Probabilistic Modeling

Today

- Discrete random variables
- Continuous random variables
- P.d.f.'s and c.d.f.'s
- Mean and variance
- Dependence and independence; joint and marginal probabilities

What is/why probabilistic modeling?

What is a random variable?

- Something that has not happened yet.
  - Does a tossed coin come up heads or tails?
  - Does the cancer recur or not?
- Something you do not know . . .
  - Did a tossed coin come up heads of tails?
  - Is X a transcription factor for gene Y?
  - How does the protein fold?

  . . . because you have not/cannot observe it directly or compute it definitively from what you have observed.

Discrete random variables

Examples

A discrete r.v. X takes values from a discrete set Ω_X.

- $X =$ result of a coin toss; $\Omega_X = \{\text{Head,Tail}\}$.
- $X =$ roll of a die; $\Omega_X = \{1, 2, 3, 4, 5, 6\}$.
- $X =$ nucleotide a position 1, chromosome 1, in a particular person; $\Omega_X = \{A, C, G, T\}$.
- $X =$ amino acid 12 in a particular person's hemoglobin; $\Omega_X = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$.
- $X =$ copy number of gene Z in a particular person; $\Omega_X = \{0, 1, 2, 3, \ldots\}$.

Probabilities

- For a discrete r.v. X, each value $x \in \Omega_X$ has a probability of occurring, denoted variously by

$$\begin{array}{ccc} \mathsf{Prob}(X = x) & \mathsf{Prob}_X(x) & \mathsf{Prob}(x) \\ \mathsf{Pr}(X = x) & \mathsf{Pr}_X(x) & \mathsf{Pr}(x) \\ \mathsf{P}(X = x) & \mathsf{P}_X(x) & \mathsf{P}(x) \end{array}$$

- $0 \leq \mathsf{P}(x) \leq 1$
- $\sum_{x \in \Omega_X} \mathsf{P}(x) = 1$
- $\mathsf{P}(X)$ denotes the *probability distribution function* for r.v. X. It can be thought of as a table.

x	A	C	G	T
$\mathsf{P}(x)$	0	0.2	0.7	0.1

Cumulative distribution functions

- If X takes values from an ordered set Ω_X (such as integers) then the *cumulative distribution function* is

$$\text{c.d.f.}(x) = \mathsf{P}(X \leq x) = \sum_{x' \leq x} \mathsf{P}(x)$$

- For example, if X is the roll of a die, then:

x	1	2	3	4	5	6
$\mathsf{P}(x)$	1/6	1/6	1/6	1/6	1/6	1/6
c.d.f.(x)	1/6	2/6	3/6	4/6	5/6	1

Mean and variance

- If Ω_X is a set of numbers, then the expected value of X is

$$\mathsf{E}(X) = \sum_{x \in \Omega_X} x\mathsf{P}(x)$$

- The variance of X is

$$\begin{aligned} \mathsf{Var}(X) &= \mathsf{E}(X^2) - (\mathsf{E}(X))^2 \\ &= \left(\sum_x x^2 \mathsf{P}(x)\right) - \left(\sum_x x\mathsf{P}(x)\right)^2 \\ &\geq 0 \end{aligned}$$

- Example: If X is a die roll, then the mean value is 3.5 and the standard deviation is approximately 3.4157.

Continuous random variables

Examples

A continuous r.v. X takes real values.

