Statistics, Probability and Noise

Dr. Clark Hochgraf

EEET-425

Today's Agenda

- Key Points From Chapter
- How to Read the Code Examples (BASIC LANGUAGE)
- In Class Problem(ICP): Speed Skills In Excel
- Chapter Summary

Chapter 2 Key Points

- Noise comes from many different sources
  - noise from the analog world
  - noise from the A/D conversion process
  - noise from digital calculations
- Noise exists and needs to be quantitatively described.
  - Signal to Noise Ratio (dB) is used for communications
  - Coefficient of Variation (%) is used for video
- Noise is random, or at least assumed to be random

Chapter 2 Key Points

- Noise is random, or at least assumed to be random
  - random signals add differently than 'normal' signals
    - Example of a normal signal is a sine wave of voltage, or a dc value of voltage.
    - Example of a random signal is thermal noise in a resistor, or static on a radio.
  - random signals add in quadrature
    - Think Pythagorean theorem
    - a^2+b^2 = c^2
    - With random signals, 2+2 does not equal 4 (at least when talking about standard deviation)

Chapter 2 Key Points

- Statistics are used to measure and quantify the amount of noise and to understand the properties of how noise adds.
- If noise is excessive, you can design a filter to reduce the effect of the noise.
  - Filters are the subject of later chapters

Chapter 2 Key Points

- Common Statistics Used to Describe Noise and Signals
  - Mean, Standard Deviation, Variance
  - Each statistic has its own purpose and its own properties
- A set of measurements of signal can be displayed using histograms to help visualize how much random noise is present.

Mean and Standard Deviation

EQUATION 2-1
Calculation of a signal's mean. The signal is contained in x_0 through x_{N-1}, i is an index that runs through these values, and μ is the mean.

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

EQUATION 2-2
Calculation of the standard deviation of a signal. The signal is stored in x_i, μ is the mean found from Eq. 2-1, N is the number of samples, and σ is the standard deviation.

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

- σ is the standard deviation – it indicates how far a signal varies from the mean μ.
- σ^2 is the variance. The variance indicates the power of the fluctuation from the mean (think rms)

Example: Mean vs Standard Deviation

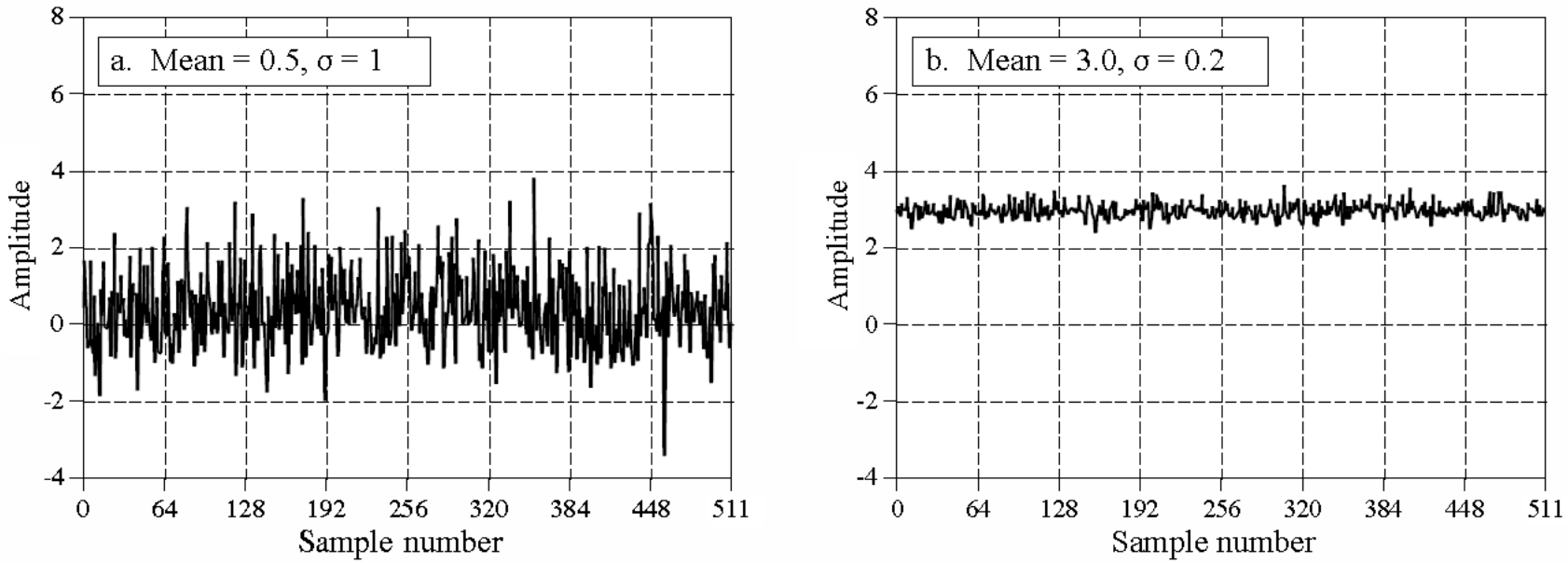


FIGURE 2-1
Examples of two digitized signals with different means and standard deviations.

- Question: How many samples in each figure?

Standard Form Calculation of Mean and Standard Deviation

```
100 CALCULATION OF THE MEAN AND STANDARD DEVIATION
110 '
120 DIM X[511]                      'The signal is held in X[0] to X[511]
130 N% = 512                        'N% is the number of points in the signal
140 '
150 GOSUB XXXX                      'Mythical subroutine that loads the signal into X[ ]
160 '
170 MEAN = 0                        'Find the mean via Eq. 2-1
180 FOR I% = 0 TO N%-1
190   MEAN = MEAN + X[I%]
200 NEXT I%
210 MEAN = MEAN/N%
220 '
230 VARIANCE = 0                    'Find the standard deviation via Eq. 2-2
240 FOR I% = 0 TO N%-1
250   VARIANCE = VARIANCE + ( X[I%] - MEAN )^2
260 NEXT I%
270 VARIANCE = VARIANCE/(N%-1)
280 SD = SQR(VARIANCE)
290 '
300 PRINT  MEAN  SD                 'Print the calculated mean and standard deviation
310 '
320 END
```

TABLE 2-1

Precision and Accuracy

- Accuracy
  - How close is the estimated mean μ to the true mean?
- Precision
  - How well do the individual measurements or sample compare with each other? It is expressed by the Signal to Noise Ratio (SNR) or by the Coefficient of Variation (CV)

Precision and Accuracy

- Precision is an measure of random noise.
- Accuracy is an indicator of systematic errors, e.g. errors in calibration.

