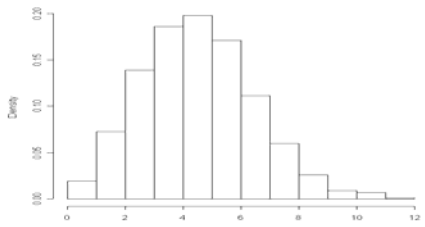
Statistics and Probability

Discrete Random Variables (Binomial, Poisson)



http://www.isrec.isb-sib.ch/~darlene/geneve/

Outline

- Review Bayes' Rule
- Random variables, density, cumulative distribution function
- Moments
- Discrete distributions
- Binomial distribution
- Poisson distribution

Bayes' Rule

For a *partition* $B_1, B_2, \ldots, B_n$ of all possible outcomes,

$$P(B_i \mid A) = P(A \mid B_i) * P(B_i) / P(A)$$

$$= \frac{P(A \mid B_i) * P(B_i)}{P(A \mid B_1) * P(B_1) + \ldots + P(A \mid B_n) * P(B_n)}$$

Huntington's disease

- Rare, autosomal dominant disorder which starts to appear in middle age (40s or 50s)
- Incidence in the *general population* is about 1/10,000, which means that the probability of carrying the defective gene is about .01%
- But for individuals with an affected parent, the probability of carrying the defective gene is 50%
- There is a test for HD, assume that the false positive rate is 2%, and the false negative rate is 1%

Huntington's disease (cont)

- Suppose a person with an affected parent takes the test
  - What is the chance the person is at risk if the test is negative?
  - What is the chance the person is at risk if the test is positive?

Huntington's disease (cont)

- Now suppose that a person without an affected parent is considering taking the test
  - What is the chance the person is at risk if the test is negative?
  - What is the chance the person is at risk if the test is positive?

Random variables (RVs)

- A *random variable* associates a numerical value with each outcome in the sample space
- RVs generally denoted by capital Roman letters: *X*, *Y*, *etc.*; particular values denoted by lower case Roman letters: *x*, *y*, *etc.*
- RVs can be *discrete* (dice) or *continuous* (height)

RV examples

- Simple examples:
  - Toss 2 dice, let *X* = sum of the numbers. Possible values for *X* are *…*
  - Choose a person at random, let *X* = height in cm. Possible values for *X* are *…*
  - *…*

Discrete or continuous??

- V = The volume of urine output per hour
- B = The amount of blood lost by a patient during the course of an operation
- S = The number of stop codons in a particular DNA sequence
- C = The number of children affected in a chicken pox epidemic
- L = The length of time that a leukemia patient's disease has been in remission

Two important functions

- Since the behavior of a RV is governed by *chance*, we cannot predict the outcome with certainty
- Instead, describe behavior of RV in terms of *probabilities*
- Two functions to accomplish this:
  - *(probability) density function*, or *pdf*, *f(x)*
  - *(cumulative) distribution function*, or *cdf*, *F(x)*

Density function for discrete RV

- The *density function* for a discrete RV *X* is

$$f(x) = P(X = x)$$

- Also sometimes called '*probability mass function*'
- The pdf for a discrete RV *X* satisfies:
  1. $f(x) \geq 0$ for all x
  2. $\sum_x f(x) = 1$

Distribution function for a RV

- The *(cumulative) distribution function* (cdf) for (any) RV *X* is

$$F(x) = P(X \leq x)$$

- The cdf satisfies:
  1. $F(x)$ is *nondecreasing* for all x
  2. $F(-\infty) = 0$
  3. $F(\infty) = 1$

Dice example

- For the sum of the numbers on 2 dice, let's work out *f(x), F(x)*

x	2	3	4	5	6	7	8	9	10	11	12
f(x)											
F(x)											

Expectation of a RV

- *Intuitive*: the *expected value* of a RV *X*, denoted *E[X]* or μ, is the long run theoretical average value of *X*
- Let's toss 2 dice… really! ☺
  1. x_1
  2. x_2
  3. x_3
  4. x_4
  5. x_5 …

