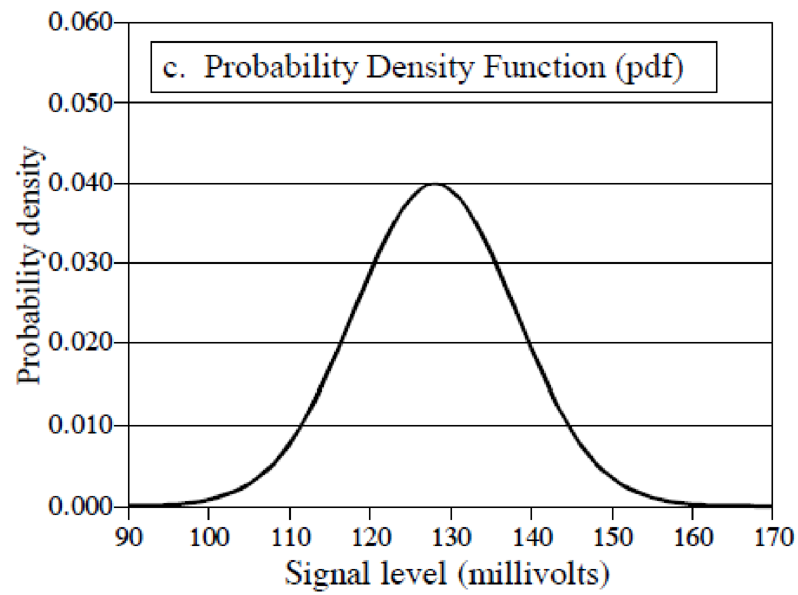




FIGURE 2-5
The relationship between (a) the histogram, (b) the probability mass function (pmf), and (c) the probability density function (pdf). The histogram is calculated from a finite number of samples. The pmf describes the probabilities of the underlying process. The pdf is similar to the pmf, but is used with continuous rather than discrete signals. Even though the vertical axis of (b) and (c) have the same values (0 to 0.06), this is only a coincidence of this example. The amplitude of these three curves is determined by: (a) the sum of the values in the histogram being equal to the number of samples in the signal; (b) the sum of the values in the pmf being equal to one, and (c) the area under the pdf curve being equal to one.