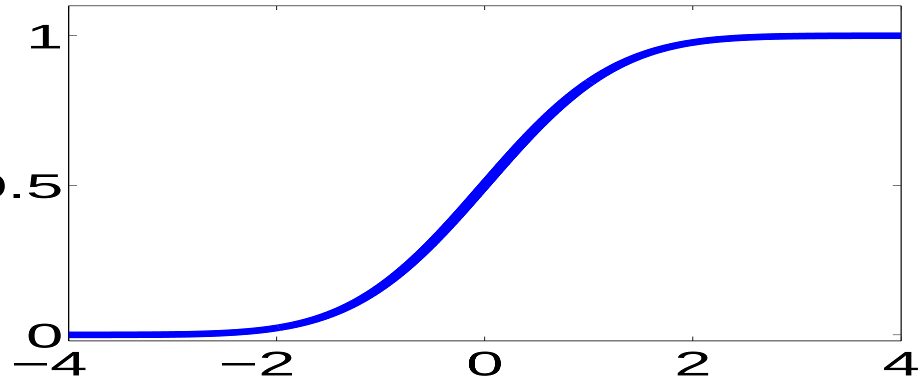
- $X =$ expression level for a gene as reported by a microarray.
- $X =$ time until a patient's cancer recurs.
- $X =$ size of a tumor.
- $X =$ mass of a peptide as reported by mass-spec.
- $X =$ binding energy between a TF and DNA. (?)
- $X =$ fraction of time a TF is bound to DNA.

Cumulative distribution functions

- Any continuous r.v. X has a *cumulative distribution function*

$$\text{c.d.f.}(x) = \mathsf{P}(X \leq x)$$

- c.d.f.(x) is a non-decreasing function; c.d.f.$(x) \leq$ c.d.f.(x') whenever $x \leq x'$.
- $\lim_{x \to -\infty}$ c.d.f.$(x) = 0$.
- $\lim_{x \to +\infty}$ c.d.f.$(x) = 1$.
- Example: The c.d.f. of a mean-zero, variance-one Gaussian r.v.:



Probability density functions

- If c.d.f.(x) is continuous and differentiable (at least, in most places) then it's derivative is the *probability density function*, analogous to the probability *distribution* function of a discrete r.v.

$$\frac{d}{dx}\text{c.d.f.}(x) = \text{p.d.f.}(x) = \mathsf{P}(x)$$

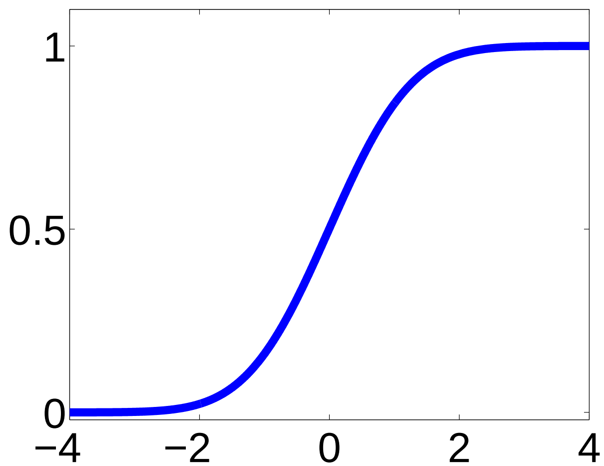
- $\mathsf{P}(x)$ is the "probability", or more properly, likelihood that X takes value x.
- $0 \leq \mathsf{P}(x) < \infty$. Observe that $\mathsf{P}(x) > 1$ is allowed, unlike for discrete r.v.'s.
- $\int_x \mathsf{P}(x)dx = 1$, similar to discrete r.v.'s.

Gaussian random variables

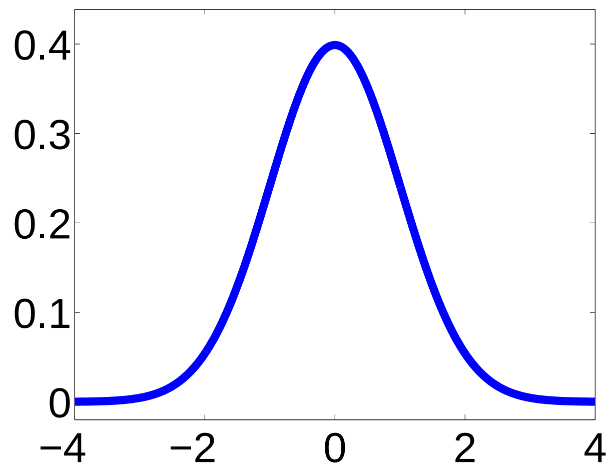
$X \sim N(\mu, \sigma)$ has mean μ and standard deviation σ.

graph

c.d.f.



p.d.f.