Expectation of a RV, formal

- *Formal*: the *expected value* of a discrete RV *X* with density *f(x)* is

  $$E[X] = \sum_x x*f(x),$$

  i.e. the sum, over all possible values of *x*, of *x*P(X = x)*
- This is just a *weighted average* of all possible values of *x*, with weights *P(X = x)*
- *Example:* Toss 2 dice, *X* = sum (again)

  E[X] = 2(1/36) + 3*(2/36) + … + 12*(1/36)*

  = ??

More about expectation

- The *expectation* (*expected value*) of a RV means the same thing as *mean* or *average*
- The expected value of a RV *X*, or the *mean of its distribution*, can be regarded as the *center of gravity* of the distribution (*i.e.*, *balance point* of a histogram)
- The expected value of a *function* of a RV *g(X)* can be computed as

  $$E[g(X)] = \sum_x g(x)*f(x)$$

Variance

- The *variance* of a *discrete* RV *X* with density *f(x)* is

  $$Var(X) = \sigma^2 = E[(X - \mu)^2] = \sum_x (x - \mu)^2 * f(x)$$
- *Example:* those same 2 dice… *X* = sum

  $$Var(X) = E[(X - \mu)^2] = E[(X - 7)^2]$$
  $$= (2 - 7)^2 * (1/36) + (3 - 7)^2 * (2/36) + \ldots + (12 - 7)^2 * (1/36)$$
  $$\approx \underline{5.83}$$

Some properties of expectations

- *Linear transformation*: let *Y = aX + b*, where *X* is a RV and *a, b* are *constants* (i.e., **NOT** RVs). Then E[*Y*] = aE[*X*] + b, and Var(*Y*) = a^2 Var(*X*)
- *Sum of RVs*: let $Y = X_1 + X_2 + \ldots + X_n$. Then $E[Y] = E[X_1+X_2+\ldots+X_n] = E[X_1] + E[X_2] + \ldots + E[X_n]$
- If in addition $X_1, X_2, \ldots, X_n$ are *independent*, then $Var(Y) = Var(X_1) + Var(X_2) + \ldots + Var(X_n)$

Another example

- The following table shows the density for the RV X, the number of persons seeking emergency room treatment unnecessarily per day in a small hospital.
- Find f(5). What probability does this represent in the context of this problem?

x	0	1	2	3	4	5
f(x)	.01	.1	.4	.3	.1	??
F(x)	??	??	??	??	??	??

Example, contd

- Find $P(X \leq 2)$; Interpret this probability in the context of this problem
- Find $P(X < 2)$ and $P(X > 3)$
- Find $E(X)$ and $E([X - \mu]^2)$
- Find the expected value and variance of X
- Find $E(X^2) - E(X)^2$. How is this quantity related to the variance of X?

Bernoulli trials

- *Bernoulli trials* are a series of events with
  - *binary outcomes* (usually referred to as 'success' and 'failure')
  - *constant* probability of success *p*
  - *independence* of outcomes
- Examples:
  - Tossing a fair coin, getting Heads: *p* = ??
  - Tossing a fair die, getting 1: *p* = ??

Binomial distribution

- The distribution of the number of `successes' *X* in a
  1. *fixed number n* of
  2. *independent*
  3. *Bernoulli trials*, each with
  4. *constant* success probability *p*

  is called Binomial(*n, p*)
- *Examples (simple)*:
  - Number of Heads in 20 tosses of a fair coin
  - Number of 12's in 5 tosses of 2 dice

Binomial, approx binomial or not (1)

- In the RNA code, UGG codes tryptophan and UGA codes a stop. In a particular segment, the word UGA appears 5 times. Assume that nucleotides U and G will not mutate, but that nucleotide A (adenine) will mutate to G (guanine) 0.1% of the time. The number of mutations in the sequence in which the word *stop* (UGA) is mutated to tryptophan (UGG) is X.

