BIOE 198MI Biomedical Data Analysis. Spring Semester 2019.
Lab 4: Introduction to Probability and Random Data

A. Random Variables

Randomness is a component of all measurement data, either from acquisition noise, uncertainties in the measurement process, or randomness in the measured object. Let random variable X, with specific values x, represent measurement data that are not exactly predictable. While deterministic variables are known exactly, random variables are known only statistically. This means the likelihood of the next data value is predictable by a frequency distribution. Examples of random variables include the number of people in a crowd with blue eyes and the number of cells undergoing division is a cell culture.

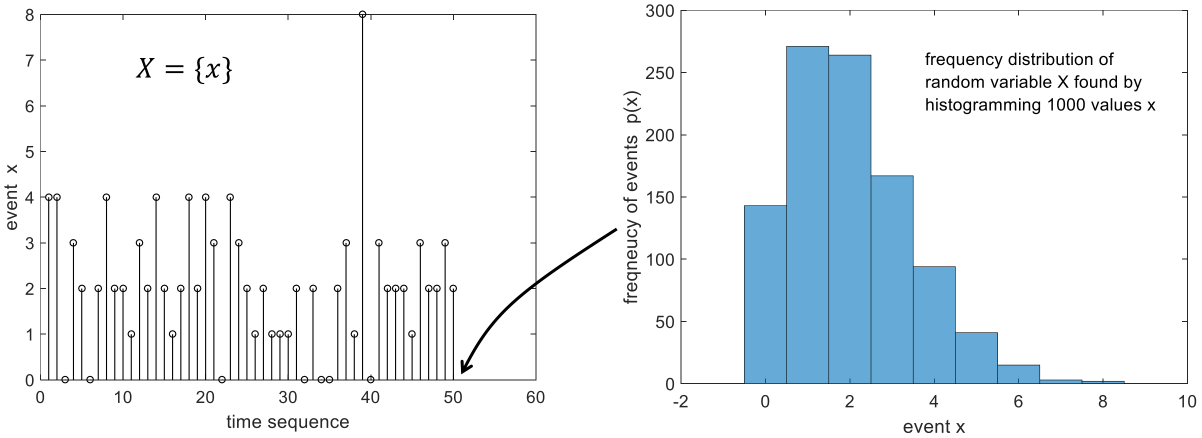

Figure 1. Illustration of random variable X having specific values x (left). Each value is 'drawn' from a frequency distribution (right) that is called a probability density function (pdf). This specific discrete random variable is from a Poisson process that will not be discussed in this lab.

B. Random Process. Random process X is a dependent random variable expressed as a function of deterministic independent variables like time t and position (ξ, γ). A random process of all three variables is written as $X(t, \xi, \gamma)$. Examples include image data recorded over time; e.g., a movie.

C. Simulating 1-D Deterministic and Random Processes using Normally Distributed Process. In the following script, we simulate a voltage signal with noise. Let's generate $N = 1001$ noise samples $x(t)$ and plot just the first 51 samples. Then histogram the entire sequence and plot the normal pdf from which those samples were drawn to show there is a match.

```
%% Normal pdf describing measurement noise X(t)
clear all;close all;
t=0:0.01:0.5;          %time axis [s] for plotting
y=10*sqrt(0.5-t);      %arbitrary deterministic voltage y(t)
subplot(2,1,1);plot(t,y);xlabel('t [s]');ylabel('y(t) [V]')
N=1001;                %number of samples for histogram
x=randn(1,N);          %generate normal random samples
subplot(2,1,2);stem(t,x(1:51));xlabel('t [s]');ylabel('X(t) [V]')
subplot(2,1,1);hold on;plot(t,y+x(1:51),'o')
legend('show', 'y(t)','y(t)+x(t)');hold off
```

```
figure;histogram(x,'normalization','pdf')
xlabel('X [V]');ylabel('pdf(X)')
hold on; z=pdf('norm',-4:0.1:4,0,1);plot(-4:0.1:4,z);hold off
figure;histfit(x);xlabel('X [V]');ylabel('hist(X)')
%
```

There are many things to note from this script and about random variable distributions in general.