# The Histogram, Pmf and Pdf

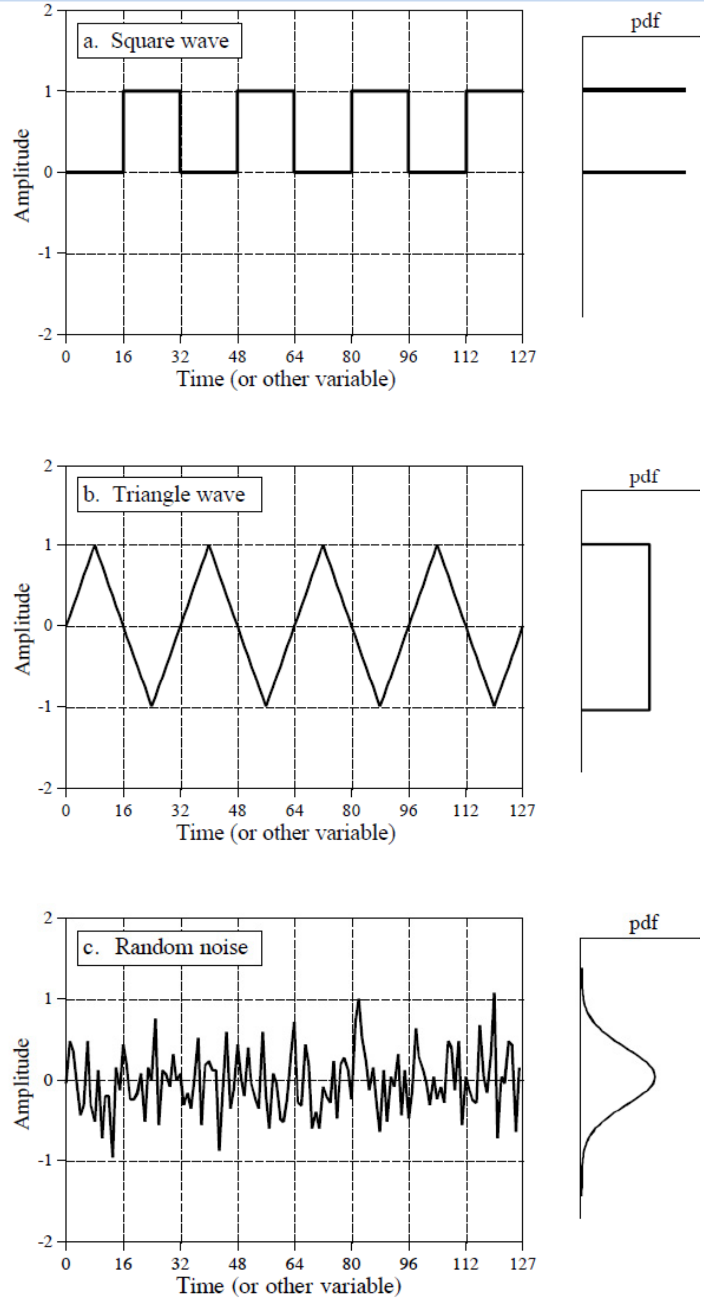
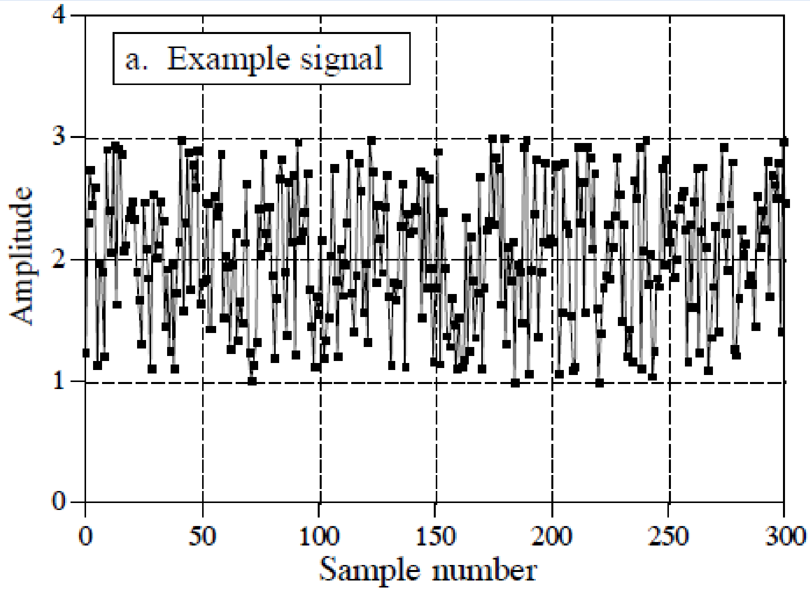



FIGURE 2-6
Three common waveforms and their probability density functions. As in these examples, the pdf graph is often rotated one-quarter turn and placed at the side of the signal it describes. The pdf of a square wave, shown in (a), consists of two infinitesimally narrow spikes, corresponding to the signal only having two possible values. The pdf of the triangle wave, (b), has a constant value over a range, and is often called a *uniform* distribution. The pdf of random noise, as in (c), is the most interesting of all, a bell shaped curve known as a *Gaussian*.

# The Histogram, Pmf and Pdf

- Problem in calculating histogram
  - Number of levels each sample can take on >> number of samples in signal
  - Always true for signals represented in floating point notation
  - Previously described approach involves counting number of samples that have each of possible quantization levels
    - Not possible with floating point data → billions of possible levels nearly all of which have no samples
  - Solution: binning
    - Done by arbitrarily selecting length of histogram to be some convenient number called bins
    - Value of each bin → total number of samples having a value within a certain range

