



BINF702 FALL 2008

Chapter 4 Discrete Probability Distributions



Section 4.1 - Introduction

- Ex. 4.1 – Retinitis pigmentosa is a progressive ocular disease that in some cases eventually results in blindness. The three main genetic types of the disease are the dominant, the recessive, and the sex-linked. Each genetic type has a different rate or progression, the dominant mode being the slowest to progress and the sex-linked mode the fastest. Suppose the prior history of disease in a family is unknown. However, 1 of 2 male children is affected, whereas 0 of 1 female children are affected. Can this information be used to identify the disease type?
 - We can use a particular type of discrete probability distribution, the binomial distribution to aid us in answering this question.



Section 4.2 – Random Variables

- Def. 4.1 – A random variable is a numeric function that assigns probabilities to different events in a sample space.
- Def. – A random variable for which there exists a discrete set of values with specified probabilities is a discrete random variable.
- Def. – A random variable whose possible values cannot be enumerated is a continuous random variable.
- Can you give me an example of a DRV?
- Can you give me an example of a CRV?



Section 4.3 – The Probability-Mass Function for a Discrete Random Variable

- Def. – A probability mass function is a mathematical relationship, or rule, that assigns to any possible value r of a discrete random variable X the probability $Pr(X = r)$. The assignment is made for all values r that have positive probability. The probability mass function is sometimes also called a probability distribution or a discrete probability distribution.
- Def. – Can you think of a probability mass function?



Section 4.3 – The Probability Mass Function for a Discrete Random Variable

- Def. – A properly normalized frequency distribution can be considered as an estimate to a probability mass function. There exist statistical tests to compare this estimate to the theoretical values as determined by the probability mass function. Once such test is the goodness of fit test which we will discuss in Chapter 10.



Section 4.4 – The Expected Value of a Discrete Random Variable

- De. 4.5 – The expected value of a discrete random variable X is defined as

$$E(X) = \mu = \sum_{i=1}^R x_i \Pr(X = x_i)$$

- Def. – Given X is the value obtained by a die upon a single toss then what is p.m.f. and $E(X)$?
- Def. – Given X is the value obtained by a coin upon a single toss where the event H is $X = 0$ and the event tails is $X = 1$ then what is p.m.f. and $E(X)$?
- Def. – How can you rationalize these fractional values with the integer values of the DRV?



Section 4.5 –The Variance of a Discrete Random Variable

- Def. 4.6 – The variance of a discrete random variable X , denoted by $\text{Var}(X)$, is defined by

$$\text{Var}(X) = \sigma^2 = \sum_{i=1}^R (x_i - \mu)^2 \Pr(X = x_i)$$

- Eq. 4.1 – A short form for the population variance is given by

$$\sigma^2 = E(X - \mu)^2 = \sum_{i=1}^R \Pr(X = x_i) - \mu^2$$



Section 4.5 – The Variance of a Discrete Random Variable

- Def. – Given X is the value obtained by a die upon a single toss then what is $Var(X)$?
- Def. – Given X is the value obtained by a coin upon a single toss where the event H is $X = 0$ and the event tails is $X = 1$ then what is $Var(X)$?



Section 4.5 – The Variance of a Discrete Random Variable

- Eq. 4.2 – Approximately 95% of the probability mass falls within two standard deviations (2σ) of the mean of a random variable.
- N.B. – This equation holds exactly when $X \sim N(\mu, \sigma^2)$ and $2s$ is replaced with 1.96σ .
- In R

```
> x = rnorm(1000)
> (sum((x > -1.96) & (x < 1.96)))/1000
[1] 0.955
```



Section 4.6 – The Cumulative-Distribution Function of a Discrete Random Variable

- De. 4.7 – The cumulative-distribution function (cdf) of a random variable X is denoted by $F(X)$ and, for a specific value x of X , is defined by