In Class Problem (ICP):

- Building Skills In Excel

In Class Problem (ICP)

- Noise adding in quadrature

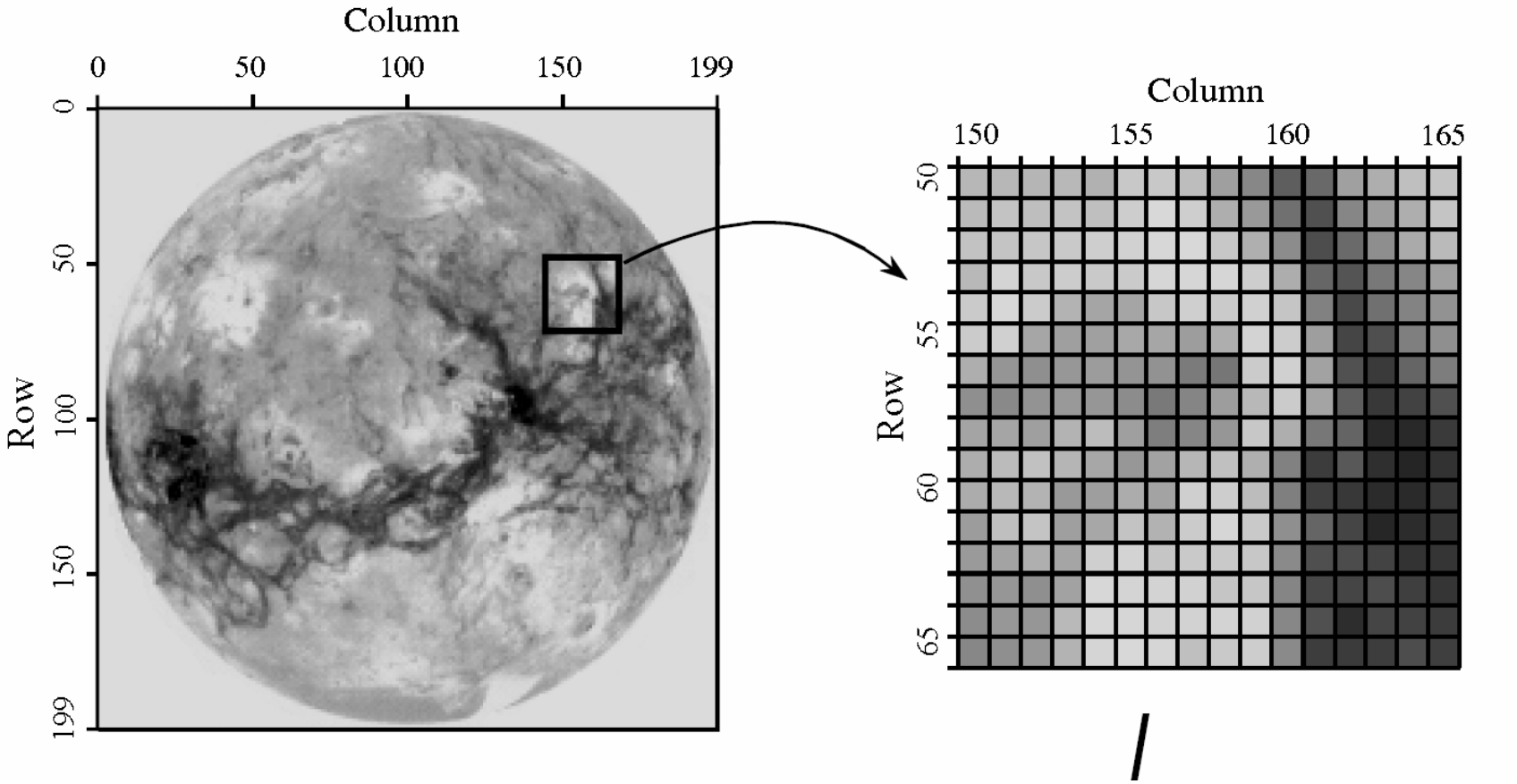
Chapter Summary

- Signals terminology – quantifying noise
- Mean, variance, standard deviation
- Noise adding in quadrature
- Histogram, PMF, PDF
- Digital Noise Generation
- Precision and Accuracy

Signal Terminology

- Horizontal axis
  - The independent variable, domain
  - Often we'll use "Sample Number". Later, for frequency domain, we'll use ""Frequency" which generally refers to a fraction of sampling rate (runs from 0 to 0.5).
  - Domains can be time, frequency, spatial or other
    - Spatial domain is used for images.

Spatial Domain



Signal Terminology

- Vertical axis
  - The dependent variable, range
  - Amplitude or other characteristic

Examples of Common Signals and Their Standard Deviations

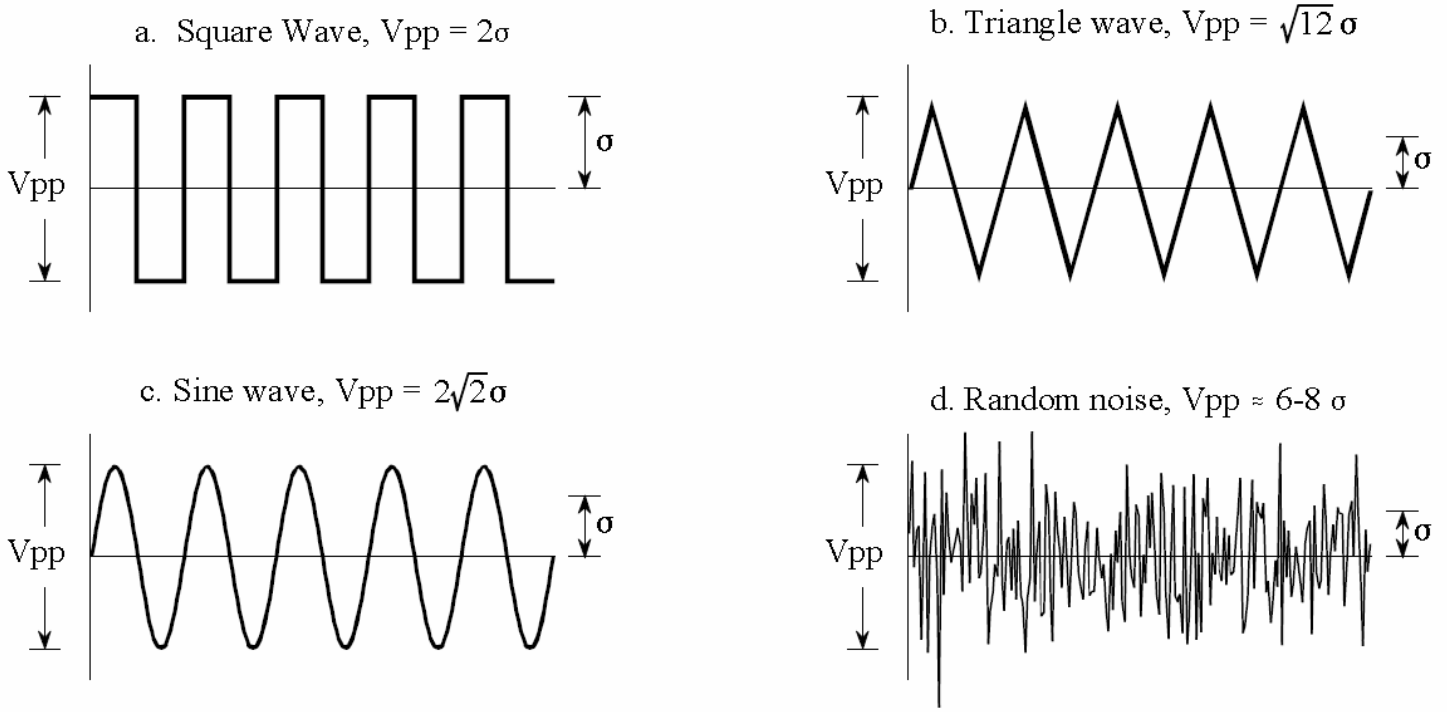
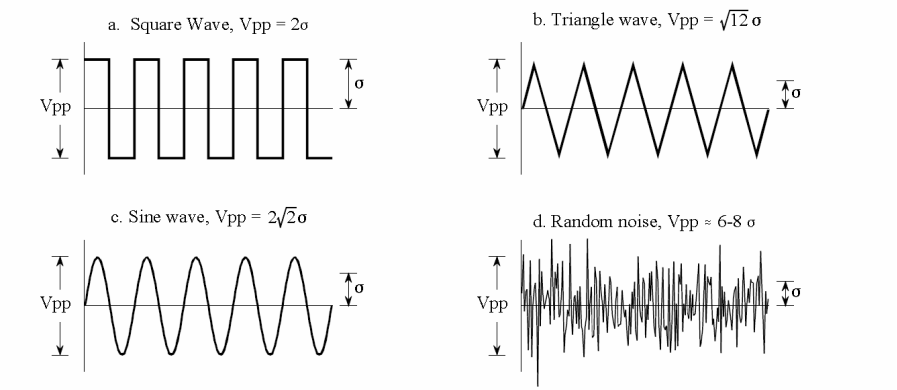


FIGURE 2-2
Ratio of the peak-to-peak amplitude to the standard deviation for several common waveforms. For the square wave, this ratio is 2; for the triangle wave it is $\sqrt{12} = 3.46$; for the sine wave it is $2\sqrt{2} = 2.83$. While random noise has no *exact* peak-to-peak value, it is *approximately* 6 to 8 times the standard deviation.