# The Histogram, Pmf and Pdf



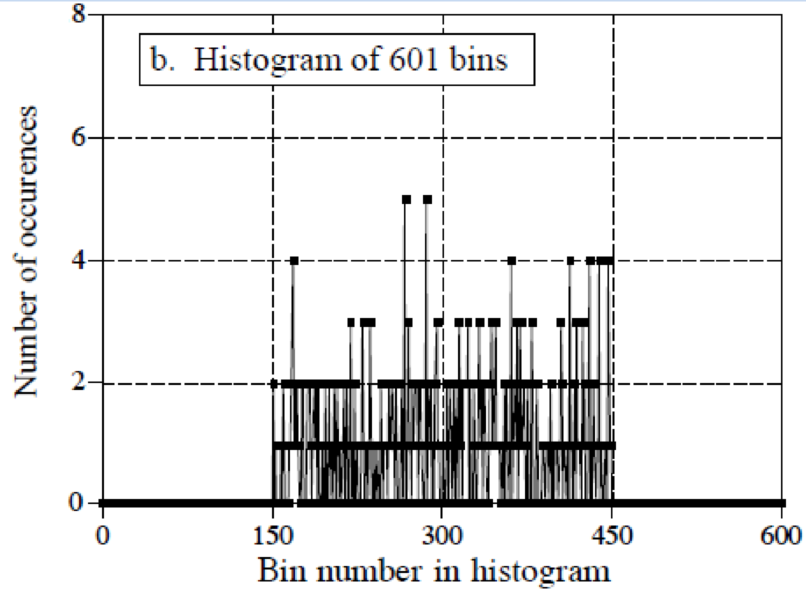
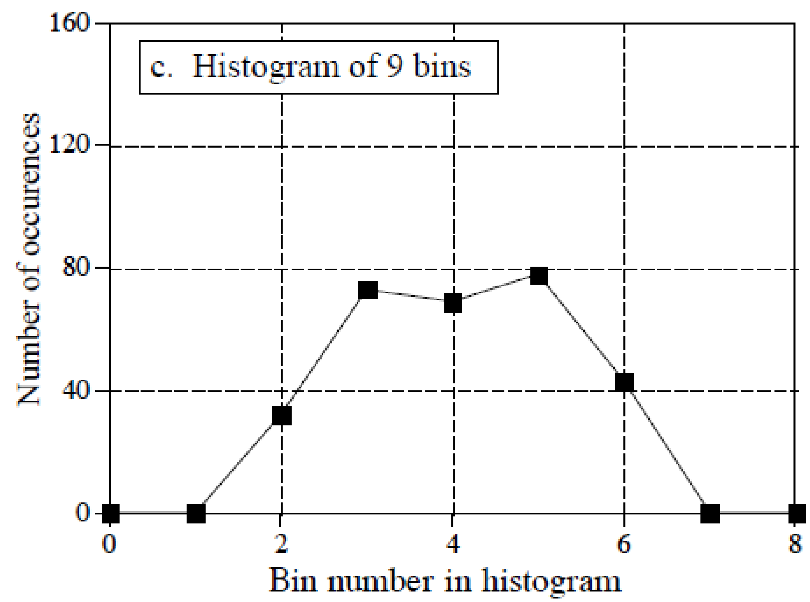




FIGURE 2-7
Example of binned histograms. As shown in (a), the signal used in this example is 300 samples long, with each sample a floating point number uniformly distributed between 1 and 3. Figures (b) and (c) show binned histograms of this signal, using 601 and 9 bins, respectively. As shown, a large number of bins results in poor resolution along the *vertical axis*, while a small number of bins provides poor resolution along the *horizontal axis*. Using more samples makes the resolution better in both directions.