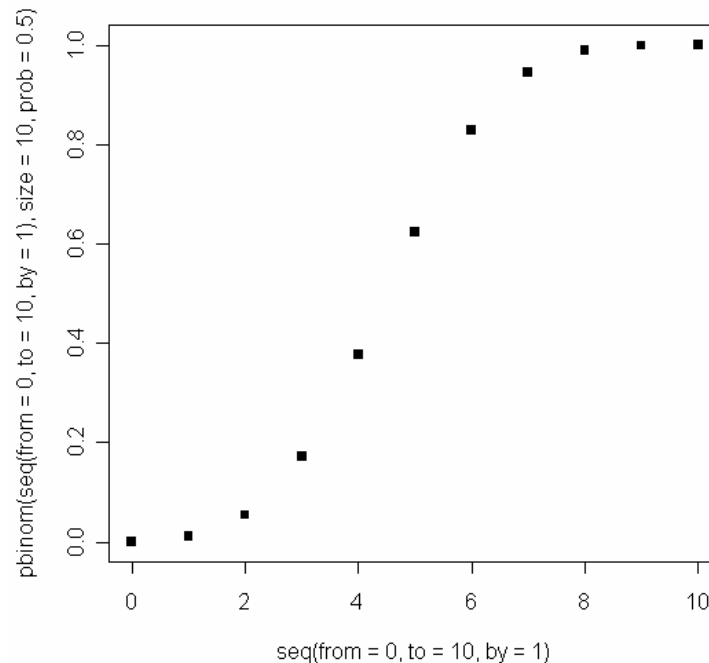
$$\Pr(X \leq x) = F(x)$$

- N.B. – The plot of a cdf for a discrete random variable looks like a step function while the cdf for a continuous random variable is a continuous curve.
- In the case where X is the value obtained during the roll of a die then what is $F(3)$?

Section 4.6 – The Cumulative-Distribution Function of a Discrete Random Variable

- In R

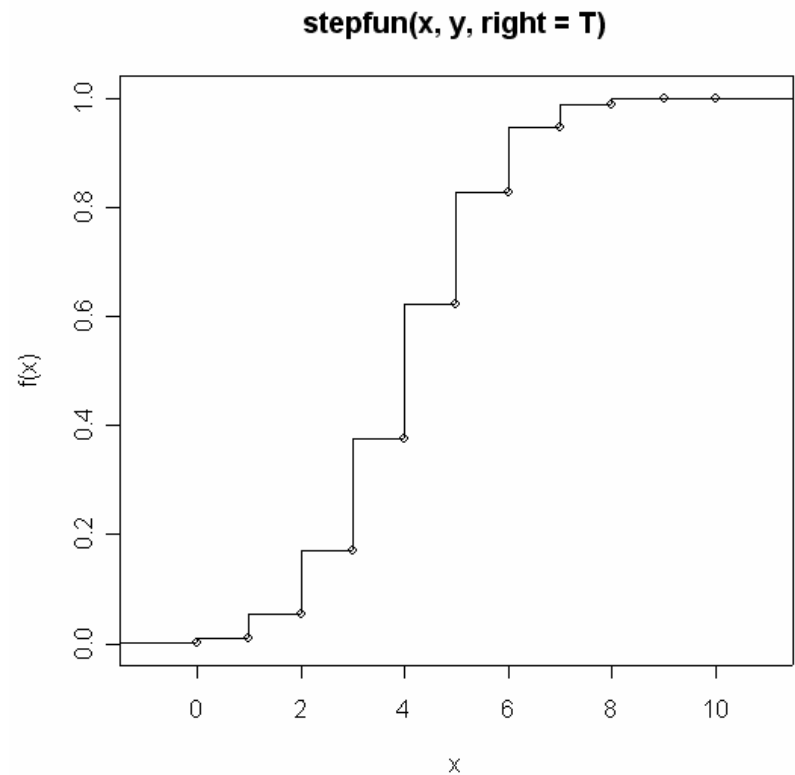
```
>plot(seq(from=0,to=10,by=1),pbinom(seq(from=0,to=10,by=1),size=10,prob=.5),pch=15)
```



- We will revisit this example later tonight.