- We can generate random-variable samples drawn from a standard normal probability density function (pdf) using `randn(N,M);`
- We can also model the normal pdf from which samples are drawn. That pdf has mean μ and standard deviation σ and is found using `normpdf(x,m,s)` or `pdf('norm',x,m,s)`. The general equation is $p(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$ but the standard normal equation is $p(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$. where $x = \sigma z + \mu$.
- Also $\int_{-\infty}^{\infty} dx\, p(x) = 1$ not only for normal pdfs but for any pdf; e.g., Poisson, binomial….
- The probability that x falls between values a and b is $\Pr(a < x < b) = \int_a^b dx\, p(x)$ (fig 2).
- `stem(x,y)` is how to plot ‘samples’.
- What happens when N in the code is varied?
- What is the difference between `histogram` and `histfit` functions? (Hint: `help histfit`).
- What are the inputs to `pdf`? (See 2nd to last line in code above)
- The horizontal axis in pdf plot often represents measurements and therefore has units.

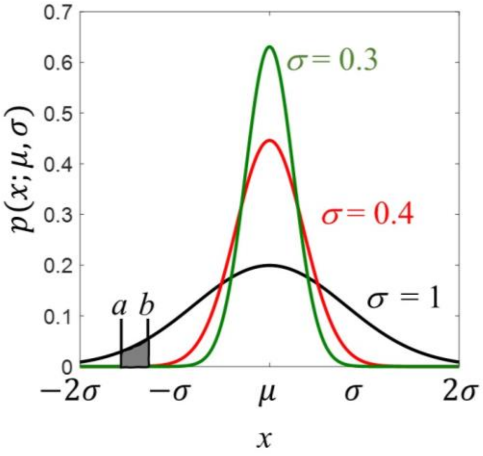

Figure 2. Three normal pdfs all with mean μ but with three different σ. The probability that samples fall within the range of a and b is $\Pr(a < x < b)$. The peaks are different because the areas

Exercise 1: (Lab 1 basic skills revisited)

(a) Using equations above for $p(z;0,1)$ and $p(x;\mu = 1,\sigma = 2)$ to plot each. Do not use `pdf` or similar intrinsic function from Matlab; program the equations. Be sure to label axes and add a legend. (b) Repeat using `myplot`.

D. Visualization Options. One alternative way to visualize the relationship between random samples and the frequency distribution from which they are drawn (the pdf) is found in Fig. 3. Here we have a noisy time-varying voltage, $v(t)$, measured at an outlet of a house with problems. Samples are summarized by the histogram where the histogram axis is aligned with the voltage axis. 100 s of data is sampled at 10 ms (100 s / 0.01 s = 10,000 samples). The mean voltage is $\bar{v}(t) = 120$ V, and the standard deviation is $\sigma = 10$ V.

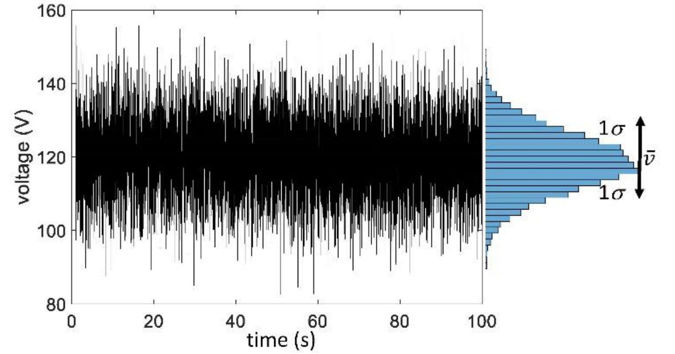

Fig. 3. Noisy voltage trace on left. A histogram of 10^4 voltage values on the right is over 100 s of measurements (Δt=0.01s).

E. Probability and pdf area: introducing cumulative distribution function (CDF).

- The total area under any pdf must equal one. In words, the probability of something occurring is certain. In math terms, $\mathrm{Pr}(-\infty < v < \infty) = \int_{-\infty}^{\infty} dv\, p(v) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} dv\, e^{-(v-\mu)^2/2\sigma^2} = 1.0 = CDF(\infty)$.
- What is the probability of finding voltages in Fig 3 at any instant of time with values between $\pm\sigma$? A. $\mathrm{Pr}(-\sigma < v < \sigma) = \int_{-\sigma}^{\sigma} dv\, p(v) \cong 0.68$. I found that number via CDFs. Let's explore.
- Cumulative distribution function (CDF) is $CDF(a) = \mathrm{Pr}(-\infty < v < a) = \int_{-\infty}^{a} dv\, p(v)$. See Fig 4 (left). In Matlab: `cdf('norm',0,0,1), normcdf(0,0,1), and cdf('norm',120,120,10)` all equal 0.5. Do you see why?

