Instrumentation and data acquisition Spring 2010

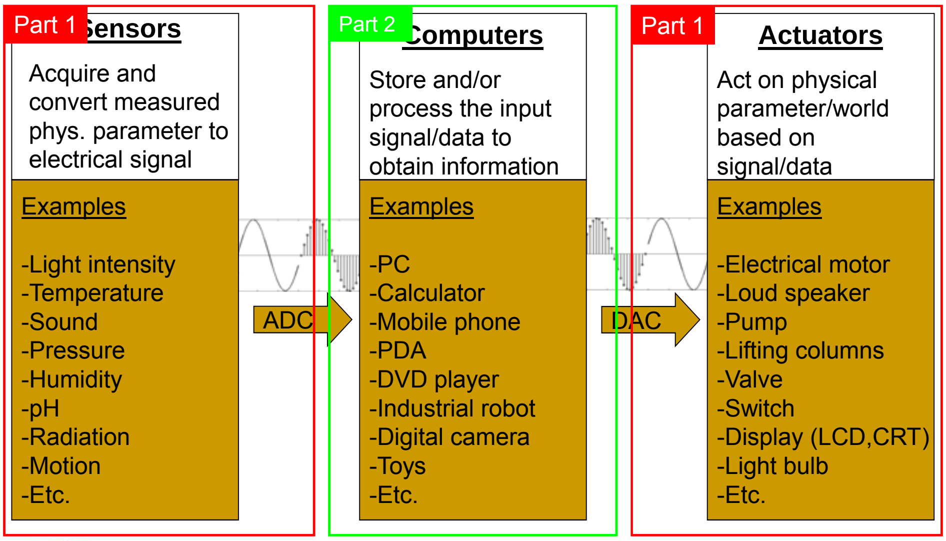
Lecture 4: Digital representation and data analysis

Zheng-Hua Tan

Multimedia Information and Signal Processing
Department of Electronic Systems
Aalborg University, Denmark
zt@es.aau.dk

Thanks to Christian Fischer Pedersen for providing some of the slides.

Acquire, process and output data



Part I: Digital representation of information

- Digital representation of information
  - Audio and image
- Acquired signal and noise
- Mean, standard deviation and SNR
- Histogram, PMF and PDF

Representation of digital audio

- Temporal resolution, i.e. sampling frequency
  - Audible frequency range: ~ 20Hz-20kHz
  - Nyquist-Shannon theorem: 2*20kHz = 40kHz
  - Actual frequency due to production: 44.1 kHz
- Bit depth, i.e. amplitude quantization
  - 16-bit linear PCM
    - Digital audio stored in computers: Windows WAV, Apple AIF, Sun AU
    - Compact Disc – Digital Audio
  - 16 bit/sample per channel ~ 2^16 = 65,536

Representation of digital audio

- Bit-rate stereo
  - 16*2*44100 ~ 1.4 Mbit/s
- A CD can store up to 74 minutes of music

Total amount of data =

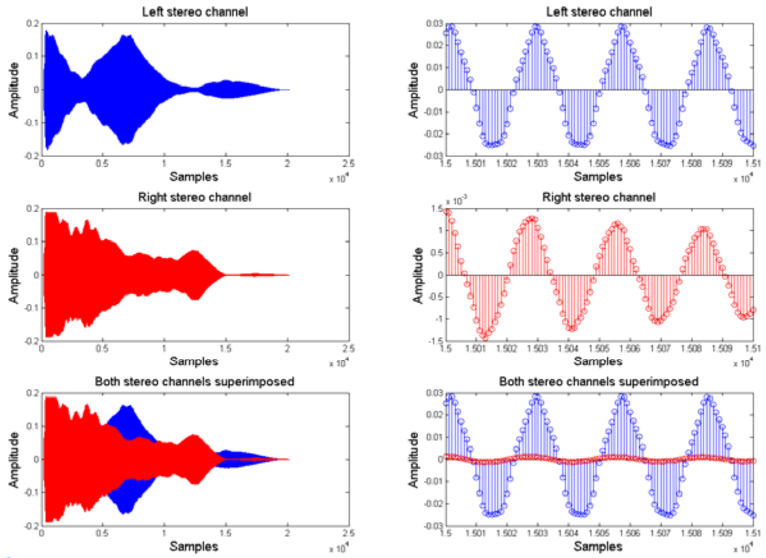
44,100 samples/(channel*second) * 2 bytes/sample *

