Lecture 2: Random Variables and Probability Distributions

May 3, 2012

GENOME 560, Spring 2012

Su-In Lee, CSE & GS

suinlee@uw.edu

Course Announcements

- A course mailing list has been created
  - genome560a_sp12@u.washington.edu
  - The registered students are already subscribed
- Problem Set 1 has been posted
  - Due next Thursday (5/10) before class
  - Please start as soon as possible
- Please go to the course website
  - http://www.cs.washington.edu/homes/suinlee/genome560
  - Check Announcements!

Outline

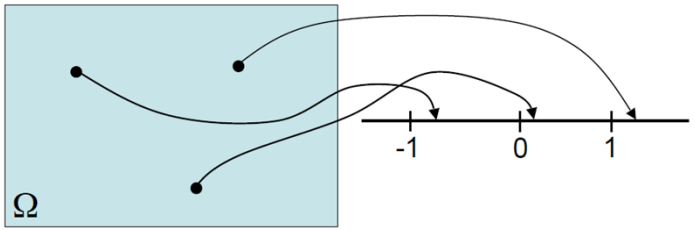
- Random variables
- Overview of probability distributions important in genetics and genomics
- R exercises
  - How to use R for calculating descriptive statistics and making graphs
  - Working with distributions in R

Random Variables (RV)

- A rv is a variable whose value results from the measurement of a quantity that is subject to variations due to chance (i.e. randomness).
  - e.g. dice throwing outcome, expression level of gene A
- More formally…

Random Variables (RV)

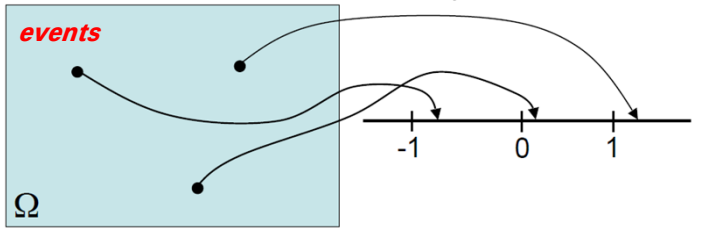
- A rv is any rule (i.e. function) that associates a number with each outcome in the sample space



5

What Does That Mean?

- Say that you throw a die
  - There are 6 possible outcomes (or ***events***)
  - Associate each event with a **number** ∈ {1,2,3,4,5,6}
  - A rv is what associates each dice throwing outcome with a number



- Let's consider an expression level of gene "A"
  - There are infinite number of **events**
  - Associate each event with a continuous-valued number representing expression level of gene A

6

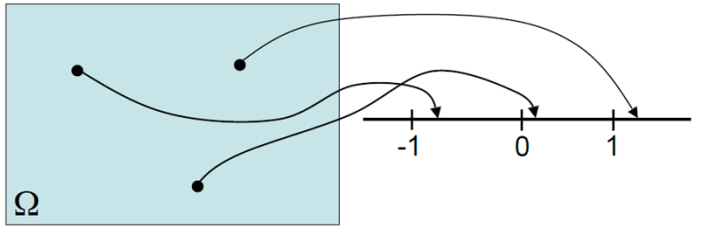
Two Types of Random Variables

- A **discrete** rv has a countable number of possible values
  - e.g. dice throwing outcome, genotype on a SNP, etc
- A **continuous** rv all values in an interval of numbers
  - e.g. expression level gene A, blood glucose level, etc

7

Random Variables (RV)

- A rv is any rule (i.e. function) that associates a number with each outcome in the sample space
- ***A rv associates each outcome with a probability…***

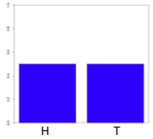


8

Probability Distribution

- **Discrete**
  - Let X be a discrete rv. Then the probability mass function (pmf), f(x), of X is:

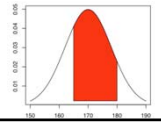
$$f(x) = \begin{cases} P(X = x), & x \in \Omega \\ 0, & x \notin \Omega \end{cases}$$