Section 4.6 – The Cumulative-Distribution Function of a Discrete Random Variable

```
> x = seq(from=0,by=1,to=10)
> y = numeric(length=12)
> y[1:11]=pbinom(seq(from=0,to=10,by
  =1),size=10,prob=.5)
> y[12]=1
> plot(stepfun(x,y,right=T))
```





Section 4.7 – Permutations and Combinations

- Def. 4.8 – The number of permutations of n things taken k at a time is

$${}_n P_k = n(n-1)\cdots(n-k+1)$$

- N.B. – This represents the number of ways of selecting k items out of n , where the order of selection is important.
- Def. 4.9 – $n!$ = n factorial is defined as $n(n-1) \dots (2)(1)$
- N. B. – $0! = 1! = 1$



Section 4.7 – Permutations and Combinations

- Eq. 4.3 – Alternative Form for Permutations – An alternate formula expressing permutations in terms of factorials is given by

$${}_n P_k = \frac{n!}{(n-k)!}$$

- Def. 4.10 – The number of combinations of n things taken k at a time is given by

$${}_n C_k = \binom{n}{k} = \frac{(n)(n-1)\cdots(n-k+1)}{k!}$$



Section 4.7 – Permutations and Combinations

- Def. 4.11 – The number of combinations of n things taken k at a time is

$${}_n C_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- Eq. – 4.4 – For any nonnegative integers n, k , where $n \geq k$,

$$\binom{n}{k} = \binom{n}{n-k}$$



Section 4.8 – The Binomial Distribution

- Ex. 4.15 – (Infectious Disease) – One of the most common laboratory tests performed on a routine medical examination is a blood count. The two main aspects to a blood count are (1) counting the number of white blood cells (the “white count”) and (2) differentiating white blood cells that do exist into five categories – namely neutrophils, lymphocytes, monocytes, eosinophiles, and basophils (called the “differential”). Both the white counts and the differential are extensively used in making clinical diagnoses. We concentrate here on the differential, particularly on the distribution of the number of neutrophils k out of 5 white blood cells.

Section 4.8 – The Binomial Distribution

- **Ex. 4.23** – Given that the probability that any one cell is a neutrophil then what is the probability that the second and fifth cells considered will be neutrophils with the remaining nonneutrophils.

- $p = \text{event neutrophil, } q = \text{event nonneutrophil}$
 - $(q)(p)(q)(q)(p) = (.6)^2(.4)^3$

- **Ex. 4.24 (Infectious Disease)** What is the probability that any 2 cells out of 5 will be neutrophils?

- **XX000**
- **XOX00**
- **X0OX0**
- **X00OX**
- **OXX00**
- **OXOX0**
- **OX0OX**
- **00XX0**
- **00XOX**
- **000XX**

$$\binom{5}{2} (.6^2)(.4^3) = .230$$

ways probability of each way



Section 4.8 – The Binomial Distribution

- Eq. 4.5 – The distribution of the number of successes in n statistically independent trials, where the probability of success on each trial is p , is known as the binomial distribution and has a probability-mass function given by

$$\Pr(X = k) = \binom{n}{k} p^k q^{n-k}, k = 0, 1, \dots, n$$



Section 4.8 – The Binomial Distribution

- Using Binomial tables (Table 1 of the Appendix)
- Given $n = 10$, $p = .2$, and $k = 2$ then what is $\Pr(X = 2)$
- **XXXXX**
- When $p > .5$ we can use Table 1 via the following identity

$$\Pr(X = k) = \binom{n}{k} p^k q^{n-k} = \binom{n}{n-k} q^{n-k} p^k = \Pr(Y = n - k)$$

Where $Y \sim \text{Binom}(n, q)$



Section 4.8 – The Binomial Distribution

- How would one calculate $\Pr(X = 0)$ where $p = .6$ and $q = .4$
- Using the identity stated on the previous slide we obtain
- ????



The Binomial Distribution in R

Description:

Density, distribution function, quantile function and random generation for the binomial distribution with parameters 'size' and 'prob'.