Binomial, approx binomial or not (2)

- A biologist has 8 plants available for experimentation. The experiment calls for the use of only 4 plants. Unknown to the biologist, 3 of the plants are diseased. She randomly selects 4 plants to use in the experiment. Variable X is the number of diseased plants selected.

Binomial, approx binomial or not (3)

- In a study of the migratory habits of Canadian geese, approximately 5% of the entire population has been tagged. During a given day, 8 geese are captured. The number that are tagged is X.

Binomial, approx binomial or not (4)

- A couple is determined to have a daughter. They decide to continue having children until a daughter is born, at which time they will produce no more children. The number of children born before the birth of the first daughter is X.

(BREAK)

pdf for the Binomial distribution

- This is straightforward to figure out from first principles
- For $X \sim \text{Bin}(n, p)$,

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

Expected value and variance for binomial distribution

- If I tossed a fair coin 10 times, how many Heads would you expect? What if I tossed it 20 times? 100 times?
- It turns out that your intuitive guess is right... if $X \sim \text{Bin}(n, p)$, $E[X] = n*p$
- You can compute this directly from the definition of expected value
- The *variance* of a binomial RV is not so intuitive... it is $\text{Var}(X) = n*p*(1-p)$

Calculations for binomials

- Find the probability of getting 4 or more heads in 6 tosses of a fair coin

4 steps to solve problems with RVs

1. *Identify* the RV of interest
2. Determine the *distribution* of the RV
3. *Translate* the question
4. *Answer* the question

Using the 4 steps

- Let's solve the previous question ...

A more complicated problem

- What is the probability that among 5 families, each with 6 children, that at least 3 of the families have 4 or more girls?
- To solve this use the 4 steps, make any assumptions clear ...

Another example

- In a microbial mutagenesis assay, a plate of bacteria is exposed to a test compound, and the number of revertants is counted after incubation
- Say we want to find the probability that an assay of the compound will produce
  a. no revertants
  b. 3 revertants
  c. more than 5 revertants
- How could we answer this? We will need some assumptions ...

Another distribution

- A *Poisson process* is a probabilistic mechanism giving rise to the occurrence of events in a specific time interval (or region of space)
  - the probability of occurrence in an infinitesimally small area (or time interval) is λ*area (or interval length)
  - the probability that more than one event occurs in the small area/time is negligible
  - events over disjoint regions are independent
- The constant λ is the *rate parameter*

pdf for Poisson distributed RV

- For a RV $X \sim \text{Poisson}(\lambda)$,
  $f(x) = P(X = x) = e^{-\lambda} \lambda^x / x!$
- $E[X] = \lambda$
- $\text{Var}(X) = \lambda$

Solving the problem

- Say that we can assume that the number of revertants has a Poisson distribution with $\lambda = 9$
  a. $P(X = 0) = e^{-9}(9^0)/0! = .00012$
  b. $P(X = 3) =$??
  c. $P(X > 5) =$??

Example, cont

- Suppose now that we are interested in Y = number of assays (out of 5) with *no revertants*
- What is a reasonable probability model for Y?
- What is the chance of no revertants in exactly 2 of 5 assays?

Another example

- Some strains of paramecia produce and secrete 'killer' particles that will cause the death of a sensitive individual if contact is made. All paramecia unable to produce killer particles are sensitive. The mean number of killer particles emitted by a killer paramecium is 1 every 5 hours.
- What is the prob. that a killer paramecium would emit no such particles in 2 ½ hours??
- What is the probability that it would emit at least one killer particle??

Yet another example

- By damaging the chromosomes in the egg or sperm, mutations can be caused which lead to abortions, birth defects, or other genetic defects. Suppose that the probability that such a mutation is produced by radiation is 0.10.
- Of the next 150 mutations caused by damage to the chromosomes, how many would you expect to have been produced by radiation??
- What is the probability that exactly 10 were produced by radiation??