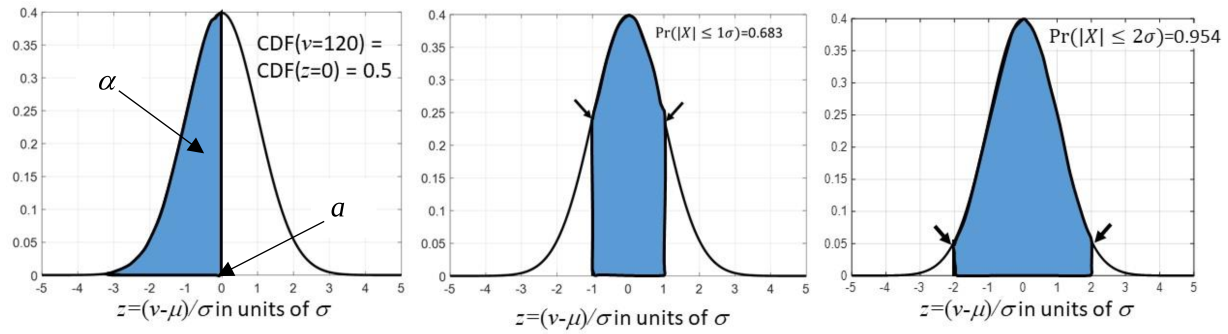

Fig 4. Standard normal pdfs (converted from the voltage pdf in Fig 3. Illustrated are (left) Pr(-∞ < X < 0) = 0.5; (center) Pr(-σ < X < σ) = 0.68; (right) Pr(-2σ < X < 2σ) = 0.95.

- Exercise 2: Type the following.

```
clear all;close all
m=120;s=10;
x=m-4*s:0.01:m+4*s;
p=normpdf(x,m,s);
myplot(x,p);xlabel('x');ylabel('y')
```

  Find Pr(-σ<X<σ), Pr(-2σ<X<2σ) , Pr(-3σ<X<3σ) , and Pr(-4σ<X<4σ) ,.
- In contrast we might have the probability value α and now wish to know the value of $z = a$. If CDF(a) = α , then we are seeking CDF^{-1}(α) = CDF^{-1}(CDF(a)) = a. For this, we use the Matlab function `norminv(0.5,0,1)` or `icdf('norm',0.5,0,1)` which we find is zero. As part of Exercise 2:

  Find $x = a$, for $a = \mu - \sigma$. and μ = 120V, σ = 10V. Without using Matlab further, tell me the value of $x = a$, for $a = \mu + \sigma$.

I hope that you have a clearer sense of what is meant by “statistically knowing” a random variable. This case assumes either a normal random voltage $v(t)$ or its standard normal equivalent $z(t)$.

F. Mean, variance and standard deviation of a normal random variable

- Note that the expression $X{\sim}p(x;\mu,\sigma)$ is shorthand to indicate X is a normal random variable that is specified entirely by just two parameters, μ and σ.
- We can (but won’t) show μ is the population mean or ensemble mean, and σ is population standard deviation or ensemble standard deviation. Both values are model parameters that are difficult to measure exactly.
- σ^2 indicates the ensemble variance for a normal random process.
- The ensemble mean is the most likely value and the peak of the normal pdf. (Median is the middle value in a pdf and mode is the value that occurs most often. For asymmetric distributions, the mean, median, and mode are not the same.)
- The ensemble standard deviation is a measure of the normal pdf width.
- Estimate μ : the sample mean from N measurements or samples is $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$.
- Estimate σ^2 : the sample variance from N samples is $s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2$.
- The sample standard deviation is the square root of the sample variance, $s = \sqrt{s^2}$.

Exercise 3: (a) Generate a 1000x1000 array of standard normal random values using `p=randn(1000);`first use the function `rng(‘default’)` so we all obtain the same values. Image the result using a grayscale mapping (use `colormap gray`) and generate a `colorbar`.
(b) Compute sample averages for the first 1x1 value, first 10x10 values, first 20x20 values, etc. up to 100x100 using a `for` loop. Plot the 10 sample means as a function of the number of values used in the sample average.