X = Coin toss outcome

- **Continuous**
  - Let X be a continuous rv. Then the probability density function (pdf) of X is a function f(x) such that for any two numbers a and b with a ≤ b

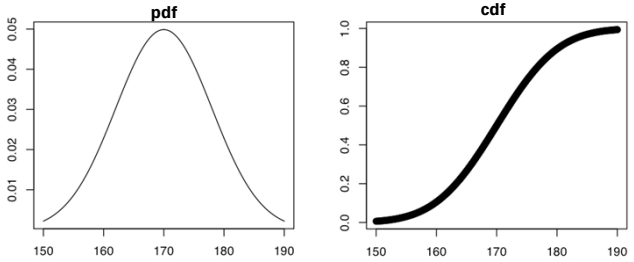
$$P(a \le X \le b) = \int_a^b f(x)dx$$



9

Cumulative Density Function

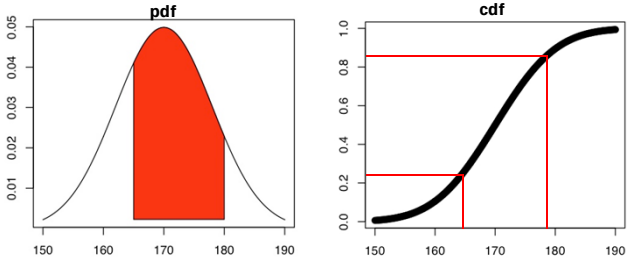
- Use CDFs to compute probabilities
- **Continuous rv:** $F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)dy$



10

Cumulative Density Function

- Use CDFs to compute probabilities
- **Continuous rv:** $F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)dy$



$$P(a \le X \le b) = F(b) - F(a)$$

11

Expectation of Random Variables

- **Discrete**
  - Let X be a discrete rv that takes on values in the set D and has a pmf f(x). Then the *expected* or *mean* value of X is:

$$\mu_X = E[X] = \sum_{x \in D} x \cdot f(x)$$

- For example, let's say that X is a rv representing the outcome of a die throw

  - X can be 1, 2, 3, 4, 5, or 6; so D = {1,2,3,4,5,6}
  - What is the expected value of X?
  - X = 1 with probability 1/6, X = 2 with prob. 1/6, X = 3 with prob. 1/6, …, X = 6 with prob. 1/6

12

Expectation of Random Variables

- **Discrete**
  - Let X be a discrete rv that takes on values in the set D and has a pmf f(x). Then the *expected* or *mean* value of X is:

$$\mu_X = E[X] = \sum_{x \in D} x \cdot f(x)$$

- **Continuous**
  - The expected or mean value of a continuous rv X with pdf f(x) is:

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

13

Variance of Random Variables

- **Discrete**
  - Let X be a discrete rv with pmf f(x) and expected value μ. The variance of X is:

$$\sigma_X^2 = V[X] = \sum_{x \in D} (x - \mu)^2 = E[(X - \mu)^2]$$

- **Continuous**
  - The variance of a continuous rv X with pdf f(x) and mean μ is:

$$\sigma_X^2 = V[X] = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx = E[(X - \mu)^2]$$

14

Example of Expectation and Variance

- Let $L_1, L_2, \ldots, L_n$ be a sequence of *n* nucleotides and define the rv X_i as:

$$X_i \begin{cases} 1, \text{ if } L_i = A \\ 0, \text{ otherwise} \end{cases}$$

- pmf is then: $P(X_i = 1) = P(L_i = A) = p_A$

  $P(X_i = 0) = P(L_i = C \text{ or } G \text{ or } T) = 1 - p_A$

- $E[X] = 1 \times p_A + 0 \times (1 - p_A) = p_A$

- $Var[X] = E[X - \mu]^2 = E[X^2] - \mu^2$

  $= [1^2 \times p_A + 0^2 \times (1 - p_A)] - p_A^2$

  $= p_A (1 - p_A)$

15

The Distributions We'll Study Today

- Binomial distribution
- Hypergenomtric Distribution
- Poisson Distribution
- Normal Distribution

