Discrete Probability Distributions

Random Variables

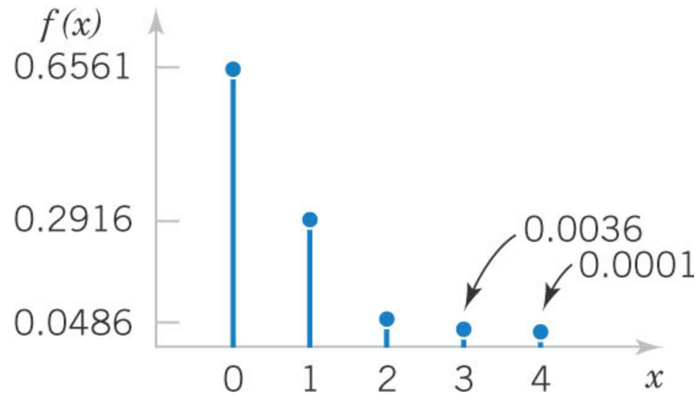
- A variable that associates a number with the outcome of a random experiment is called a random variable.
- Notation: random variable is denoted by an uppercase letter, such as X. After the experiment is conducted, the measured value is denoted by a lowercase letter, such a x. Both X and x are shown in italics, e.g., $P(X=x)$.

Continuous & Discrete Random Variables

- A discrete random variable is usually integer number
  - N - the number of p53 proteins in a cell
  - D - the number of nucleotides different between two sequences
- A continuous random variable is a real number
  - C=N/V – the concentration of p53 protein in a cell of volume V
  - Percentage (D/L)*100% of different nucleotides in protein sequences of different lengths L
    (depending on the set of L's may be discrete but dense)

Probability Mass Function (PMF)

- I want to compare all 4-mers in a pair of human genomes
- *X* – random variable: the number of nucleotide differences in a given 4-mer
- Probability Mass Function: f(x) or P(X=x) – the probability that the # of SNPs is exactly equal to x



Probability Mass Function for the # of mismatches in 4-mers

$P(X=0)$ =	0.6561
$P(X=1)$ =	0.2916
$P(X=2)$ =	0.0486
$P(X=3)$ =	0.0036
$P(X=4)$ =	0.0001
$\sum_x P(X=x)$=	1.0000

Cumulative Distribution Function (CDF)

	P(X≤x)	P(X>x)
	0.6561	0.3439
	0.9477	0.0523
	0.9963	0.0037
	0.9999	0.0001
	1.0000	0.0000

Cumulative Distribution Function CDF: $F(x)=P(X \leq x)$

Example:

$F(3)=P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) = 0.9999$

Complementary Cumulative Distribution Function (tail distribution) or CCDF: $F_{>}(x)=P(X>x)$

Example: $F_{>}(3)= P(X>3) = 1- P(X \leq 3) = 0.0001$

Mean or Expected Value of X

The mean or expected value of the discrete random variable X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \sum_x x \cdot P(X = x) = \sum_x x \cdot f(x)$$

- The mean = the weighted average of all possible values of *X*. It represents its "center of mass"

- The mean may, or may not, be an allowed value of X

- It is also called the arithmetic mean (to distinguish from e.g. the geometric mean discussed later)

- Mean may be infinite if X any integer and $P(X=x)>1/x^2$

Variance $V(X)$: square of a typical deviation from the mean $\mu = E(X)$

$V(X) = \sigma^2$, where σ is called standard deviation

$$\sigma^2 = V(X) = E\left((X-\mu)^2\right) =$$

$$= E(X^2 - 2\mu X + \mu^2) = E(X^2) -$$

$$- 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 =$$

$$= E(X^2) - \mu^2 = E(X^2) - (E(X))^2$$

Variance of a Random Variable

If X is a discrete random variable with probability mass function $f(x)$,

$$E\left[h(X)\right]=\sum_x h(x)\cdot P(X=x)=\sum_x h(x)f(x) \qquad (3\text{-}4)$$

If $h(x)=(X-\mu)^2$, then its expectation, $V(x)$, is the variance of X.

$\sigma=\sqrt{V(x)}$, is called standard deviation of X

$$\sigma^2=V(X)=\sum_x (x-\mu)^2 f(x) \text{ is the definitional formula}$$

$$=\sum_x\left(x^2-2\mu x+\mu^2\right)f(x)$$

$$=\sum_x x^2 f(x)-2\mu\sum xf(x)+\mu^2\sum_x f(x)$$

$$=\sum_x x^2 f(x)-2\mu^2+\mu^2$$

$$=\sum_x x^2 f(x)-\mu^2 \text{ is the computational formula}$$

Variance can be infinite if X can be any integer and $P(X=x) \geq 1/x^3$

Skewness of a random variable

- Want to quantify how asymmetric is the distribution around the mean?
- Need any odd moment: $E[(X-\mu)^{2n+1}]$
- Cannot do it with the first moment: $E[X-\mu]=0$
- Normalized 3-rd moment is skewness: $\gamma 1=E[(X-\mu)^3/\sigma^3]$
- Skewness can be infinite if X takes unbounded integer values and $P(X=x) \geq 1/x^4$