2 channels * 60 seconds/minute * 74 minutes

= 783,216,000 bytes

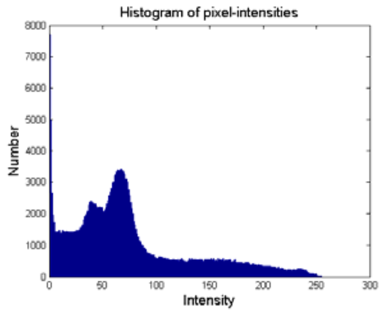
- There are CD-Rs that can hold 700 megabytes (734,003,200 bytes) of error corrected data, or 80 minutes of stereo 16 bit 44.1 kHz audio (846,739,200 bytes) (807.51 megabytes) without error correction code.

Representation of digital audio



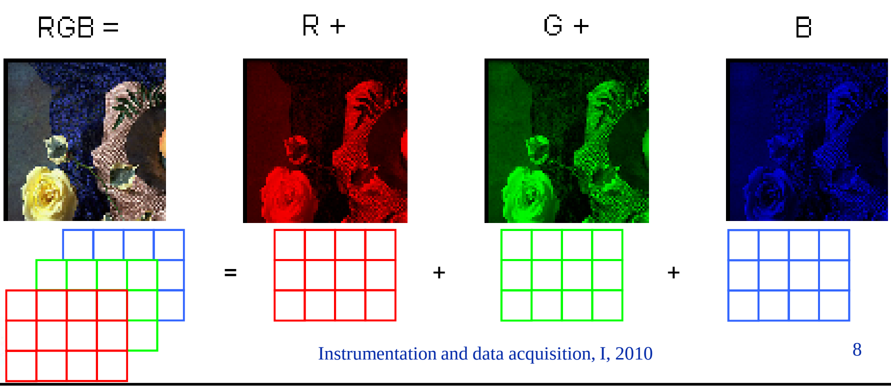
Rep. of digital grayscale bitmap image

- Spatially mapped array, e.g.
  - Resolution: 512x512x1 [uint8];
  - Bit-depth=8 => 2^8 = 256^1 = 256 intensities/pixel
  - Space as independent variables.: I(x,y)



Rep. of digital colour bitmap image

- Spatially mapped array of bits, e.g.
  - Resolution: 300x300x3 [uint8];
  - Bit-depth=24 => 2^24=256^3=16,777,216 colors/pixel
  - Space as independent vars.: {r(x,y),g(x,y),b(x,y)}