In Class Problem:

- What is the power in the fluctuation for each of these signals?
- A)Square wave with Vpp= 2
- B)Sine wave with Vpp= 2.828
- C)Triangular wave with Vpp= 3.464
- D) Random noise with Vpp = 7



Implementation Issues with Mean and Standard Deviation Calculations

- A limitation of the mean and standard deviation calculations are that they can create excessive round off error by subtracting two numbers that are very close in value e.g. when mean is much greater than standard deviation. (e.i. Xi – μ is small.)

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

- A second issue is that to get the mean or std. deviation you must recalculate using all the past values of xi. This makes the calculation slow and requires a lot of memory to store all the past values of xi

In Class Problem: Round off error

- Assume that you have collected 100,000,000 samples of a voltage signal and stored the values as single precision floating point numbers. If each calculation of Xi-μ has a biased additive round off error of 1 part in 40 million, what is the largest possible error in the computation of the variance?

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

Alternative Calculation Using Running Statistics

EQUATION 2-3
Calculation of the standard deviation using running statistics. This equation provides the same result as Eq. 2-2, but with less round-off noise and greater computational efficiency. The signal is expressed in terms of three accumulated parameters: N, the total number of samples; *sum*, the sum of these samples; and *sum of squares*, the sum of the squares of the samples. The mean and standard deviation are then calculated from these three accumulated parameters.

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

or using a simpler notation,

$$\sigma^2 = \frac{1}{N-1}\left[sum\ of\ squares - \frac{sum^2}{N}\right]$$

Code for Using Running Statistics

```
100 'MEAN AND STANDARD DEVIATION USING RUNNING STATISTICS
110 '
120 DIM X[511]                      'The signal is held in X[0] to X[511]
130 '
140 GOSUB XXXX                      'Mythical subroutine that loads the signal into X[ ]
150 '
160 N% = 0                          'Zero the three running parameters
170 SUM = 0
180 SUMSQUARES = 0
190 '
200 FOR I% = 0 TO 511               'Loop through each sample in the signal
210   '
220   N% = N%+1                     'Update the three parameters
230   SUM = SUM + X[I%]
240   SUMSQUARES = SUMSQUARES + X[I%]^2
250   '
260   MEAN = SUM/N%                 'Calculate mean and standard deviation via Eq. 2-3
270   IF N% = 1 THEN SD = 0: GOTO 300
280   SD = SQR( (SUMSQUARES - SUM^2/N%) / (N%-1) )
290   '
300   PRINT  MEAN   SD              'Print the running mean and standard deviation
310   '
320 NEXT I%
330 '
340 END
```

TABLE 2-2

In Class Problem: Running Statistics

- As one new value of the input signal xi is acquired, how many calculations (multiply, add, divide, square root) are required to compute the new standard deviation using
  - 1) the basic calculation
  - 2) the running statistics calculation

- What impact does this have on calculation time?

Signal to Noise Ratio

- Now that we've defined how to quantify a signal in terms of mean and variance, we can apply this to describing how clean/noisy a signal is.
- In some situations, the mean describes what is being measured, while the standard deviation represents the noise.
- To describe how noisy a measurement we can compute the **Signal to Noise Ratio** (SNR) which is the mean divided by the standard deviation.
- Another term used is the **Coefficient of Variation** (CV) which is the standard deviation divided by the mean expressed as a percentage.
  - E.g. if SNR =20 db (10x), the Coefficient of Variation is 10%
  - Higher SNR = lower Coefficient of Variation

Signals vs Underlying Processes

- Statistics describe a data set, for example a set of samples of a signal. Probability describes the underlying process that created that data set.
  - Coin flip example:
    - Flipping a coin 1000 times (1= heads, 0 =tails) generates a data set whose mean might be 0.498. However, the process that generated that signal has a mean of exactly 0.5, determined by the probability of heads (50%) or tails (50%). They are close but subtly different.

Statistical Noise, Statistical Variation

- Statistical variation arises in that the probability is constant (0.5 for the coin toss case), but the statistics (mean) for any given set of samples can vary.
- How big is this statistical variation? The size of the error is dependent upon the standard deviation and the number of samples.

EQUATION 2-4
Typical error in calculating the mean of an underlying process by using a finite number of samples, N. The parameter, σ, is the standard deviation.

$$Typical\ error = \frac{\sigma}{N^{1/2}}$$

Typical Error

- To reduce the typical error, you can increase the number of samples, N.

$$Typical\ error = \frac{\sigma}{N^{1/2}}$$

- What does this formula really mean? It states that the error in the mean between the samples you took and the true mean will get smaller as you take more samples. As N goes to infinity, the error between the estimated mean and the true mean goes to zero.
- How is this formula used? You can determine how many samples are needed to get a certain level of error.