16

Binomial Distribution

- Experiment consists of *n* trials
  - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- Trials are independent and identical (called *i.i.d*)
- Each trial can result in one of the two same outcomes
  - e.g., head or tail in each toss of a coin
  - Generally called "success" and "failure"
  - Probability of success is *p*, probability of failure is 1-*p*
- Constant probability for each observation
  - e.g., Probability of getting a trial is the same each time we toss the coin

17

Binomial Distribution: Example 1

- Let's say that we toss a coin *n* (=100) times
  - It is a biased coin; the chance of Head is 0.4
- A rv *X* represents the number of heads
  - What is the probability that *X* = *k*? What if *k* > 100 ?
- The probability of *k* particular tosses coming up Heads out of *n* tosses (say TTHTHHTTT…HT) is

$$(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p)\dots p(1-p) = p^k(1-p)^{n-k}$$

- There are $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$ different ways to choose *k* coins to be Heads out of *n* tosses. Each of these choices is mutually exclusive, so we add up the above probability that many times, so the total probability of all ways of getting *k* Heads out of n tosses is

$$\binom{n}{k} p^k(1-p)^{n-k} = \frac{n!}{k!\,(n-k)!} p^k(1-p)^{n-k}$$

18

Our Coin Example

- In our numerical example (*n* = 100, *p* = 0.4)
- Probability of *k* Heads is

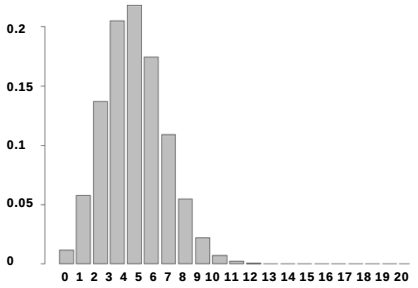
$$\binom{100}{48}(0.4)^{48}(1-0.4)^{100-48} = \frac{100!}{48!\,52!}0.4^{48}0.6^{52}$$
$$= 93{,}206{,}558{,}875{,}049{,}876{,}949{,}581{,}681{,}100 \times 7.92282 \times 10^{-20} \times 2.90981 \times 10^{-12}$$
$$= 0.0214878$$

(which really is best done with a computer and/or logarithms!)

19

The Histogram of a Binomial Dist.

- This is for *n* = 20 and *p*=0.2



20

Binomial Distribution

- **pmf:**

$$P\{X = x\} = \binom{n}{x} p^x (1-p)^{n-x}$$

- **cdf:**

$$P\{X \le x\} = \sum_{y=0}^{x} \binom{n}{y} p^y (1-p)^{n-y}$$

- **E(x) = np**
- **Var(x) = np(1-p)**

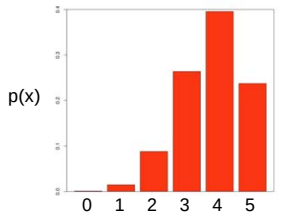
21

Binomial Distribution: Example 2

- A couple, who are both carriers for a recessive disease, wish to have 5 children. They want to know the probability that they will have four healthy kids
  - What is *p* and *n*?

$$P\{X = 4\} = \binom{5}{4} 0.75^4 \times 0.25^1$$

$$= 0.395$$



22

Binomial Distribution: Example 3

- Wright-Fisher model: There are *i* copies of the A allele in a population of size 2*N* in generation *t*. What is the distribution of the number of A alleles in generation (t+1) ?
  - What is *p* and *n*?
  - The probability of *j* copies of A allele in generation (t+1) is

$$p_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \quad j = 0, 1, \ldots, 2N$$

23

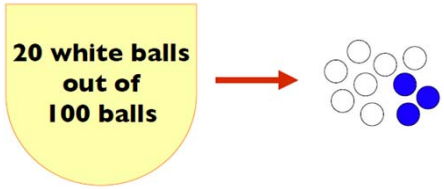
Hypergeometric Distribution

- Population to be sampled consists of *N* finite individuals, objects, or elements
- Each individual can be characterized as a success or failure, *m* successes in the population
- A sample of size *k* is drawn and the rv of interest is X = number of successes

24

Hypergeometric Distribution

- Similar in spirit to Binomial distribution, but from a **finite** population **without** replacement

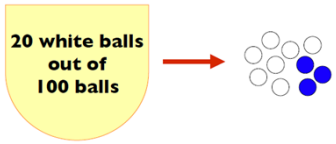


If we randomly sample 10 balls, what is the probability that 3 or less are blue?