# The Normal Distribution

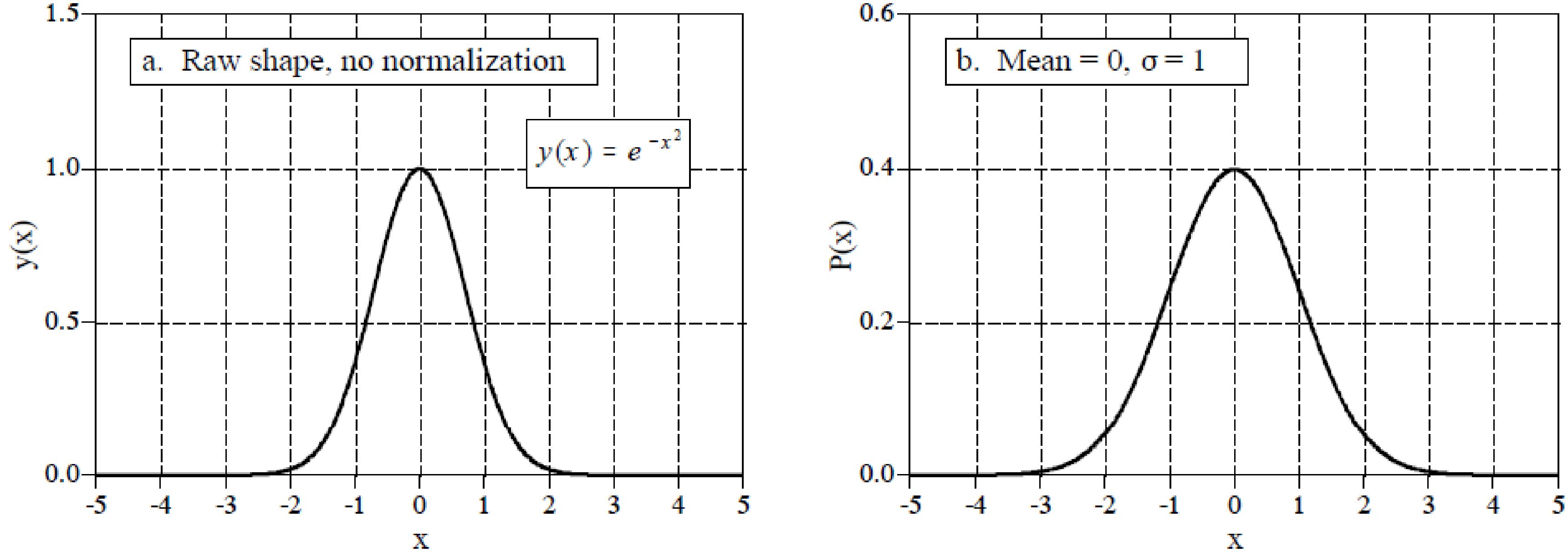
- Signals formed from random processes
  - Usually have a bell shaped pdf
  - Called normal distribution or Gauss distribution
  - Basic shape of curve generated from

    $$y(x) = e^{-x^2}$$

  - Adding adjustable μ and σ + normalizing (area under curve = 1) → general form of normal distribution

    $$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

# The Normal Distribution



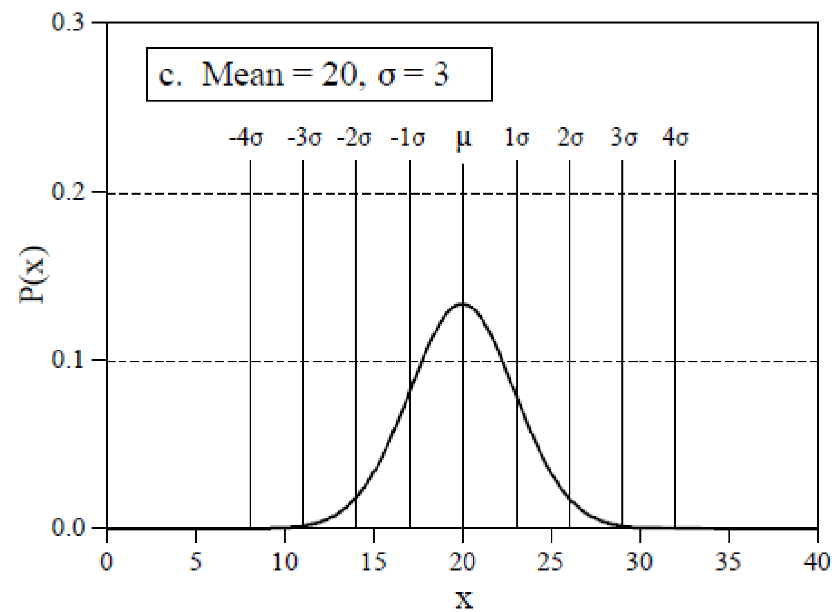




FIGURE 2-8
Examples of Gaussian curves. Figure (a) shows the shape of the raw curve without normalization or the addition of adjustable parameters. In (b) and (c), the complete Gaussian curve is shown for various means and standard deviations.