Non-stationary Processes

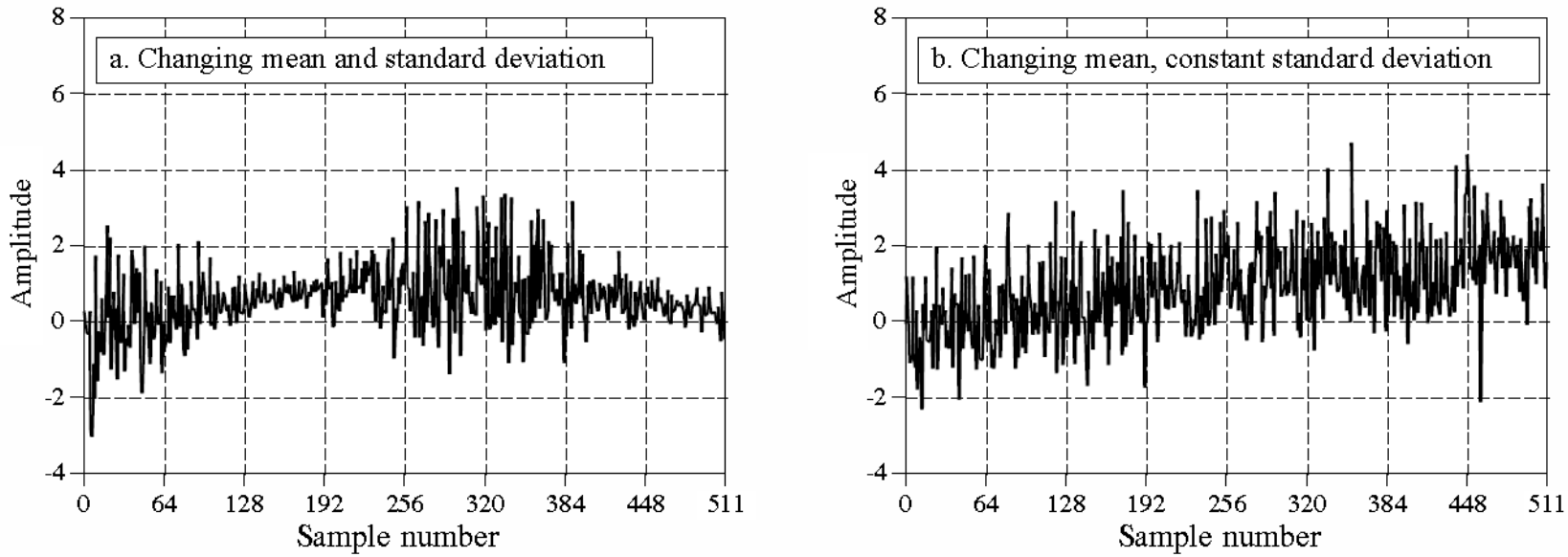
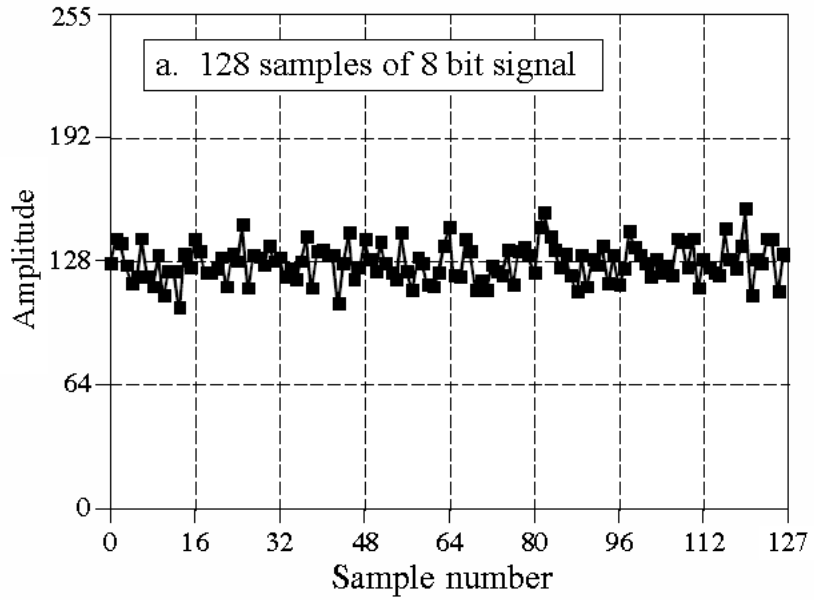


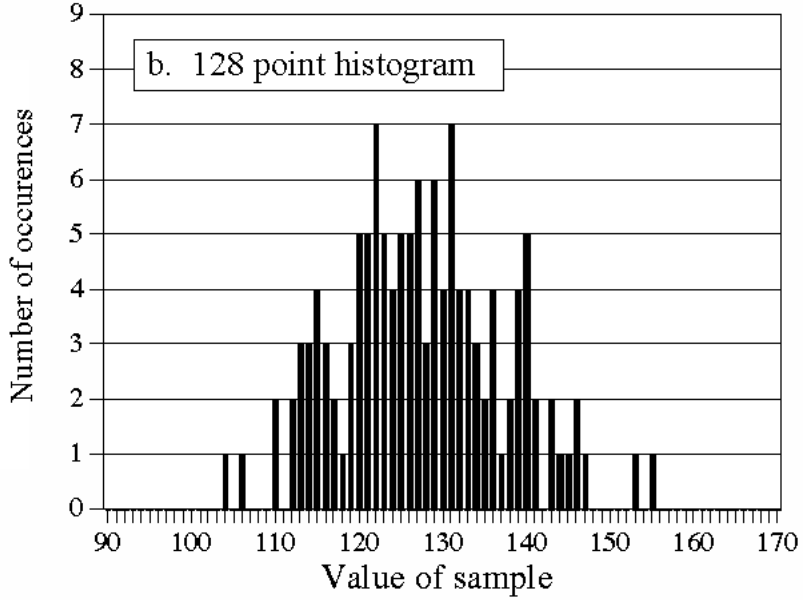
FIGURE 2-3
Examples of signals generated from nonstationary processes. In (a), both the mean and standard deviation change. In (b), the standard deviation remains a constant value of one, while the mean changes from a value of zero to two. It is a common analysis technique to break these signals into short segments, and calculate the statistics of each segment individually.

- What is the impact of using short segments (fewer samples) on the typical error of the mean and standard deviation?

Histograms

- Describes the number of samples in the data set that have the given value. E.g. 8 bit A/D samples

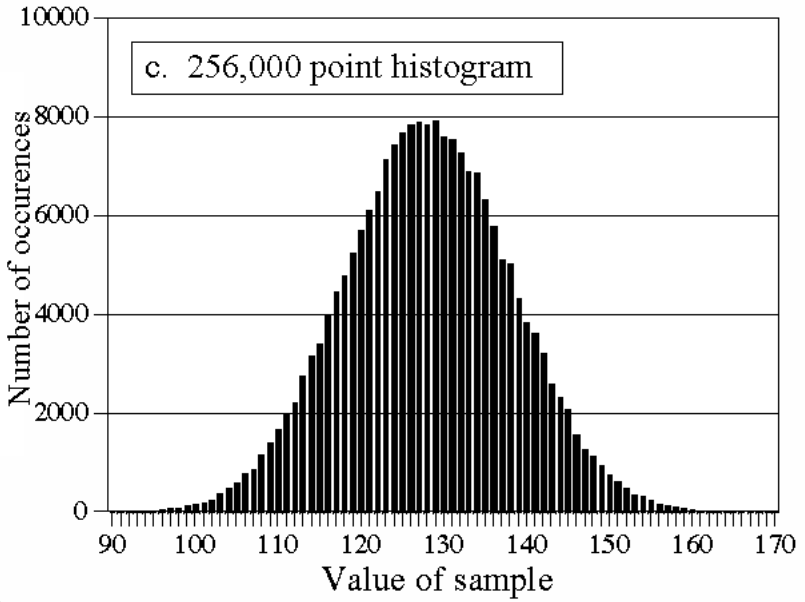




Histogram With More Data Samples

- More samples reveals the underlying distribution, smoother.

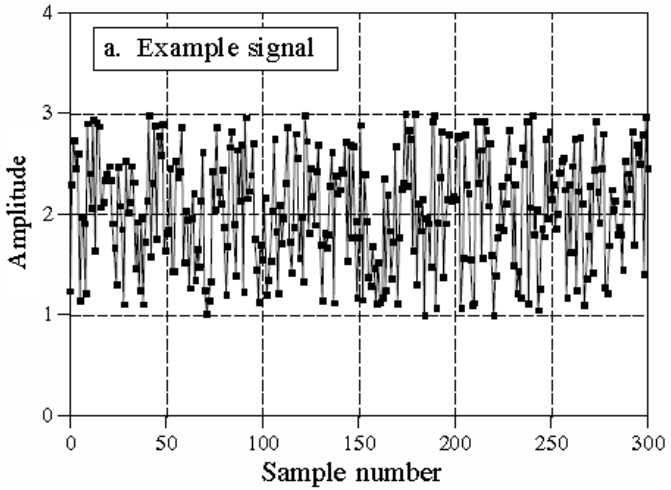
FIGURE 2-4
Examples of histograms. Figure (a) shows 128 samples from a very long signal, with each sample being an integer between 0 and 255. Figures (b) and (c) show histograms using 128 and 256,000 samples from the signal, respectively. As shown, the histogram is smoother when more samples are used.