Another Matlab tool of interest is `rng` used to “seed” a random number generator. Run these twice.

```
rng('default');c=randn(3,1)  %resets seed to reproduce c w/ea call
rng('shuffle');c=randn(3,1)  %seed randn with clock time to vary c
```

Assignment (Introduction Section of Lab Report):
In this assignment, we apply waveform averaging to noisy measurements to improve “data quality.” The first stage of the assignment is to simulate M = 100 voltage waveforms $y(t)$= $x(t)$+ $n(t)$, each consisting of the same deterministic signal $x(t)$ plus an independent, zero-mean noise traces $n(t)$. The next stage is to average signals as directed, and then measure the standard deviation of the waveform-averaged noise $n(t)$= $y(t)$- $x(t)$. Finally, report on changes to the signal-to-noise ratio (SNR) as a function of the number of waveforms averaged. Theory suggests that the sample standard deviation s at each time point in a voltage waveform with additive normally distributed independent noise should decrease by a factor of $1/\sqrt{N}$ where N is the number of waveforms averaged. Let’s use signal simulations to test that idea.

Simulate 100 noisy sinusoidal voltage waveforms, where in each waveform the deterministic signal is the same 1 Hz, 10-cycle sine wave $x(t) = \sin(2\pi u_0 t)$. Add an independent, normally distributed noise trace, $n(t){\sim}p(x;0,\sigma)$, to the same sinusoidal signal to model 100 noisy measurements of a signal. How much noise do we add? That question is specified by the signal-to-noise ratio (SNR). Set SNR = 1 dB. Then find the modeled noise amplitude σ by solving for σ in the equation, $SNR\ (\text{dB}) = 10\log_{10}\left(\frac{MSV}{\sigma^2}\right)$, where $MSV = \frac{1}{10T_0}\int_0^{10T_0} dt\ \sin^2(2\pi\, u_0 t)$ is the mean-squared voltage (MSV) value of $x(t)$ measured in volts squared, and $u_0 = \frac{1}{T_0}$ is the sinusoidal frequency in Hz. Note that $\sin^2(2\pi u_0 t) = \frac{1}{2}(1 - \cos(4\pi u_0 t))$.

Simulate the data (Methods Section of Lab Report)…

(a) First, compute MSV using your calculus skills.

(b) Second, compute σ by combining the SNR equations and $MSV.$

(c) Generate a time axis in [s], the sinusoidal signal in [V], and an $M \times N$ matrix of standard normal random noise series, $z(t)$ where $M = 100$ and $N = 10/\Delta t + 1$. You must also select a sampling interval Δt [s]. Be sure to scale each row of the standard normal random samples in array z to find, $n(t) = \sigma z(t)$. Now the noise will correspond to the correct SNR.

(d) Use a `for` loop to add signal $x(t)$ to each independent noise trace $n(t)$. You should generate 100 waveforms $y(t) = x(t) + n(t)$. Now you have the simulated data! In a 4×1 subplot, plot one of the recorded waveforms that you just simulated. Be sure axis labels have units.

Analyze the waveform data (Results Section of Lab Report)…

(e) Compute the standard deviation of the noise in the first waveform using $std(n;1) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_1(t_i) - x(t_i))^2}$. In Matlab: `s = std(y(1,:)-x).` Recall the additive noise was generated to be zero mean. However, each measured waveform contains the same non-zero-mean deterministic signal x, which must be subtracted; e.g., $n_1(t) = y_1(t) - x(t)$.

(f) Average the first j waveforms, e.g., at $j = 2$, $\bar{y}(t;2) = (y_1(t) + y_2(t))/2$. Hint: `ybar=mean(y(1:j,:).` Repeat for each j over range $2 \le j \le 100$ using $\bar{y}(t;j) = \frac{1}{j}\sum_{k=1}^{j} y_k\,(t)$. For example, $\bar{y}(t;25)$ represents the mean temporal waveform from averaging 25 recorded waveforms.

(g) For each j, compute the noise standard deviation, $std(n;j) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}(t_i;j) - x(t_i))^2}$.

(h) Plot both $\bar{y}(t;100)$ and $std(n;j)$ in the subplot matrix as plots 2 and 3, respectively.

(i) Compute and plot $SNR\ (\,j\,)$ as a function of the number of waveforms averaged, j.

Discuss findings (Discussion Section of Lab Report)…

(j) The sample standard deviation of noise for one waveform is s. Theory tell us the standard deviation for an average of j waveforms should be $s_{\bar{x}} = s/\sqrt{j}$. On top of the plot of sample std versus j in part (g) above, plot the theoretical prediction. Plot every 10th point to more clearly show the match. The plot belongs in RESULTS but the discussion belongs in the DISCUSSION section

(k) Summarize the findings and discuss the tradeoffs when deciding how many waveforms to average. How many waveforms must be averaged to increase the SNR by 20dB? If the cost of each experiment (one waveform) is $1000, how many experiments would you advise the company request?