# The Normal Distribution

- Cumulative distribution function (cdf)
  - Integral of pdf
    - Used to find probability that a signal is within a certain range of values
  - Problem with Gaussian
    - Cannot be integrated using elementary methods
    - Solution: calculating by numerical integration – providing a table for use in calculating probabilities

# The Normal Distribution

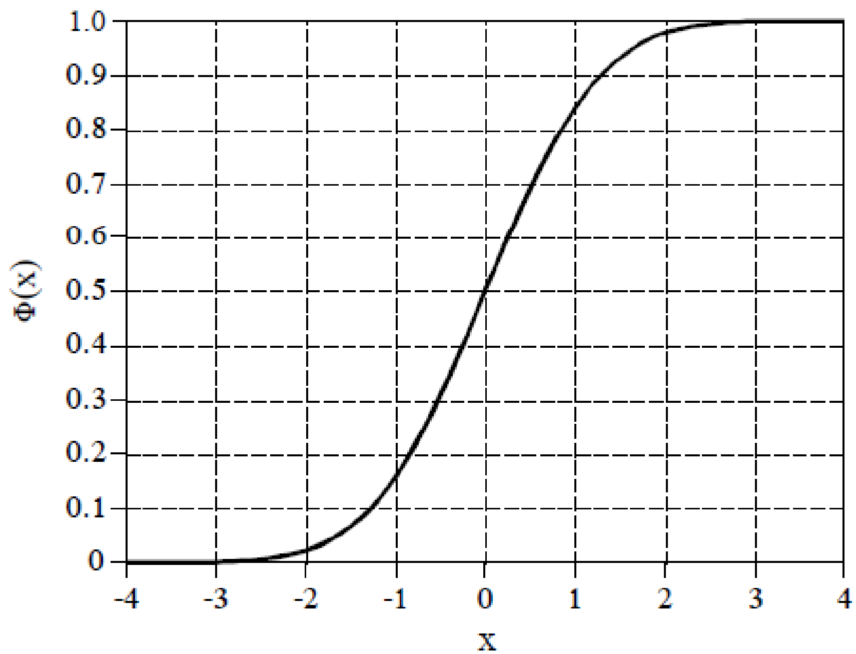



FIGURE 2-9 & TABLE 2-5
$\Phi(x)$, the cumulative distribution function of the normal distribution (mean = 0, standard deviation = 1). These values are calculated by numerically integrating the normal distribution shown in Fig. 2-8b. In words, $\Phi(x)$ is the probability that the value of a normally distributed signal, at some randomly chosen time, will be less than $x$. In this table, the value of $x$ is expressed in units of standard deviations referenced to the mean.

| x | Φ(x) | x | Φ(x) |
|---|---|---|---|
| -3.4 | .0003 | 0.0 | .5000 |
| -3.3 | .0005 | 0.1 | .5398 |
| -3.2 | .0007 | 0.2 | .5793 |
| -3.1 | .0010 | 0.3 | .6179 |
| -3.0 | .0013 | 0.4 | .6554 |
| -2.9 | .0019 | 0.5 | .6915 |
| -2.8 | .0026 | 0.6 | .7257 |
| -2.7 | .0035 | 0.7 | .7580 |
| -2.6 | .0047 | 0.8 | .7881 |
| -2.5 | .0062 | 0.9 | .8159 |
| -2.4 | .0082 | 1.0 | .8413 |
| -2.3 | .0107 | 1.1 | .8643 |
| -2.2 | .0139 | 1.2 | .8849 |
| -2.1 | .0179 | 1.3 | .9032 |
| -2.0 | .0228 | 1.4 | .9192 |
| -1.9 | .0287 | 1.5 | .9332 |
| -1.8 | .0359 | 1.6 | .9452 |
| -1.7 | .0446 | 1.7 | .9554 |
| -1.6 | .0548 | 1.8 | .9641 |
| -1.5 | .0668 | 1.9 | .9713 |
| -1.4 | .0808 | 2.0 | .9772 |
| -1.3 | .0968 | 2.1 | .9821 |
| -1.2 | .1151 | 2.2 | .9861 |
| -1.1 | .1357 | 2.3 | .9893 |
| -1.0 | .1587 | 2.4 | .9918 |
| -0.9 | .1841 | 2.5 | .9938 |
| -0.8 | .2119 | 2.6 | .9953 |
| -0.7 | .2420 | 2.7 | .9965 |
| -0.6 | .2743 | 2.8 | .9974 |
| -0.5 | .3085 | 2.9 | .9981 |
| -0.4 | .3446 | 3.0 | .9987 |
| -0.3 | .3821 | 3.1 | .9990 |
| -0.2 | .4207 | 3.2 | .9993 |
| -0.1 | .4602 | 3.3 | .9995 |
| 0.0 | .5000 | 3.4 | .9997 |