EQUATION 2-5
The sum of all of the values in the histogram is equal to the number of points in the signal. In this equation, H_i is the histogram, N is the number of points in the signal, and M is the number of points in the histogram.

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

Selection of the Number of Bins



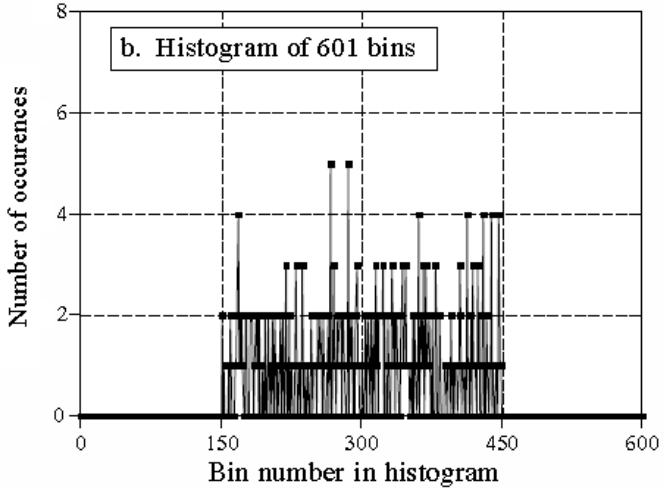
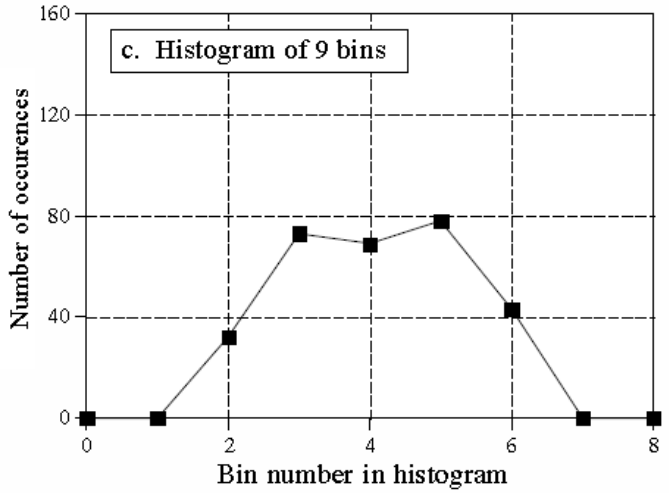


FIGURE 2-7
Example of binned histograms. As shown in (a), the signal used in this example is 300 samples long, with each sample a floating point number uniformly distributed between 1 and 3. Figures (b) and (c) show binned histograms of this signal, using 601 and 9 bins, respectively. As shown, a large 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.



Calculation of Mean, Standard Deviation From Histogram

EQUATION 2-6
Calculation of the mean from the histogram. This can be viewed as combining all samples having the same value into groups, and then using Eq. 2-1 on each group.

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

EQUATION 2-7
Calculation of the standard deviation from the histogram. This is the same concept as Eq. 2-2, except that all samples having the same value are operated on at once.

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

- Hi is the number of samples in bin number i.

Calculation of Mean, Standard Deviation From Histogram

```
100 'CALCULATION OF THE HISTOGRAM, MEAN, AND STANDARD DEVIATION
110 '
120 DIM X%[25000]                 'X%[0] to X%[25000] holds the signal being processed
130 DIM H%[255]                   'H%[0] to H%[255] holds the histogram
140 N% = 25001                    'Set the number of points in the signal
150 '
160 FOR I% = 0 TO 255             'Zero the histogram, so it can be used as an accumulator
170   H%[I%] = 0
180 NEXT I%
190 '
200 GOSUB XXXX                    'Mythical subroutine that loads the signal into X%[ ]
210 '
220 FOR I% = 0 TO 25000 'Calculate the histogram for 25001 points
230   H%[ X%[I%] ] = H%[ X%[I%] ] + 1
240 NEXT I%
250 '
260 MEAN = 0                      'Calculate the mean via Eq. 2-6
270 FOR I% = 0 TO 255
280   MEAN = MEAN + I% * H%[I%]
290 NEXT I%
300 MEAN = MEAN / N%
310 '
320 VARIANCE = 0                  'Calculate the standard deviation via Eq. 2-7
330 FOR I% = 0 TO 255
340   VARIANCE = VARIANCE + H%[I%] * (I%-MEAN)^2
350 NEXT I%
360 VARIANCE = VARIANCE / (N%-1)
370 SD = SQR(VARIANCE)
380 '
390 PRINT  MEAN  SD               'Print the calculated mean and standard deviation.
400 '
410 END
```

TABLE 2-3

- 25000 samples of an 8 bit number
- 256 bins (8 bits-> 0 255)
- Add to bin count if number is in bin
- For each bin, add number of hits in bin times bin value to the mean
- Variance is number of hits in each bin times bin value minus mean