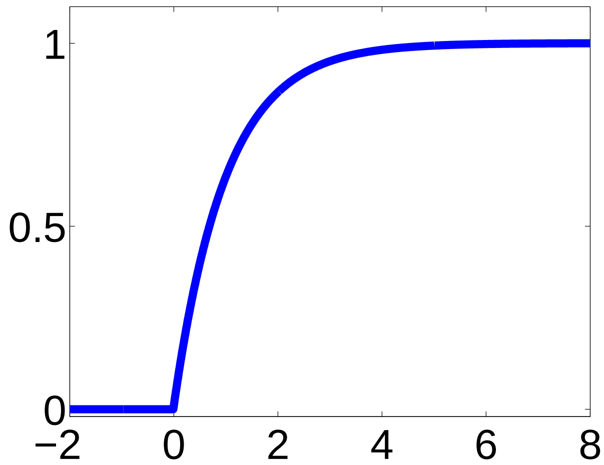
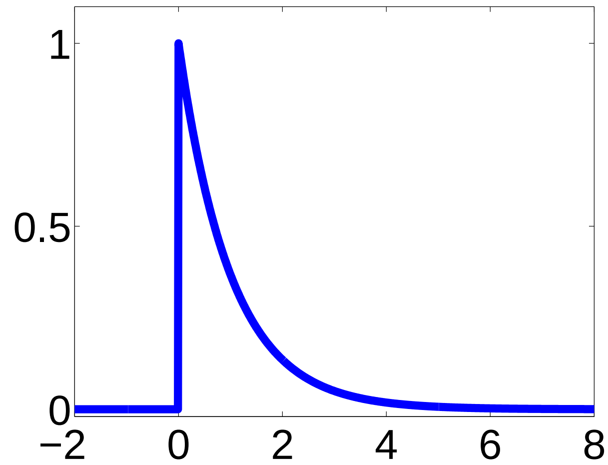
formula

?

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

Exponential random variables

$X \sim$ Exp(λ) has mean $1/\lambda$ and standard deviation $1/\lambda$.

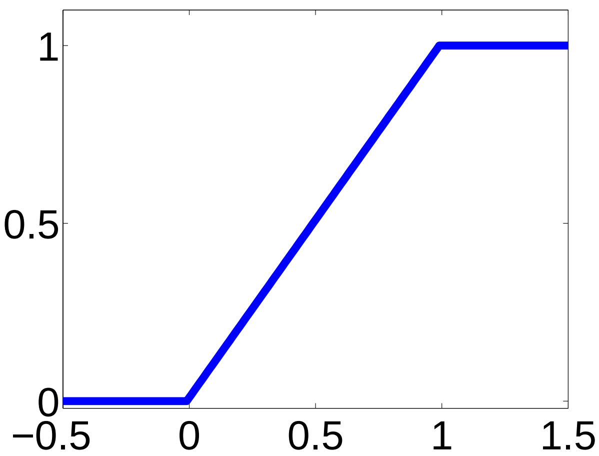
	c.d.f.	p.d.f.
graph		
formula	$1 - e^{-\lambda x}$	$\lambda e^{-\lambda x}$

Uniform random variables

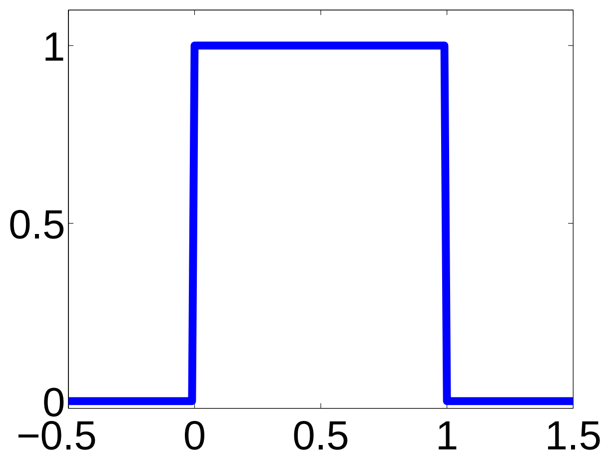
$X \sim U(a, b)$ has mean $\frac{a+b}{2}$ and standard deviation $\frac{(b-a)}{\sqrt{12}}$.