# Digital Noise Generation

- Generate signals that resemble various types of random noise
  - To test performance of algorithms that must work in presence of noise
- Random number generator
  - Heart of digital noise generation

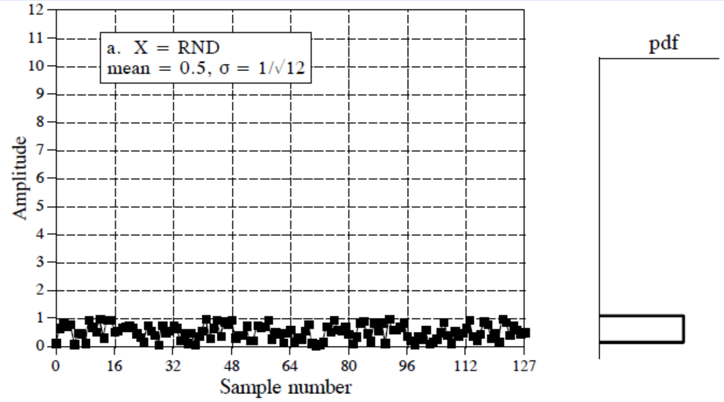
# Digital Noise Generation

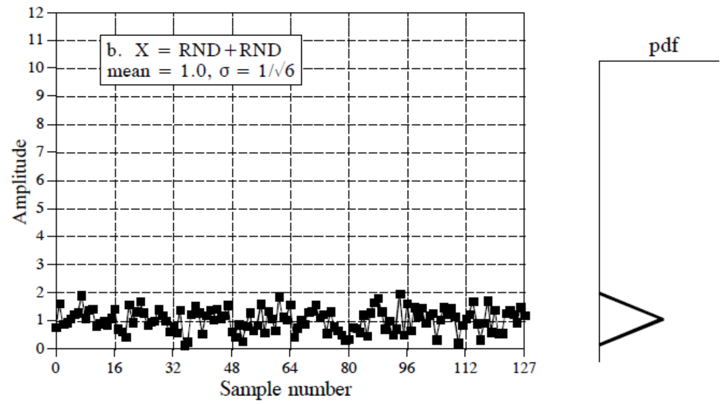
- First method
  - Central limit theorem
    - A sum of random numbers becomes normally distributed as more and more of random numbers are added together
- In Programming languages
  - X = RND → 0 < X < 1, uniform distribution
    - $\mu = 0.5, \sigma = 1/\sqrt{12} = 0.29$
  - X = RND + RND → 0 < X < 2, triangular distribution
    - $\mu = 1, \sigma = 1/\sqrt{6}$
  - X = RND + … + RND (12 times), 0 < X < 12, Gaussian distribution
    - $\mu = 6, \sigma = 1$
    - Can be used to create a normally distributed noise signal

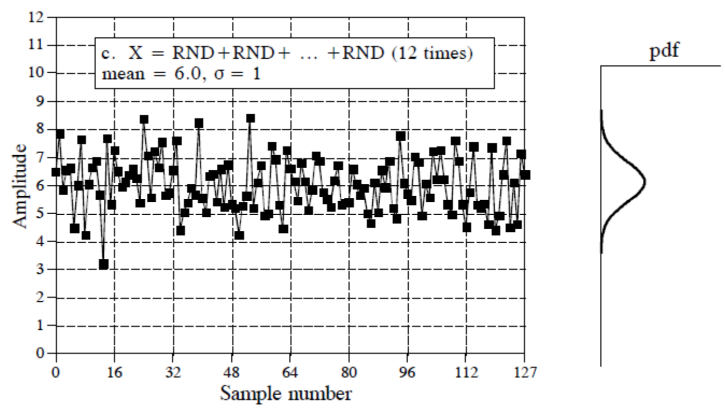
# Digital Noise Generation

- For each sample in signal
  - Add twelve random numbers
  - Subtract six to make mean equal to zero
  - Multiply by standard deviation desired
  - Add desired mean

