Public Service Announcements

• HW1 has been graded and you should get an e-mail from the gradescope
• HW2 will be posted next Tuesday
• The midterm is on March 18 and will cover material in HW1 and HW2
• You should receive invitations to register for misterm from CBTF
• Only 6 of you went to office hours yesterday: based on the results of Matlab exercise many more need it. Watch Zoom recording of office hours
What I want you to do in Matlab exercises

• Stats=6;
• \( r2=[2,2,1,3,2,3] \);
• mean(r2); \( (2+2+1+3+2+3)/6=2.17 \)
• % How to approximate the PMF? Bin variables
  • \([a,b]=\text{hist}(r2,1:3); \ % b=[1,2,3]; a=[1,3,2];\)
  • \( p\_\text{dist}=a./\text{sum}(a); \ % p\_\text{dist}=[1/6, 3/6, 2/6];\)
  • figure; \text{plot}(b,p\_\text{dist},'ko-');
  • % mean=sum
  \( p\_\text{dist}(x)*x=1*1/6+2*3/6+3*2/6=2.17, \text{ same as in } (2+2+1+3+2+3)/6 \)
Matlab exercise: Poisson distribution

• Generate a sample of size 100,000 for Poisson-distributed random variable X with $\lambda = 2$

• Plot the approximation to the Probability Mass Function based on this sample

• Calculate the mean and variance of this sample and compare it to theoretical calculations:
  $E[X] = \lambda$ and $V[X] = \lambda$
Continuous Probability Distributions

Uniform Distribution
Important terms & concepts for discrete random variables

- Probability Mass Function (PMF)
- Cumulative Distribution Function (CDF)
- Complementary Cumulative Distribution Function (CCDF)
- Expected value
- Mean
- Variance
- Standard deviation
- Uniform distribution
- Bernoulli distribution/trial
- Binomial distribution
- Poisson distribution
- Geometric distribution
- Negative binomial distribution

*Boldface and underlined are the same for continuous distributions*
Continuous & Discrete Random Variables

• A **discrete random variable** is usually integer number
  – N – the number of proteins in a cell
  – D - number of nucleotides different between two sequences

• A **continuous random variable** is a real number
  – C=N/V – the concentration of proteins in a cell of volume V
  – Percentage D/L*100% of different nucleotides in protein sequences of different lengths L (depending on set of L’s may be discrete but dense)
Probability Mass Function (PMF)

• $X$ – discrete random variable

• Probability Mass Function: $f(x) = P(X=x)$ – the probability that $X$ is exactly equal to $x$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$P(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.6561</td>
</tr>
<tr>
<td>1</td>
<td>0.2916</td>
</tr>
<tr>
<td>2</td>
<td>0.0486</td>
</tr>
<tr>
<td>3</td>
<td>0.0036</td>
</tr>
<tr>
<td>4</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

$\sum_x P(X=x) = 1.0000$
Probability Density Function (PDF)

Density functions, in contrast to mass functions, distribute probability continuously along an interval.

Figure 4-2 Probability is determined from the area under $f(x)$ from $a$ to $b$. 
Probability Density Function

For a continuous random variable $X$, a **probability density function** is a function such that

1. $f(x) \geq 0$ means that the function is always non-negative.

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

3. $P(a \leq X \leq b) = \int_{a}^{b} f(x) dx = \text{area under } f(x) dx \text{ from } a \text{ to } b$
Normalized histogram approximates PDF

A histogram is graphical display of data showing a series of adjacent rectangles. Each rectangle has a base which represents an interval of data values. The height of the rectangle is a number of events in the sample within the base.

When base length is narrow, the histogram could be normalized to approximate PDF \( f(x) \):

\[
\text{height of each rectangle} = \frac{(\# \text{ of events within base})}{(\text{total} \# \text{ of events})/\text{width of its base}}.
\]

Normalized histogram approximates a probability density function.
Cumulative Distribution Functions (CDF & CCDF)

The **cumulative distribution function (CDF)** of a continuous random variable $X$ is,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u) du \quad \text{for} \quad -\infty < x < \infty \quad (4-3)$$

One can also use the **inverse cumulative distribution function** or **complementary cumulative distribution function (CCDF)**

$$F_{>}(x) = P(X > x) = \int_{x}^{\infty} f(u) du \quad \text{for} \quad -\infty < x < \infty$$

Definition of CDF for a continuous variable is the same as for a discrete variable.
Density vs. Cumulative Functions

• The probability density function (PDF) is the derivative of the cumulative distribution function (CDF).

\[ f(x) = \frac{dF(x)}{dx} = \frac{dF_>(x)}{dx} \]

as long as the derivative exists.
Mean & Variance

Suppose $X$ is a continuous random variable with probability density function $f(x)$. The mean or expected value of $X$, denoted as $\mu$ or $E(X)$, is

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x) \, dx$$

(4-4)

The variance of $X$, denoted as $V(X)$ or $\sigma^2$, is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx = \int_{-\infty}^{\infty} x^2 f(x) \, dx - \mu^2$$

The standard deviation of $X$ is $\sigma = \sqrt{\sigma^2}$. 
Gallery of Useful Continuous Probability Distributions
Continuous Uniform Distribution

• This is the simplest continuous distribution and analogous to its discrete counterpart.