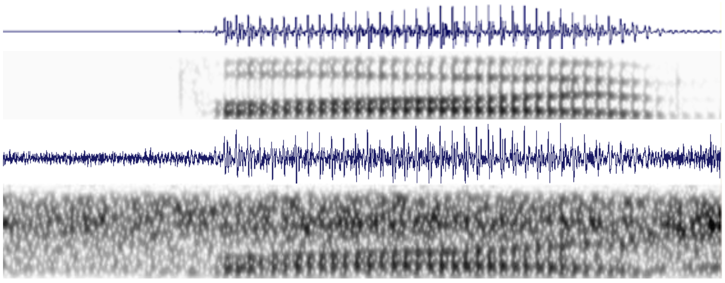
Rep. of digital colour bitmap image

R: 83 G: 97 B: 86	R: 94 G:106 B: 97	R: 80 G: 91 B: 79	R: 78 G: 90 B: 73	R: 84 G: 95 B: 81	R: 89 G:101 B: 87	R:101 G:113 B:100	R:101 G:113 B:100	R: 88 G: 99 B: 89	R: 79 G: 89 B: 79
R:102 G:115 B:105	R:105 G:118 B:108	R: 84 G: 97 B: 85	R: 90 G:103 B: 86	R: 88 G: 99 B: 86	R: 92 G:105 B: 91	R:106 G:118 B:106	R: 96 G:108 B: 97	R: 77 G: 89 B: 79	R: 64 G: 75 B: 67
R:101 G:114 B:104	R:104 G:117 B:108	R: 78 G: 92 B: 79	R: 84 G: 98 B: 81	R: 92 G:106 B: 91	R:102 G:116 B:102	R:109 G:123 B:110	R: 78 G: 91 B: 83	R: 55 G: 68 B: 61	R: 57 G: 70 B: 62
R:102 G:114 B:104	R:107 G:122 B:112	R: 81 G: 96 B: 82	R: 84 G: 97 B: 81	R: 95 G:109 B: 95	R:103 G:116 B:103	R:103 G:116 B:106	R: 72 G: 84 B: 76	R: 63 G: 75 B: 71	R: 65 G: 77 B: 74
R: 96 G:110 B:100	R:107 G:123 B:113	R: 88 G:104 B: 91	R: 88 G:104 B: 85	R: 84 G: 99 B: 85	R: 72 G: 86 B: 74	R: 81 G: 95 B: 85	R: 89 G:102 B: 95	R:116 G:130 B:126	R: 95 G:108 B:109
R:103 G:119 B:107	R:110 G:128 B:114	R: 99 G:116 B:102	R:100 G:117 B: 99	R: 80 G: 96 B: 81	R: 60 G: 76 B: 63	R: 78 G: 95 B: 83	R:103 G:119 B:110	R:124 G:140 B:132	R: 95 G:110 B:106
R:118 G:135 B:122	R:115 G:131 B:113	R:104 G:121 B:104	R:103 G:118 B:107	R: 87 G:103 B: 91	R: 79 G: 95 B: 82	R: 92 G:108 B: 96	R: 95 G:111 B: 98	R: 76 G: 92 B: 79	R: 67 G: 83 B: 69
R:115 G:131 B:119	R:106 G:122 B:107	R: 89 G:106 B: 91	R: 83 G: 99 B: 90	R: 82 G: 98 B: 88	R: 76 G: 92 B: 82	R: 78 G: 94 B: 84	R: 74 G: 90 B: 80	R: 67 G: 83 B: 73	R: 69 G: 85 B: 75
R:108 G:123 B:115	R: 98 G:115 B:102	R: 85 G:101 B: 90	R: 83 G: 99 B: 91	R: 84 G: 99 B: 91	R: 77 G: 92 B: 84	R: 75 G: 90 B: 83	R: 72 G: 88 B: 80	R: 71 G: 86 B: 78	R: 73 G: 88 B: 80

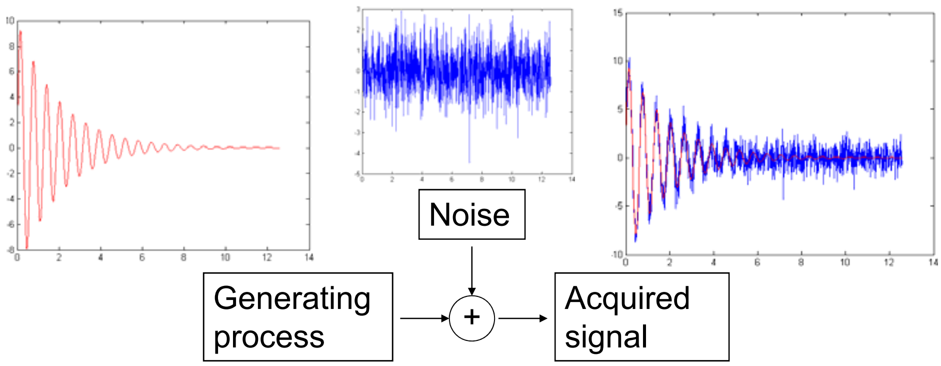
Part II: Acquired signal and noise

- Digital representation of information
- Acquired signal and noise
- Mean, standard deviation and SNR
- Histogram, PMF and PDF