graph

c.d.f.



p.d.f.



formula

$$\begin{cases} 1 & x \geq b \\ \frac{x-a}{b-a} & a < x < b \\ 0 & x \leq a \end{cases}$$

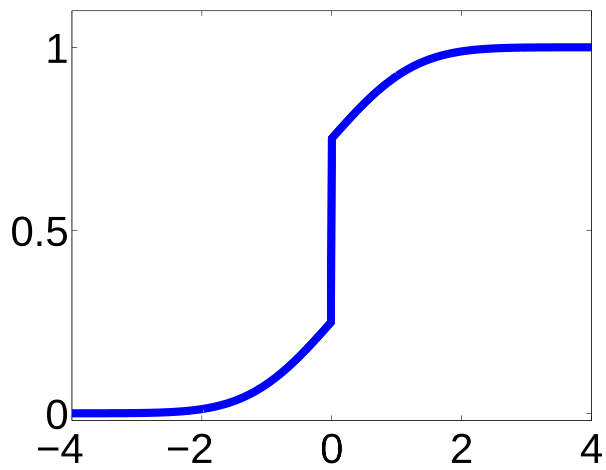
$$\begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

A continuous r.v. with no p.d.f.

- Suppose X equal to zero with probability $\frac{1}{2}$ and otherwise is distributed according to $N(0,1)$.
- Then the c.d.f. is

$$\text{c.d.f.}(x) = \begin{cases} \frac{1}{2} f(x) & x < 0 \\ \frac{1}{2} f(x) + \frac{1}{2} & x \geq 0 \end{cases}$$

where $f(x)$ denotes the c.d.f. of a $N(0,1)$ r.v.



- There is no p.d.f. because of the discrete jump in the c.d.f.

Mean and variance

- We will almost always restrict attention to continuous r.v.'s with p.d.f.'s.
- Then, the expected value is defined as

$$\mathsf{E}(X) = \int_x x\mathsf{P}(x)dx$$

- Variance is

$$\begin{aligned}\mathsf{Var}(X) &= \mathsf{E}(X^2) - (\mathsf{E}(X))^2 \\ &= \int_x x^2\mathsf{P}(x)dx - \left(\int_x x\mathsf{P}(x)dx\right)^2 \\ &\geq 0\end{aligned}$$

[In]dependent random variables

Example

- Let $X_1 =$ true iff a rolled die comes out even.
- Let $X_2 =$ true iff the same rolled die comes out odd.

$$\mathsf{P}(X_1 = \text{true}) = \mathsf{P}(X_1 = \text{false}) = \frac{1}{2}$$

$$\mathsf{P}(X_2 = \text{true}) = \mathsf{P}(X_2 = \text{false}) = \frac{1}{2}$$

- What is the probability $\mathsf{P}(X_1 = \text{true and } X_2 = \text{true})$?

Example

- Let $X_1 =$ true iff a rolled die comes out even.
- Let $X_2 =$ true iff the same rolled die comes out odd.

$$\mathsf{P}(X_1 = \text{true}) = \mathsf{P}(X_1 = \text{false}) = \frac{1}{2}$$

$$\mathsf{P}(X_2 = \text{true}) = \mathsf{P}(X_2 = \text{false}) = \frac{1}{2}$$

- What is the probability $\mathsf{P}(X_1 = \text{true and } X_2 = \text{true})$?
- We know it is zero.
- But there is no way of knowing just from $\mathsf{P}(X_1)$ and $\mathsf{P}(X_2)$.

$\Rightarrow$ There are several ways we can specify the relationships between variables. They all come down to specifying *joint probability distributions/densities*.