# Digital Noise Generation



# Digital Noise Generation

- Second method
  - Random number generator invoked twice to obtain $R_1$ and $R_2$
  - A normally distributed random number, X, is found

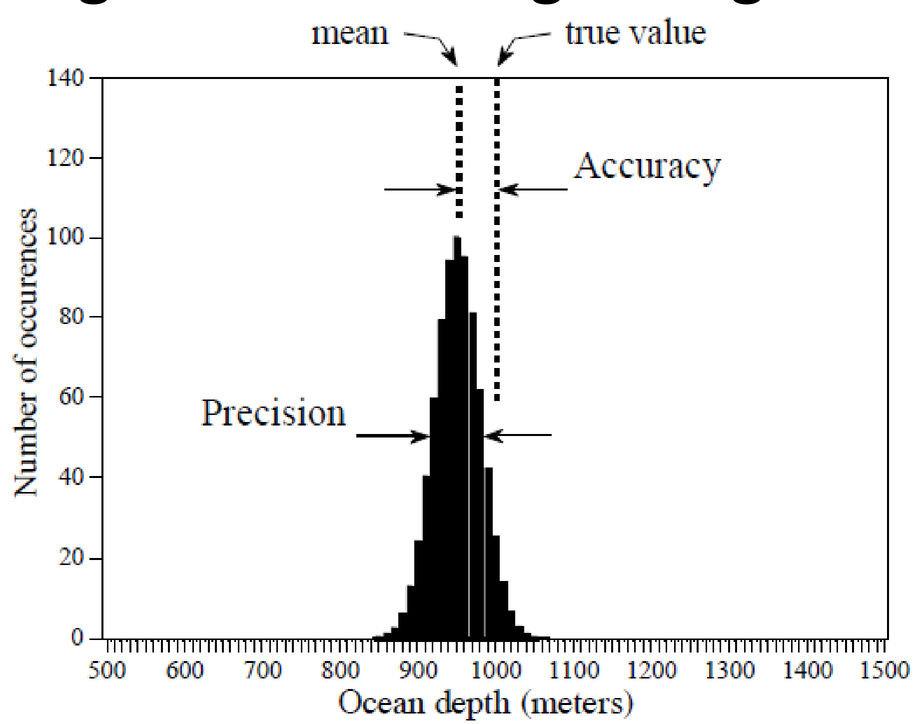
$$X = (-2\log_e R_1)^{1/2} \cos(2\pi R_2)$$

    - $\mu = 0, \sigma = 1$
  - To generate normally distributed random signals
    - Take each number generated by equation above
    - Multiply it by desired standard deviation
    - Add desired mean

# Precision and Accuracy

- Example
  - An oceanographer measuring water depth using a sonar system
  - Location is exactly 1000 meters deep (true value)
  - Oceanographer takes many successive readings
  - Measurements are arranged as following histogram

FIGURE 2-11
Definitions of accuracy and precision. Accuracy is the difference between the true value and the mean of the under-lying process that generates the data. Precision is the spread of the values, specified by the standard deviation, the signal-to-noise ratio, or the CV.



# Precision and Accuracy

- Accuracy
  - Amount of shift of mean from true value
- Precision
  - Width of distribution, expressed by standard deviation, signal-to-noise ratio, or CV
- Poor repeatability
  - A measurement that has good accuracy, but poor precision
  - Poor precision results from random errors
    - Random errors change each time measurement is repeated
    - Precision is a measure of random noise
    - Averaging several measurements always improves precision

# Precision and Accuracy

- precise measurement but with poor accuracy
  - Poor accuracy results from systematic errors
    - Systematic errors become repeated in exactly same manner each time measurement is conducted
    - Accuracy dependant on how system is calibrated
    - Averaging individual measurements does not improve accuracy