Probability Mass Function Vs Histogram

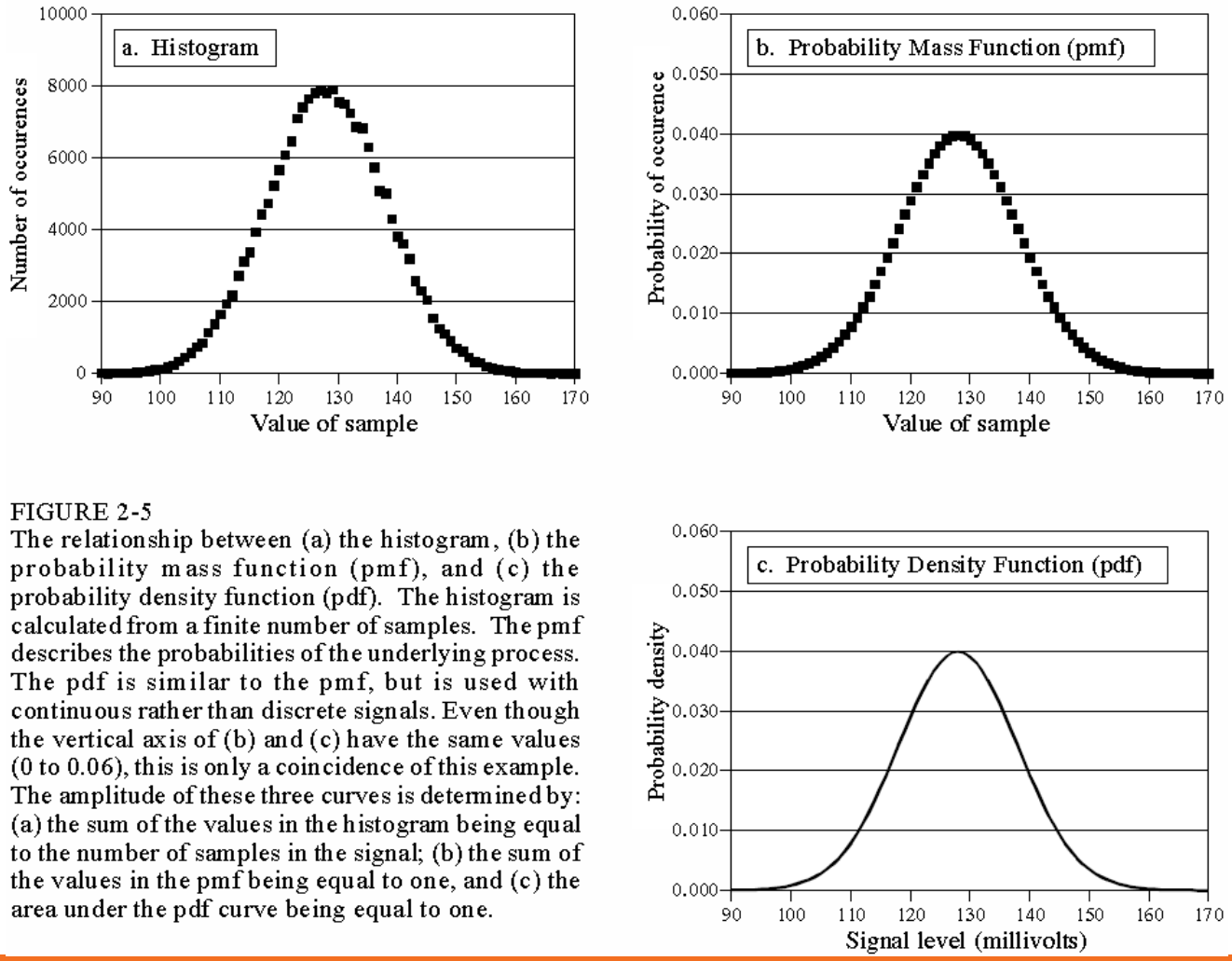
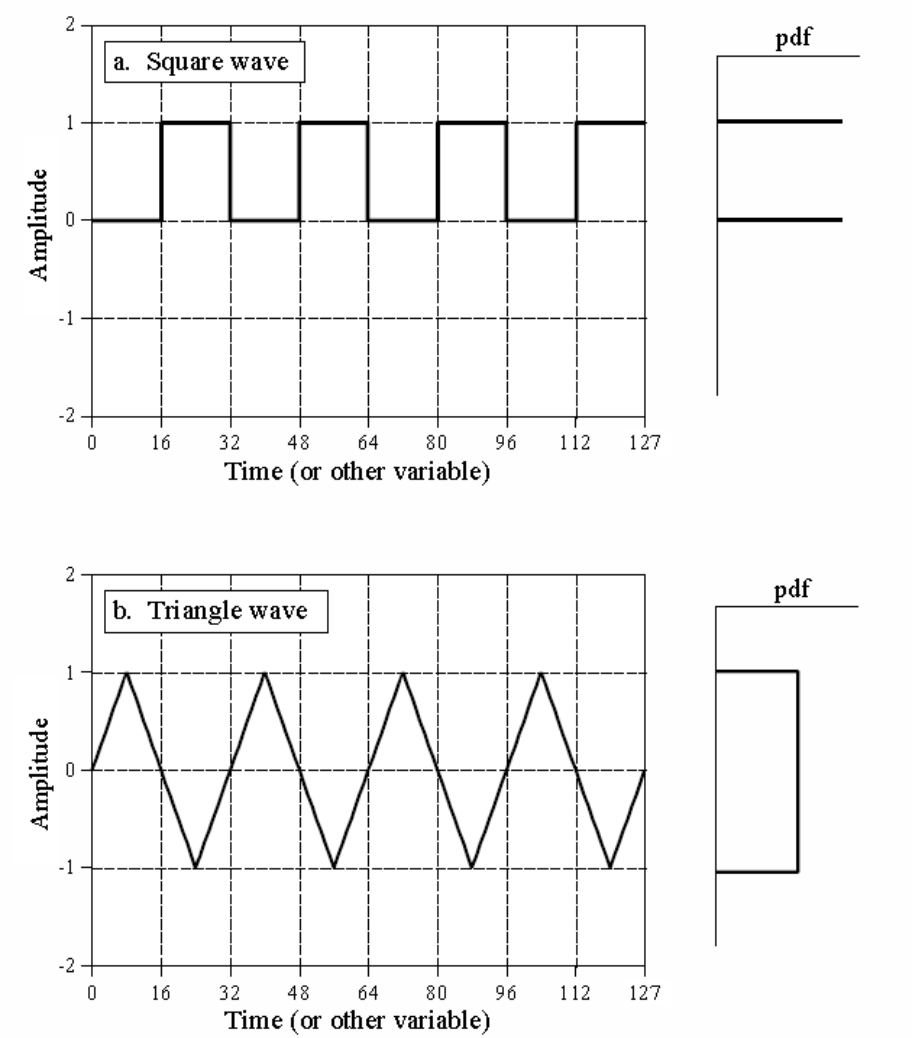


FIGURE 2-5
The relationship between (a) the histogram, (b) the probability mass function (pmf), and (c) the probability density function (pdf). The histogram is calculated from a finite number of samples. The pmf describes the probabilities of the underlying process. The pdf is similar to the pmf, but 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.

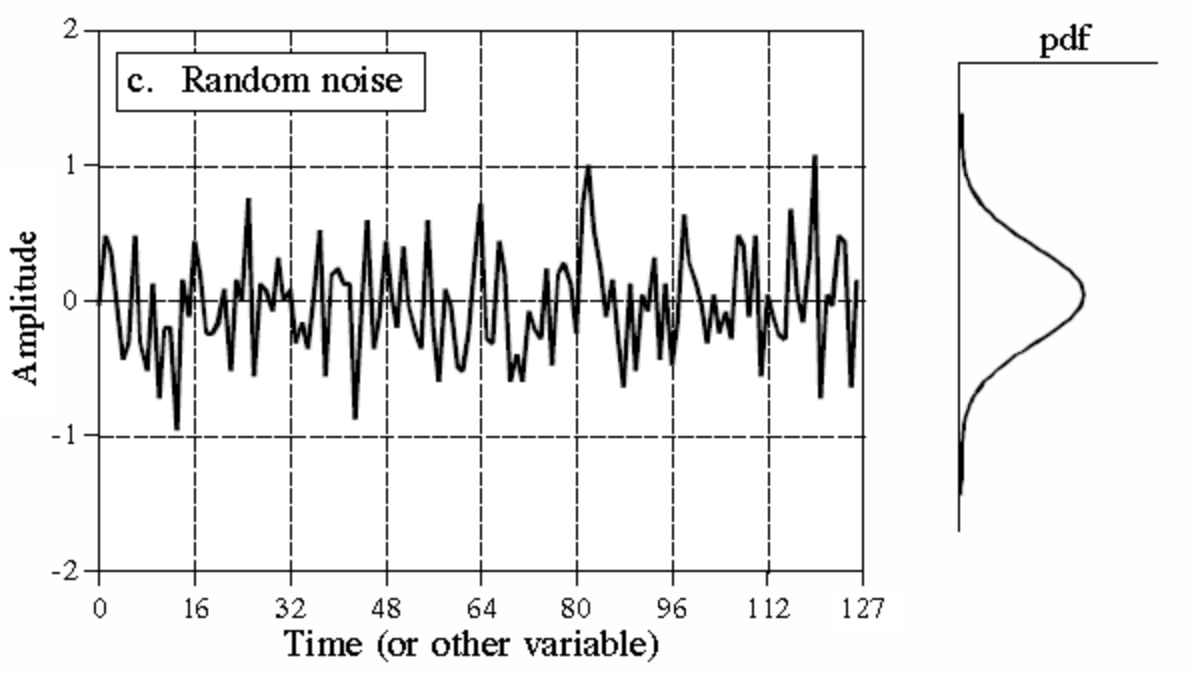
- (a) Histogram – based on finite number of samples – a statistical estimate of the underlying probability
- (b) Probability Mass Function – the underlying probability for a signal that takes on discrete values
- (c) Probability Density Function – the underlying probably for signal that is a continuous function