Joint probabilities

- When considering r.v.'s $X_1, X_2, \ldots, X_m$, the joint probability function specifies the probability of every combination of values.

$$\mathrm{P}(X_1 = x_1 \text{ and } X_2 = x_2 \text{ and } \ldots \text{ and } X_m = x_m)$$

- When the r.v.'s are discrete, the joint probability can be viewed as a table.

	even=true	odd=true
odd=true	0	1/2
odd=false	1/2	0

	die=1	2	3	4	5	6
even=true	0	1/6	0	1/6	0	1/6
even=false	1/6	0	1/6	0	1/6	0

Marginal probabilities

Given r.v.'s $X_1, X_2, \ldots X_m$ with joint probability $\mathrm{P}(x_1, x_2, \ldots, x_m)$.

- The marginal probability of a r.v. X_i is

$$\mathrm{P}(X_i = x_i) = \sum_{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m} \mathrm{P}(x_1, x_2, \ldots, x_m)$$

- That is, you get the marginal probability by summing (or integrating) over all possible values of the other r.v.'s.

	die=1	2	3	4	5	6	P(even)
even=true	0	1/6	0	1/6	0	1/6	1/2
even=false	1/6	0	1/6	0	1/6	0	1/2
P(die)	1/6	1/6	1/6	1/6	1/6	1/6	

- Similarly for the marginal probability of a subset of the r.v.'s.

Independent r.v.'s

- Two r.v.'s X and Y are independent if and only if

$$\mathsf{P}(X = x \text{ and } Y = y) = \mathsf{P}(X = x)\mathsf{P}(Y = y)$$

  for all x and y.

- This is often abbreviated as $\mathsf{P}(X, Y) = \mathsf{P}(X)\mathsf{P}(Y)$.

Conditional probablity

- For two r.v.'s X and Y, $\mathsf{P}(X = x|Y = y)$ denote the probability that $X = x$ *given that* $Y = y$.
  - $\mathsf{P}(\text{die=1}|\text{odd} = \text{true}) = 1/3$.
  - $\mathsf{P}(\text{die=1}|\text{odd} = \text{false}) = 0$.
- The conditional probability can be defined (and computed) as

$$\mathsf{P}(x|y) = \frac{\mathsf{P}(x, y)}{\mathsf{P}(y)}$$

  as long as $\mathsf{P}(y) > 0$.

- This is sometimes used as

$$\mathsf{P}(x) = \sum_y \mathsf{P}(x, y) = \sum_y \mathsf{P}(x|y)\mathsf{P}(y)$$

Conditional probability (2)

Conditional probabilities are interesting because we often observe something and want to infer something/make a guess about something unobserved but related.

- P$($cancer recurs$|$tumor measurements$)$
- P$($TF binds$|$TF and DNA properties$)$
- P$($Gene expressed $> 1.3|$TF concentrations$)$

Bayes' Rule

(Or possibly Bayes's Rule.)

- Bayes' Rule: $\mathsf{P}(x|y) = \frac{\mathsf{P}(y|x)\mathsf{P}(x)}{\mathsf{P}(y)}$.
- E.g., suppose we know based on past data collected:

$$\mathsf{P}(\text{tumor measurements}|\text{cancer})$$

$$\mathsf{P}(\text{tumor measurements}|\text{not cancer})$$

$$\mathsf{P}(\text{cancer} \qquad\qquad \mathsf{P}(\text{not cancer})$$

$$\mathsf{P}(\text{cancer}|\text{tumor meas.}) = \frac{\mathsf{P}(\text{tumor meas.}|\text{cancer})\mathsf{P}(\text{cancer})}{\mathsf{P}(\text{tumor meas.})}$$

$$= \frac{\mathsf{P}(\text{tumor meas.}|\text{cancer})}{\mathsf{P}(\text{tumor meas.}|\text{cancer})\mathsf{P}(\text{cancer}) + \mathsf{P}(\text{tumor meas.}|\text{not cancer})\mathsf{P}(\text{not cancer})}$$