Acquired signals



Acquired signal vs. generating process



Acquired signal vs. generating process

- Example: Coin flip
- Generating process
  - Assign: Head=1, Tail=0
  - Constant probability: P(H)=P(T)=50% => μ=0.5
- Acquired signal
  - Flip a coin 1000 times
  - Create a signal: (flip_no,outcome)
  - Varying statistical mean: μ≠0.5 but μ~0.5
  - Statistical dispersion: variance, standard deviation

Acquired signal vs. generating process

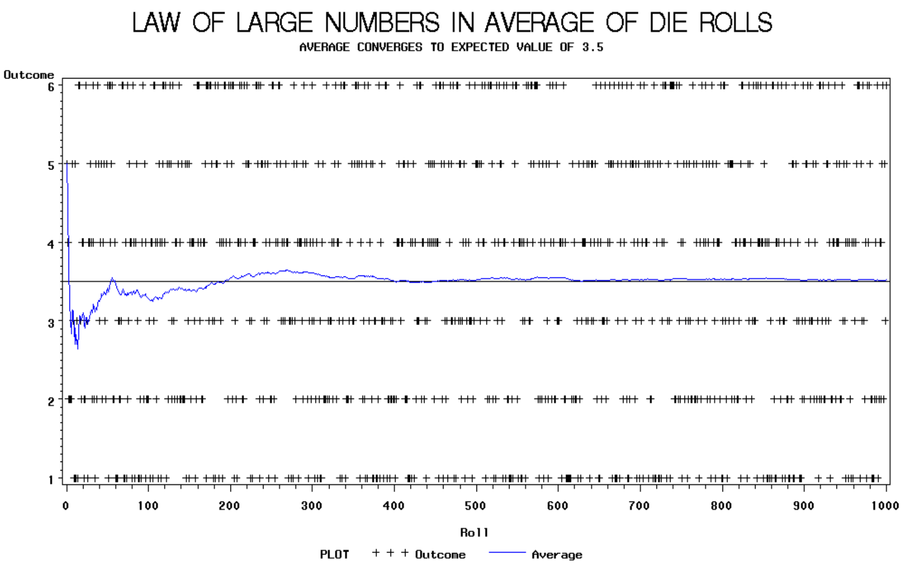


Figure: Courtesy of Shivan Bird

Data analysis

- Inference, noise and other undesirable components exist in acquired data.
  - Imperfections in the data acquisition system
  - Inherent part of the signal being measured
  - By-product of some DSP operation
- There is a need to reduce them and characterise signals and the processes that generate them

  -> statistics and probability theory

Part III: Mean, standard deviation and SNR

- Digital representation of information
- Acquired signal and noise
- Mean, standard deviation and SNR
- Histogram, PMF and PDF

Sample indexing in discrete-time signals

- Number of samples in a signal: N
- Each element in the set *x* is a sample
- The samples can be indexed in two ways

