Continuous Probability Distributions

Uniform Distribution
R2 is a discrete random variable taking values 0,1,2,3. I generated a sample of 100,000 random values and stored it in the array r2 of size 100,000x1.

Then I computed the histogram: [counts, bins]=hist(r2,0:3)

Let f(x) be the PMF of R2.

Which pair of variables best approximates f(x):

A. x = counts, f(x)=bins./counts
B. x = counts, f(x)=bins./sum(bins)
C. x=bins./sum(bins), f(x)=counts
D. x=bins, f(x)=counts./sum(counts)
E. I have no idea

Get your i-clickers
R2 is a discrete random variable taking values 0,1,2,3. I generated a sample of 100,000 random values and stored it in the array r2 of size 100,000x1. Then I computed the histogram: `[counts, bins]=hist(r2,0:3)`.

Let $f(x)$ be the PMF of R2. Which pair of variables best approximates $f(x)$?

A. $x = \text{counts}$, $f(x) = \text{bins}./\text{counts}$
B. $x = \text{counts}$, $f(x) = \text{bins}./\text{sum(bins)}$
C. $x = \text{bins}./\text{sum(bins)}$, $f(x) = \text{counts}$
D. $x = \text{bins}$, $f(x) = \text{counts}./\text{sum(counts)}$
E. I have no idea

Get your i-clickers
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>$P(X=x)$</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 a 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 \( f(x) \):

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

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

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

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

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

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

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

$$
\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx \quad (4-4)
$$

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: \( f(x) = \frac{1}{(b-a+1)} \)
- Mean and Variance:
  \[ \mu = E(x) = \frac{(b+a)}{2} \]
  \[ \sigma^2 = V(x) = \frac{((b-a+1)^2-1)}{12} \]

Continuous:

- PMF: \( f(x) = \frac{1}{(b-a)} \)
- Mean and Variance:
  \[ \mu = E(x) = \frac{(b+a)}{2} \]
  \[ \sigma^2 = V(x) = \frac{(b-a)^2}{12} \]
X is a **continuous** random variable with a uniform distribution between 0 and 3.  
What is Probability(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
X is a \textcolor{red}{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
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.

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

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

$$p = \frac{\Gamma \Delta x}{L} = x/\Delta x$$

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

$\text{Poisson}$
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 \( r \cdot x \):

\[
P_N(N_x=n) = \exp(-r \cdot x) (r \cdot x)^n / n! \quad \text{E}(N_x) = r x
\]

• 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: \( n = x / \Delta x \gg 1 \)
  – Mean is constant: \( r = \text{E}(N_x) = p_{\text{success}} \cdot n = r \Delta x \cdot x / \Delta x = r \cdot x \)
  – 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

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

- $X$ is a continuous random variable
- CCDF: $P_X(X>x) = P_{N_X=0} = \exp(-r \cdot x)$.
  - Remember: $P_{N_X=n} = \exp(-r \cdot x) \frac{(r \cdot x)^n}{n!}$
- 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: 

\[ \eta \text{ - rate per unit of length} \]

\[ N(x) \text{ - discrete number of events in time } x \]

Poisson: \[ P(N(x) = n) = \frac{(\eta \cdot x)^n}{n!} e^{-\eta \cdot x} \]

Time interval \( X \) between successive events is a continuously distributed random variable.

Its PDF if \[ f(x) = e^{-\eta x} \]
Exponential Mean & Variance

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

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

(4–15)

Note that, for the:

• Poisson distribution: $\text{mean} = \text{variance}$
• Exponential distribution: $\text{mean} = \text{standard deviation} = \text{variance}^{0.5}$
Biochemical Reaction Time

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

\[
F_>(x) = e^{-2x} \text{ for } 0 \leq x
\]

Here the rate of this process is \( r=2 \text{ min}^{-1} \) and \( 1/r=0.5 \text{ min} \) is the average time between successive products of this enzyme

• What is the PDF?

\[
f(x) = -\frac{dF_>(x)}{dx} = -\frac{d}{dx} e^{-2x} = 2e^{-2x} \text{ for } 0 \leq x
\]

• What proportion of reactions will not generate another product within 0.5 minutes of the previous product?

\[
P(X > 0.5) = F_>(0.5) = e^{-2 \times 0.5} = 0.37
\]
We observed our enzyme for 1 minute and no product has been generated: The product is “overdue”

What is the probability that a product will not appear during the next 0.5 minutes?

A. 0.32
B. 0.37
C. 0.08
D. 0.24
E. I have no idea

Get your i-clickers
Memoryless property of the exponential distribution

\[ P(X > t+s | X > s) = P(X > t) \]

\[ P(X > t+s | X > s) = \frac{P(X > t+s, X > s)}{P(X > s)} = \frac{P(X > t+s)}{P(X > s)} = \frac{\exp(-\lambda s)}{\exp(-\lambda s)} = \exp(-\lambda t) = \]

\[ P(X > t) \]

Exponential is the only memoryless distribution