PDF For Square and Triangle Waves

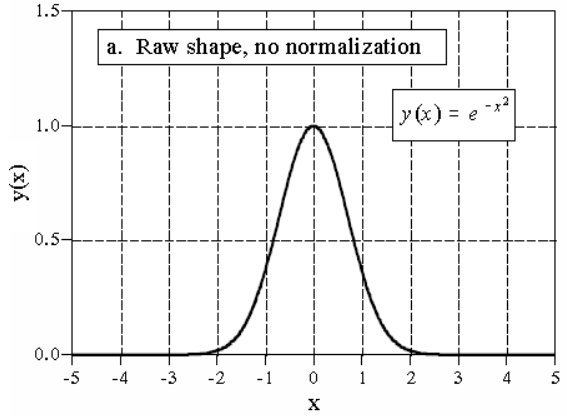


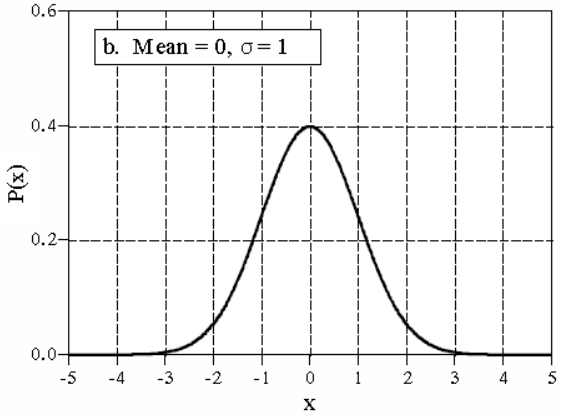
PDF for Random Noise

- PDF for this signal is a normal distribution



Normal Distribution





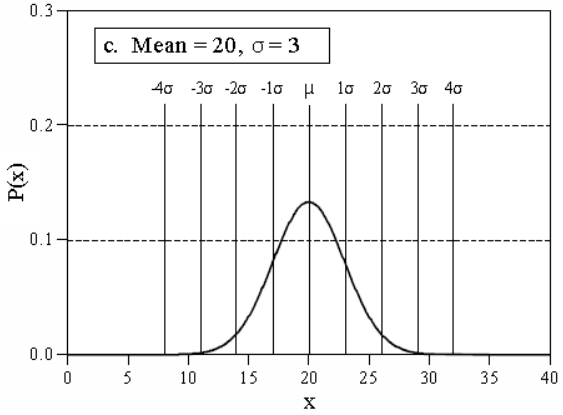


FIGURE 2-8
Examples of Gaussian curves. Figure (a) shows the shape of the raw curve without normalization or the addition of adjustable parameters. In (b) and (c), the complete Gaussian curve is shown for various means and standard deviations.

EQUATION 2-8
Equation for the *normal distribution*, also called the *Gauss distribution*, or simply a *Gaussian*. In this relation, $P(x)$ is the probability distribution function, μ is the mean, and σ is the standard deviation.

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

Characteristics of the Normal or Gaussian Distribution

- Due the exponential to the square nature, the tails of the probability density drops off very rapidly. The likelihood of values far from the mean, e.i. 4 sigma away from the mean, is very low. This is why the signal appears to have a bounded peak to peak value of 6-8 times sigma

Cumulative Distribution Function

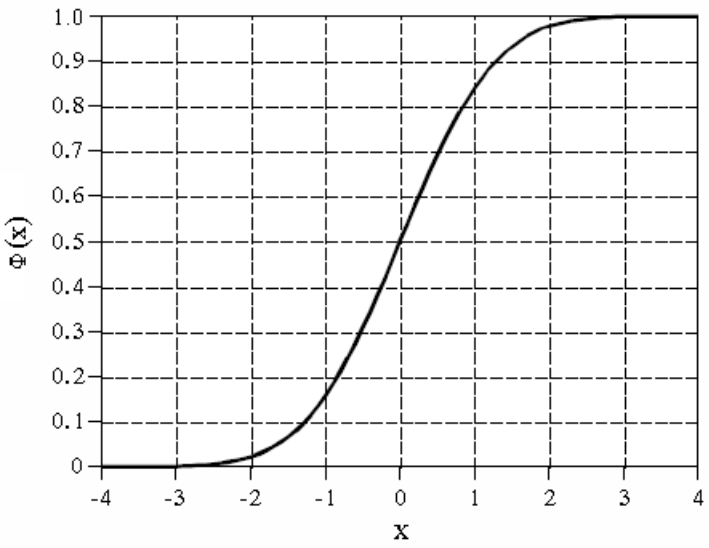


FIGURE 2-9 & TABLE 2-5
$\Phi(x)$, the cumulative distribution function of the normal distribution (mean = 0, standard deviation = 1). These values are calculated by numerically integrating the normal distribution shown in Fig. 2-8b. In words, $\Phi(x)$ is the probability that the value of a normally 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.

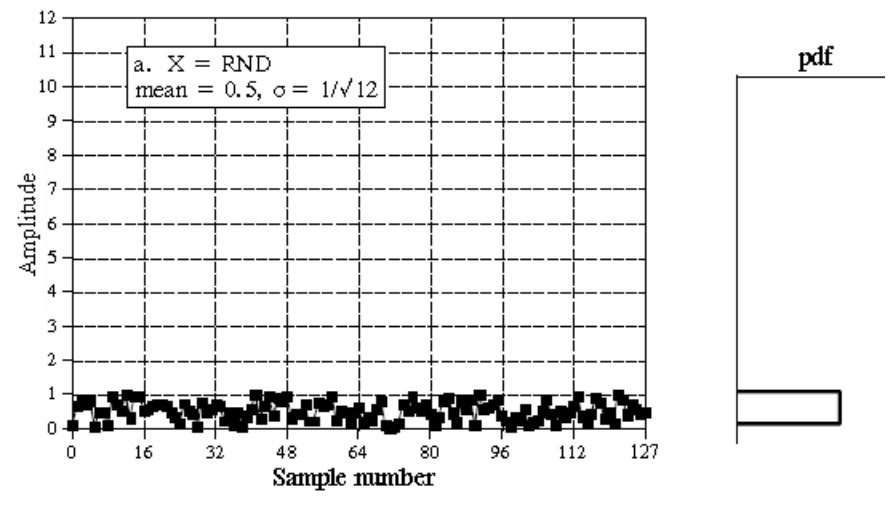
x	Φ(x)	x	Φ(x)
-3.4	.0003	0.0	.5000
-3.3	.0005	0.1	.5398
-3.2	.0007	0.2	.5793
-3.1	.0010	0.3	.6179
-3.0	.0013	0.4	.6554
-2.9	.0019	0.5	.6915
-2.8	.0026	0.6	.7257
-2.7	.0035	0.7	.7580
-2.6	.0047	0.8	.7881
-2.5	.0062	0.9	.8159
-2.4	.0082	1.0	.8413
-2.3	.0107	1.1	.8643
-2.2	.0139	1.2	.8849
-2.1	.0179	1.3	.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