Geometric mean of a random variable

- Useful for very broad distributions (many orders of magnitude)?
- Mean may be dominated by very unlikely but very large events. Think of a lottery
- Exponent of the mean of *log X*: *Geometric mean=exp(E[log X])*
- Geometric mean usually is not infinite

Summary: Parameters of a Probability Distribution

- The mean, $\mu=E[X]$, is a measure of the center of mass of a random variable
- The variance, $V(X)=E[(X-\mu)^2]$, is a measure of the dispersion of a random variable around its mean
- The standard deviation, $\sigma=sqrt[V(X)]$, is another measure of the dispersion around mean
- The skewness, $\gamma 1=E[(X-\mu)^3/\sigma^3]$, measure of asymmetry around mean
- The geometric mean, $\exp(E[\log X])$, is useful for very broad distributions
- All can be infinite! Practically it means they increase with sample size
- <u>Different distributions can have identical parameters</u>

Summary: Parameters of a Probability Distribution

- Probability Mass Function (PMF): $f(x)=\text{Prob}(X=x)$
- Cumulative Distribution Function (CDF): $F(x)=\text{Prob}(X\leq x)$
- Complementary Cumulative Distribution Function (CCDF): $F_>(x)=\text{Prob}(X>x)$
- The mean, $\mu=E[X]$, is a measure of the center of mass of a random variable
- The variance, $V(X)=E[(X-\mu)^2]$, is a measure of the dispersion of a random variable around its mean
- The standard deviation, $\sigma=[V(X)]^{1/2}$, is another measure of the dispersion around mean. Has the same units as X
- The skewness, $\gamma 1=E[(X-\mu)^3/\sigma^3]$, a measure of asymmetry around mean
- The geometric mean, $\exp(E[\log X])$ is useful for very broad distributions

A gallery of useful
discrete probability distributions

Discrete Uniform Distribution

- Simplest discrete distribution.
- The random variable *X* assumes only a finite number of values, each with equal probability.
- A random variable *X* has a discrete uniform distribution if each of the n values in its range, say $x_1, x_2, \ldots, x_n$, has equal probability.

  $f(x_i) = 1/n$

Uniform Distribution of Consecutive Integers

- Let X be a discrete uniform random variable all integers from a to b (inclusive). There are $b - a + 1$ integers. Therefore each one gets:

$$f(x) = 1/(b-a+1)$$

- Its measures are:

$$\mu = E(x) = (b+a)/2$$

$$\sigma^2 = V(x) = [(b-a+1)^2 - 1]/12$$

Note that the mean is the midpoint of a & b.

A random variable X has the same probability for integer numbers

x =1:10

What is the behavior of its Probability Mass Function (PMF): P(X=x)?

A. does not change with x=1:10
B. linearly increases with x=1:10
C. linearly decreases with x=1:10
D. is a quadratic function of x=1:10

Get your i-clickers

A random variable X has the same probability for integer numbers

x =1:10

What is the behavior of its Cumulative Distribution Function (CDF): $P(X \leq x)$?

A. does not change with x=1:10
B. linearly increases with x=1:10
C. linearly decreases with x=1:10
D. is a quadratic function of x=1:10

Get your i-clickers

A random variable X has the same probability for integer numbers

x =1:10

What is its mean value?

A. 0.5
B. 5.5
C. 5
D. 0.1

Get your i-clickers

A random variable X has the same probability for integer numbers

$x = 1:10$

What is its skewness?

A. 0.5
B. 1
C. 0
D. 0.1

Get your i-clickers

Matlab exercise: Uniform distribution

- Generate a sample of size 100,000 for uniform random variable X taking values 1,2,3,…10
- Plot the approximation to the probability mass function based on this sample
- Calculate mean and variance of this sample and compare it to infinite sample predictions: $E[X]=(a+b)/2$ and $V[X]=((a-b+1)^2-1)/12$

Matlab template: Uniform distribution

- **b=10; a=1; % b= upper bound; a= lower bound (inclusive)'**
- **Stats=100000; % sample size to generate**
- **r1=rand(Stats,1);**
- **r2=floor(??*r1)+??;**
- **mean(r2)**
- **var(r2)**
- **std(r2)**
- **[hy,hx]=hist(r2, 1:10); % hist generates histogram in bins 1,2,3...,10**
- **% hy - number of counts in each bin; hx - coordinates of bins**
- **p_f=hy./??; % normalize counts to add up to 1**
- **figure; plot(??,p_f, 'ko-'); ylim([0, max(p_f)+0.01]); % plot the PMF**

Matlab exercise: Uniform distribution

- **b=10; a=1; % b= upper bound; a= lower bound (inclusive)'**
- **Stats=100000; % sample size to generate**
- **r1=rand(Stats,1);**
- **r2=floor(b*r1)+a;**
- **mean(r2)**
- **var(r2)**
- **std(r2)**
- **[hy,hx]=hist(r2, 1:10); % hist generates histogram in bins 1,2,3...,10**
- **% hy - number of counts in each bin; hx - coordinates of bins**
- **p_f=hy./sum(hy); % normalize counts to add up to 1**
- **figure; plot(hx,p_f, 'ko-'); ylim([0, max(p_f)+0.01]); % plot the PMF**