25

Hypergeometric Distribution

- Say that we have an urn with *N* balls in it, *M* of which are blue (the rest are white). If we draw *n* balls out of it without replacement, what is the probability that *m* of those are blue?



- It turns out to be the fraction, out of the ways we could choose *n* balls out of *N*, in which there are *m* white and (*n*-*m*) blue balls:

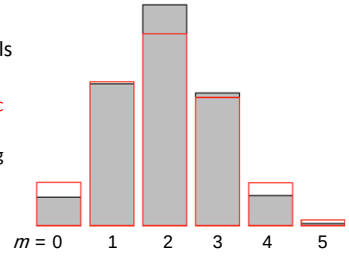
$$\frac{\binom{M}{m} \times \binom{N-M}{n-m}}{\binom{N}{n}} = \frac{M!\,(N-M)!\,n!\,(N-n)!}{N!\,m!\,(M-m)!\,(n-m)!\,\big(N-(n-m)\big)!}$$

26

Histogram of a Hypergeometric Dist.

- There are *N* (=20) balls in the urn, *M* (=8) of which are blue.
- If we draw *n* (=5) balls out of it without replacement, what is the probability that *m* of them are blue?

- Here are histograms showing the pmf of *m*, the number of blue balls (out of 5)
  - Gray boxes are the hypergeometric distribution
  - Red outlines are the corresponding binomial distribution
- What made them different?



27

Hypergeometric Distribution

- pmf of a hypergeometric rv:

$$P\{X = i \mid n,m,k\} = \frac{\binom{m}{i}\binom{n}{k-i}}{\binom{m+n}{k}}$$

Where,

k = Number of balls selected

m = Number of balls in urn considered "success"

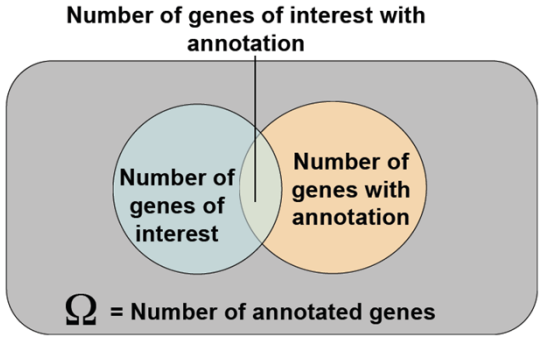
n = Number of balls in urn considered "failure"

m + n = Total number of balls in urn

28

Hypergeometric Distribution

- Extensively used in genomics to test for "enrichment":



29

Poisson Distribution

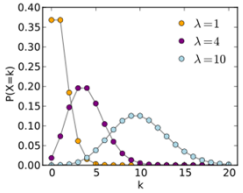
- It expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event
- Suppose someone typically gets 4 pieces of mail per day. That becomes the expectation, but there will be a certain spread: sometimes a little more, sometimes a little less, once in a while nothing at all.
- Given only the average rate, for a certain period of observation (e.g. pieces of mail per day), and assuming that the process that produce the event flow are essentially random, the Poisson distribution specifies how likely it is that the count will be 3, or 5, or 11, or any other number, during one period of observation (e.g. 1 week). That is, it predicts the degree of spread around a known average rate of occurrence.
- Poisson distribution approximates the binomial distribution when *n* (# trials) is large and *p* (change of success) is small

30

Poisson Distribution

- A rv *X* follows a Poisson distribution if the pmf of *X* is:

$$P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}$$ For i = 0, 1, 2, 3, …



- λ is frequently a rate per unit time:

  λ = αt = expected number of events per unit time t

- Safely approximates a binomial experiment when n > 100, p < 0.01, np = λ < 20)
- **E(X) = Var(X) = λ**

31

Poisson RV: Example 1

- The number of crossovers, X, between two markers is X ~ poisson (λ=d)

$$P\{X = i\} = e^{-d}\frac{d^i}{i!}$$

$$P\{X = 0\} = e^{-d}$$

$$P\{X \geq 1\} = 1 - e^{-d}$$

32

Poisson RV: Example 2

- Recent work in Drosophila suggests the spontaneous rate of deleterious mutations is ~ 1.2 per diploid genome. Thus, let's tentatively assume X ~ Poisson(λ = 1.2) for humans. What is the probability that an individual has 12 or more spontaneous deleterious mutations?

$$P\{X \geq 12\} = 1 - \sum_{i=0}^{11} e^{-1.2} \frac{1.2^i}{i!}$$

$$= 6.17 \times 10^{-9}$$

Poisson RV: Example 3

- Suppose that a rare disease has an incidence of 1 in 1000 people per year. Assuming that members of the population are affected independently, find the probability of *k* cases in a population of 10,000 (followed over 1 year) for *k*=0,1,2.

- The expected value (mean) = λ = 0.001 x 10,000 = 10

$$P(X = 0) = \frac{(10)^0 e^{-(10)}}{0!} = .0000454$$

$$P(X = 1) = \frac{(10)^1 e^{-(10)}}{1!} = .000454$$

$$P(X = 2) = \frac{(10)^2 e^{-(10)}}{2!} = .00227$$

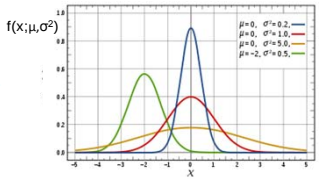
Normal Distribution

- "Most important" probability distribution
- Many rv's are approximately normally distribued
- Even when they aren't, their sums and averages often are Central Limit Theorem (CLT)

Normal Distribution

- pdf of normal distribution:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



- standard normal distribution (μ = 0, σ² = 1):

$$f(z;0,1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/2}$$

- cdf of Z

$$P(Z \leq z) = \int_{-\infty}^{z} f(y;0,1)\, dy$$

Standardizing Normal RV

- If X has a normal distribution with mean μ and standard deviation σ, we can standardize to a standard normal rv:

$$Z = \frac{X - \mu}{\sigma}$$

$$f(z;0,1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/2}$$

37

1 Digress: Sample Distributions

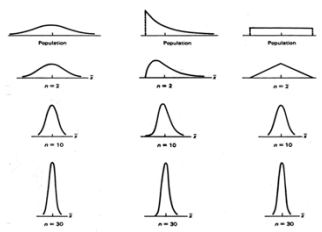
- Before data is collected, we regard observations as random variables X_1, X_2,…, X_n.
- This implies that until data is collected, any function (statistic) of the observations (mean, sd, etc) is also a random variable
- Thus, any statistic, because it is a random variable, has a probability distribution – referred to as a **sample distribution**
- Let's focus on the sampling distribution of the mean, $\overline{X}$

38

Behold The Power of the CLT

- Let X_1, X_2,…, X_n be an iid random sample from a distribution with mean μ and standard deviation σ. If n is sufficiently large:

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$



39

Example

- If the mean and standard deviation of serum iron values from healthy men are 120 and 15 mgs per 100ml, respectively, what is the probability that a random sample of 50 normal men will yield a mean between 115 and 125 mgs per 100ml?

First, calculate mean and sd to normalize (120 and 15/sqrt(50))

$$p(115 \le \bar{x} \le 125 = p\left(\frac{115-120}{2.12} \le \bar{x} \le \frac{125-120}{2.12}\right)$$

$$= p(-2.36 \le z \le 2.36)$$

$$= p(z \le 2.36) - p(z \le -2.36)$$

$$= 0.9909 - 0.0091$$

$$= 0.9818$$

40