  $n \in [0; N-1]$ or $n \in [1; N]$
- We will use the first indexing range

  $n \in [0; N-1]$
- In MATLAB

  $n \in [1; N]$

Mean and deviations

- Mean – the average value of a signal

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

- Average deviation – amplitude fluctuation

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

- Standard deviation – power fluctuation

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

- Variance – power of fluctuation

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

Unbiased estimator

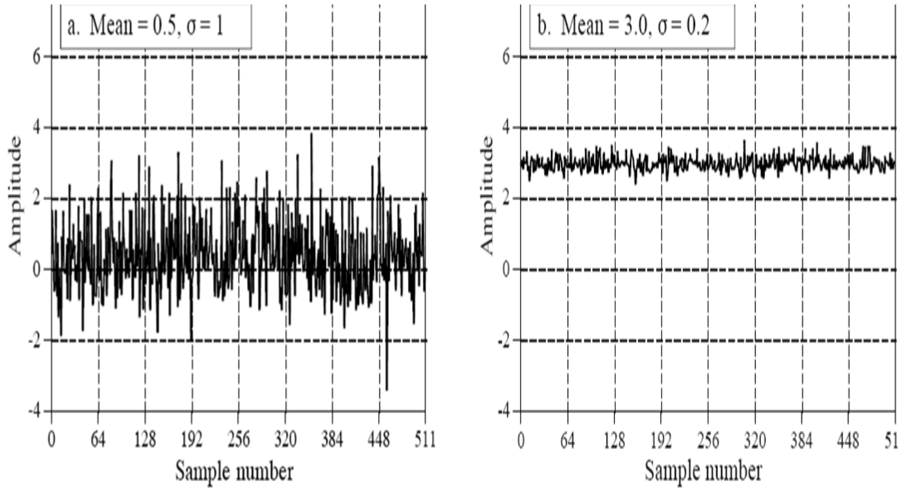
- If the underlying distribution is not known, then the sample variance may be computed as

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

- Note that the sample variance defined above is *not* an unbiased estimator for the population variance σ^2.
- In order to obtain an unbiased estimator for σ^2, it is necessary to instead define a "bias-corrected sample variance"

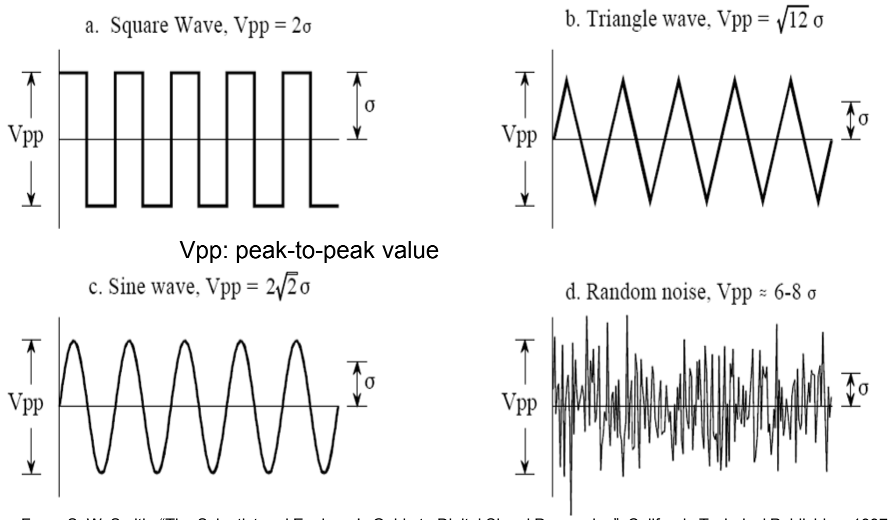
$$S_{N-1}{}^2 = \frac{1}{N-1}\sum_{i=0}^{N-1}(x_i - \mu)^2$$

Mean and deviations



From: S. W. Smith, “The Scientist and Engineer's Guide to Digital Signal Processing”, California Technical Publishing, 1997

Mean and deviations

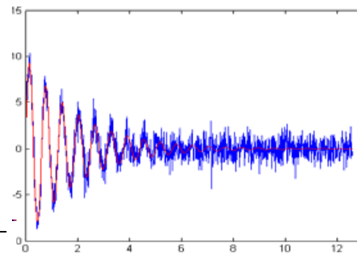


Vpp: peak-to-peak value

From: S. W. Smith, “The Scientist and Engineer's Guide to Digital Signal Processing”, California Technical Publishing, 1997