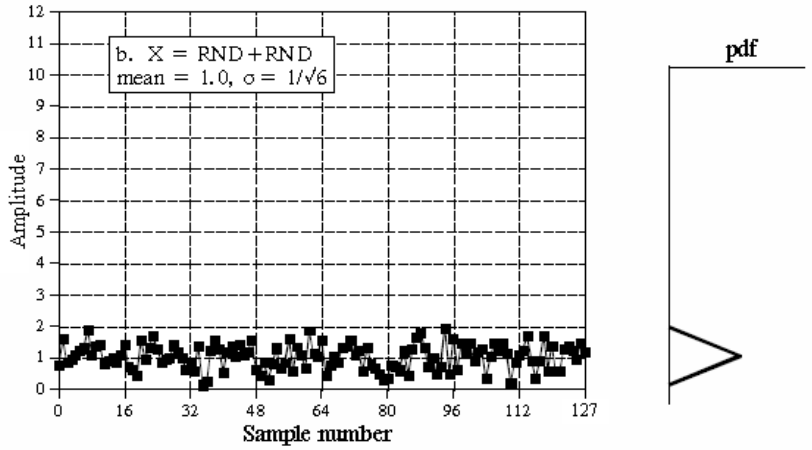
- CDF is the integral of the area under the density function

Central Limit Theorem

- The central limit theorem states that a sum of random numbers becomes normally distributed as more and more of the random numbers are added together. This is true even if the random numbers being added together are not normally distributed (e.g. they could be uniformly distributed), nor do the numbers have to come from the same distribution (they could be some that are Poisson distributed plus uniform, plus other distribution).

Generation of a Normally Distributed Random Number From a Uniformly Distributed Number





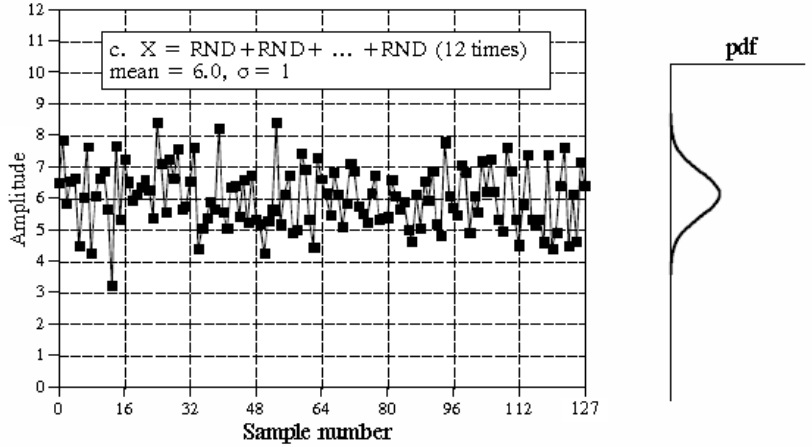


FIGURE 2-10
Converting a uniform distribution to a Gaussian distribution. Figure (a) shows a signal where each sample is generated by a random number generator. As indicated by the pdf, the value of each sample is uniformly distributed between zero and one. Each sample in (b) is formed by adding *two* values from the random number generator. In (c), each sample is created by adding *twelve* values from the random number generator. The pdf of (c) is very nearly Gaussian, with a mean of *six*, and a standard deviation of *one*.

Digital Noise Generation

- It is convenient to be able to create digital random noise for the purpose of evaluating how a DSP algorithm will perform in a noisy environment.
- Most programming languages can produce uniformly distributed random numbers.
- By repeatedly adding uniformly distributed random numbers, you can create normally (Gaussian) distributed random numbers. The mathematical theory related to this is the Central Limit Theorem.

Typical Noise Generator

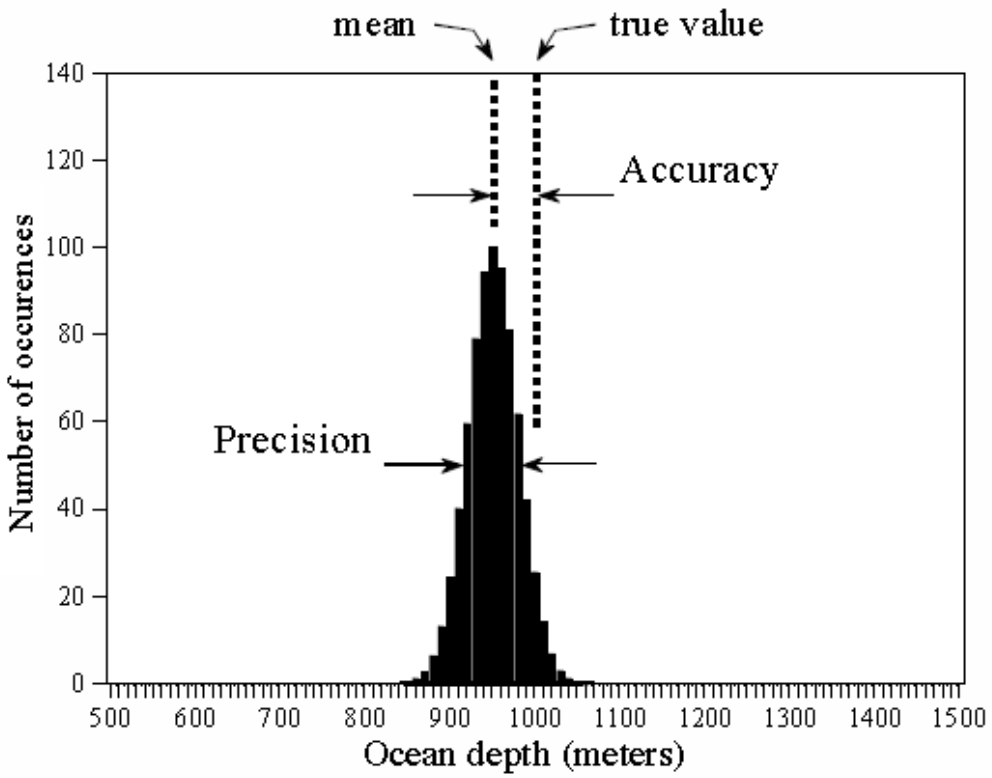
EQUATION 2-10
Common algorithm for generating uniformly distributed random numbers between zero and one. In this method, S is the seed, R is the new random number, and a,b,& c are appropriately chosen constants. In words, the quantity $aS+b$ is divided by c, and the remainder is taken as R.

$$R = (aS + b) \text{ modulo } c$$

- Generator: **gsl_rng_vax**
  This is the VAX generator MTH$RANDOM. Its sequence is,
- X_{n+1} = (a X_n + c) mod m
- with a = 69069, c = 1 and m = 2^32. The seed specifies the initial value, X_1. The period of this generator is 2^32 and it uses 1 word of storage per generator.

Precision and Accuracy Illustrated

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.



Summary of Today's Lecture

- The speed of a DSP algorithm is set by the number of multiplies and adds.
  - Different algorithms, such as running statistics, can speed up the calculation.
- Mean, variance, standard deviation
  - Variance gives the power of the deviation from the mean, not the power of the mean.
- Histogram, PMF, PDFs
  - PMF is for discrete data. PDF for continuous. Histogram is an estimate of PMF or PDF
- Digital Noise Generation
  - uniform random variables can be combined according the Central Limit Theorem to produce a normally distributed random variable.
- Precision and Accuracy
  - Precision is related to standard deviation, can be express in terms of SNR as well.

Summary of Main Points

- Random noise adds in quadrature
  - σtotal= sqrt(σ1^2 + σ2^2)
  - When many noise sources act together, the combined noise distribution is normally distributed. This is consequence of the central limit theorem.
  - Another aspect of random noise adding is quadrature is they one noise source usually dominates the total noise.
- The typical error formula is very useful in determining how many more samples you need to take to get a desired level of precision in the estimate of the true mean.