• A continuous random variable $X$ with probability density function

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$  \hspace{1cm} (4-6)

Compare to discrete

$$f(x) = \frac{1}{b-a+1}$$

Figure 4-8 Continuous uniform PDF
Comparison between Discrete & Continuous Uniform Distributions

Discrete:
- PMF: $f(x) = 1/(b-a+1)$
- Mean and Variance:
  $\mu = E(x) = (b+a)/2$
  $\sigma^2 = V(x) = [(b-a+1)^2-1]/12$

Continuous:
- PMF: $f(x) = 1/(b-a)$
- Mean and Variance:
  $\mu = E(x) = (b+a)/2$
  $\sigma^2 = V(x) = (b-a)^2/12$
X is a **continuous** random variable with a uniform distribution between 0 and 3.

What is $P(X=1)$?

A. $\frac{1}{4}$
B. $\frac{1}{3}$
C. 0
D. Infinity
E. I have no idea

Get your i-clickers
X is a **continuous** random variable with a uniform distribution between 0 and 3.

What is $P(X=1)$?

A. $\frac{1}{4}$  
B. $\frac{1}{3}$  
C. 0  
D. Infinity  
E. I have no idea

Get your i-clickers
X is a **continuous** random variable with a uniform distribution between 0 and 3.

What is \( P(X<1) \)?

A. 1/4  
B. 1/3  
C. 0  
D. Infinity  
E. I have no idea

Get your i-clickers
X is a **continuous** random variable with a uniform distribution between 0 and 3.

**What is P(X<1)?**

A. 1/4  
B. 1/3  
C. 0  
D. Infinity  
E. I have no idea

Get your i-clickers
Constant rate (Poisson) process
Constant rate (Poisson) process

Discrete events happen at rate $\Gamma$
Expected number of events in time $x$ is $\Gamma x$

The actual number of events $N_x$ is a Poisson distributed discrete random variable

$$\mathbb{P}(N = n) = \frac{(\Gamma x)^n}{n!} e^{-\Gamma x}$$

Why Poisson? Divide $x$ into many tiny intervals of length $\Delta x$

$$p = \frac{\Gamma x}{\Delta x} \quad \text{Prob}(N = n) = \left( \frac{L}{n} \right)^n (1-p)^{L-n}$$

$$E(N_x) = pL = \Gamma x$$

$\text{Poisson}$
Constant rate (AKA Poisson) processes

- Let’s assume that proteins are produced by ribosomes in the cell at a rate $r$ per second.
- The expected number of proteins produced in $x$ seconds is $r \cdot x$.
- The actual number of proteins $N_x$ is a discrete random variable following a Poisson distribution with mean $r \cdot x$:
  \[
  P_{N}(N_x=n)=\exp(-r\cdot x)(r\cdot x)^n/n! \quad E(N_x)= r x
  \]
- Why Discrete Poisson Distribution?
  - Divide time into many tiny intervals of length $\Delta x \ll 1/r$
  - The probability of success (protein production) per internal is small: $p_{\text{success}}=r\Delta x \ll 1$,
  - The number of intervals is large: $n= x/\Delta x >> 1$
  - Mean is constant: $r=E(N_x)=p_{\text{success}} \cdot n= r\Delta x \cdot x/\Delta x = r \cdot x$
  - In the limit $\Delta x \ll x$, $p_{\text{success}}$ is small and $n$ is large, thus Binomial distribution $\rightarrow$ Poisson distribution
Exponential Distribution Definition

Exponential random variable $X$ describes interval between two successes of a constant rate (Poisson) random process with success rate $r$ per unit interval.

The probability density function of $X$ is:

$$f(x) = re^{-rx} \text{ for } 0 \leq x < \infty$$

Closely related to the discrete geometric distribution

$$f(x) = p(1-p)^{x-1} = p e^{(x-1) \ln(1-p)} \approx pe^{-px} \text{ for small } p$$
What is the interval $X$ between two successes of a constant rate process?

- $X$ is a **continuous** random variable
- **CCDF:** $P_X(X>x) = P_N(N_X=0) = \exp(-r\cdot x)$.
  - Remember: $P_N(N_X=n) = \exp(-r\cdot x) \frac{(r\cdot x)^n}{n!}$
- **PDF:** $f_X(x) = -dCCDF_x(x)/dx = r \cdot \exp(-r\cdot x)$
- We started with a discrete Poisson distribution where time $x$ was a parameter
- We ended up with a **continuous exponential** distribution
To summarize constant rate processes:

- $r$ - rate per unit of length

- $N(x)$ - discrete number of events in time $x$

Poisson: $P(N(x)=n) = \frac{(r \cdot x)^n}{n!} e^{-r \cdot x}$

Time interval $X$ between successive events is a continuously distributed random variable.

Its PDF if $f(x) = e^{-r \cdot x}$
Exponential Mean & Variance

If the random variable $X$ has an exponential distribution with rate $r$,

$$
\mu = E(X) = \frac{1}{r} \quad \text{and} \quad \sigma^2 = V(X) = \frac{1}{r^2} \quad (4-15)
$$

Note that, for the:

- Poisson distribution: mean = variance
- Exponential distribution: mean = standard deviation = variance^{0.5}
Biochemical Reaction Time

• The time $x$ (in minutes) until an enzyme catalyzes a biochemical reaction and generates a product is approximated by this CDF:

$$ F(x) = 1 - e^{-x/1.4} \text{ for } 0 \leq x $$

Here 1.4 min is average time between successive products $\rightarrow$ rate of this process is $r=1/1.4$ min$^{-1}$

• What is the PDF?

$$ f(x) = \frac{dF(x)}{dx} = \frac{d}{dx}[1 - e^{-x/1.4}] = e^{-x/1.4}/1.4 \text{ for } 0 \leq x $$

• What proportion of reactions is complete within 0.5 minutes?

$$ P(X < 0.5) = F(0.5) = 1 - e^{-0.5 \text{ min}/1.4 \text{ min}} = 1 - 0.7 = 0.3 $$