Signal-to-noise ratio (SNR)

$$\mathrm{SNR} = \frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}} = \left(\frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}}\right)^2$$



- where P is average power and A is root mean square (RMS) amplitude

$$x_{\mathrm{rms}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}}$$

- SNR = 20 log (Signal RMS / Noise RMS)

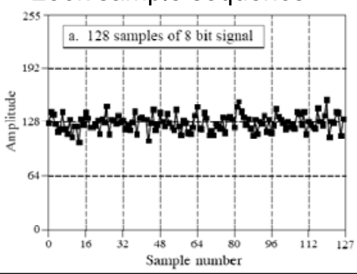
$$\mathrm{SNR(dB)} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) = 20\log_{10}\left(\frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}}\right)$$

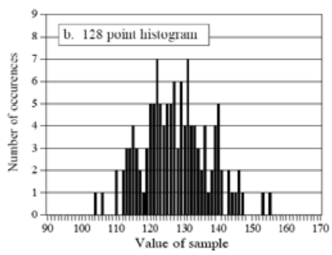
Part IV: Histogram, PMF and PDF

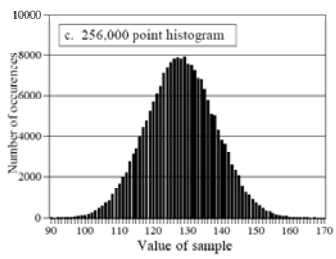
- Digital representation of information
- Acquired signal and noise
- Mean, standard deviation and SNR
- Histogram, PMF and PDF

Histogram

128 samples from a
256k sample sequence







$$|\{x(n)\}| = N = \sum_{i=0}^{M-1} H_i$$

N: No. of samples
H_i: No. of occurrence in *i*'th histogram bin
M: No. of histogram bins

From: S. W. Smith, "The Scientist and Engineer's Guide to Digital Signal Processing", California Technical Publishing, 1997

Histogram and statistics

Histogram based (for *the* histogram in previous slide)

Mean

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

Variance

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

Sample based

Mean

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

Variance

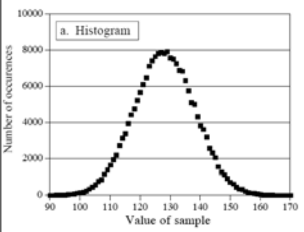
$$\sigma^2 = \frac{1}{N}\sum_{i=0}^{N-1} (x_i-\mu)^2$$

- For very large data sets it is computationally more effective to bin the data in histograms and base the statistics upon these
- Histogram's applications in image segmentation and enhancement.
- The need for PDF for modelling and classification applications.

Histogram, PMF and PDF

Histogram —infer→ PMF | PDF

Histogram



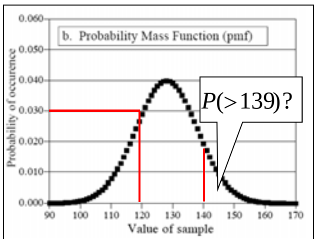
- For discrete data/signals
- From acquired signal
- Number of samples finite

-Y-axis:
No. of occur: $H_i \in [0; N-1]$

-Area under histogram

$$|\{x(n)\}| = N = \sum_{i=0}^{M-1} H_i$$

PMF



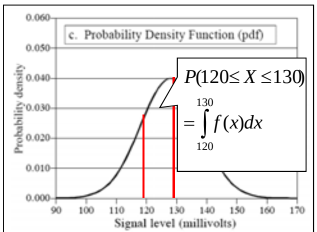
- For discrete data/signals
- From generating process
- Number of samples infinite
- Estimated from histogram

- Y-axis:
Frequency: $H_i / N \in [0;1]$

- Area under PMF

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

PDF



- For continuous data/signals
- From generating process
- Continuous thus inf. domain
- Estimated from histogram

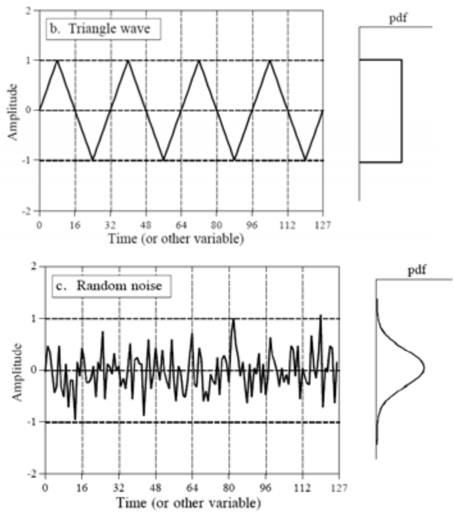
-Y-axis
Prob. density: $f : R \mapsto R$

-Area under PMF

$$1 = \int_{-\infty}^{\infty} f(x)dx$$

From: S. W. Smith, "The Scientist and Engineer's Guide to Digital Signal Processing", California Technical Publishing, 1997

Histogram, PMF and PDF – cont.



From: S. W. Smith, "The Scientist and Engineer's Guide to Digital Signal Processing", California Technical Publishing, 1997

Binned histograms

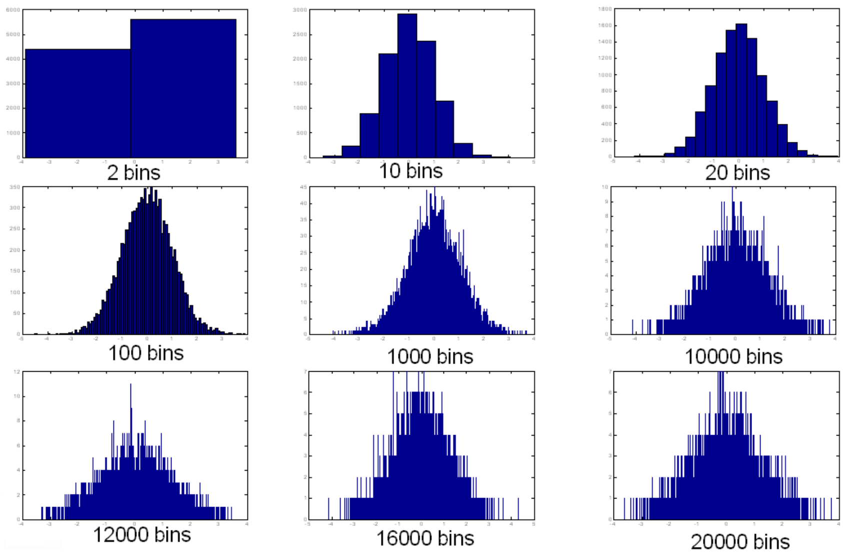
- When to use?
  - If the levels each sample can take on is much larger than the number of samples in the signal, i.e. sample space >> |{x(n)}|=N. For example,
    - Matlab double: 64bit = 2^64 = 1.84*10^19
    - Signal: 10.000 samples
    - => sample space >> No. of samples
- Why?
  - Cannot count No. samples corresponding to each quantization level as No. of quantization level very high – computationally ineffective.
  - Many quantization levels have no corresponding sample.

Binned histograms – cont.

- How many bins to use? It is a compromise!
- Too many bins
  - Difficult to estimate amplitude of underlying PMF, because only few samples (if any) fall within each bin
- Too few bins
  - Difficult to estimate the domain of underlying PMF, because of crude resolution

Binned histograms – cont.

Signal: 10000 samples --- Sample space: 2^64=1.84*10^19



Summary

- Digital representation of information
- Acquired signal and noise
- Mean, standard deviation and SNR
- Histogram, PMF and PDF