Usage:

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```



```
dbinom(x, size, prob, log = FALSE)
```

- This computes $\Pr(X=x | X \sim \text{Binom}(\text{size}, \text{prob}))$

```
P(X=2 | X~Binom(n=10, p=.2))
```

```
> dbinom(x=2, size=10, prob=.2)
```

```
[1] 0.3019899
```



```
pbinom(q, size, prob, lower.tail =  
TRUE, log.p = FALSE)
```

- This computes $\Pr(X \leq q \mid X \sim \text{Binom}(\text{size}, \text{prob}))$
- Setting `lower.tail=FALSE` computes

Careful $\Pr(X > q \mid X \sim \text{Binom}(\text{size}, \text{prob}))$

- Consider Ex. 4.27.
 - Can you calculate this using the `lower.tail=FALSE` option
 - $\Pr(X \geq 3 \mid X \sim \text{binom}(20, .05))$



Ex. 4.27 (continued)

- Can you repeat your analysis using the `lower.tail=TRUE` option



Assessing Anomalous Significance

- Note that we assess the anomalous significance via
 - $\Pr(X \geq 3)$ rather than $\Pr(X = 3)$
- Consider $\Pr(X=75|X\sim\text{Binom}(1500,.05))$

- As compared to $\Pr(X\geq 75|X\sim\text{Binom}(1500,.05))$



```
qbinom(p, size, prob, lower.tail =  
TRUE, log.p = FALSE)
```

- This function is the inverse cdf it computes q such that

$$p = \Pr(X \leq q \mid X \sim \text{Binom}(\text{size}, \text{prob}))$$

```
> pbinom(q=2,prob=.05,size=20)  
[1] 0.9245163  
> qbinom(p=0.9245163,prob=.05, size=20)  
[1] 2
```



`rbinom(n, size, prob)`

- This draws a random sample of size n from $\text{Binom}(\text{size}, \text{prob})$

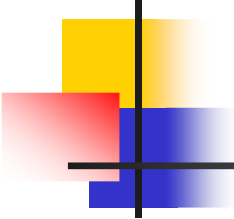
```
> rbinom(n=10, size=10, prob=.5)
[1] 4 4 4 5 6 4 4 4 1 5
```

Why this preponderance of values centered around 5?



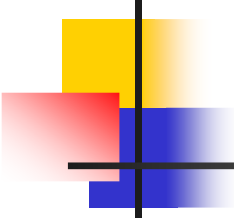
Section 4.9 - Expected Value and Variance of the Binomial Distribution

- Eq. 4.7 – The expected value and the variance of a binomial distribution are np and npq respectively.



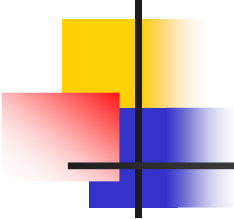
Section 4.10 – The Poisson Distribution

- Ex. 4.31 (Infectious Disease) – Consider the distribution of number of deaths attributed to typhoid fever over a long period of time, for example, 1 year. Assuming the probability of a new death from typhoid fever in any one day is very small and the number of cases reported in any two distinct periods of time are independent random variables, then the number of deaths over a one year period will follow a Poisson distribution.
- Ex. 4.32 (Bacteriology) – The preceding example concerns a rare event over time. Rare events can also be considered not only over time but also on a surface area, such as the distribution of number of bacterial colonies growing on an agar plate. Suppose we have a 100-cm² agar plate., the probability of finding any bacterial colonies at any 1 point a (or more precisely in a small area around a) is very small, and the events of finding bacterial colonies at any 2 points a_1, a_2 , are independent. The number of bacterial colonies over the entire agar plate will follow a Poisson distribution.



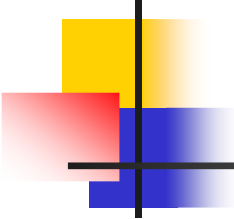
Section 4.10 – The Poisson Distribution

- Assumption 4.1 – Assume that
 1. The probability of observing 1 death is directly proportional to the length of the time interval Δt . That is, $\Pr(1 \text{ death})$ is approximately equal to $\lambda\Delta t$ for some constant λ .
 2. The probability of observing 0 deaths over Δt is approximately $1 - \lambda\Delta t$.
 3. The probability of observing more than 1 death over this time interval is essentially 0.



Section 4.10 – The Poisson Distribution

- Assumption 4.2 (Stationarity) – Assume the number of deaths per unit time is the same throughout the entire time interval. Thus an increase in the incidence of the disease as time goes on within time period t would violate the assumption. Note that t should not be overly long, because this assumption is less likely to hold as t increases.
- Assumption 4.3 (Independence) – If a death occurs within one time subinterval, then it has no bearing on the probability of death in the next time subinterval. This assumption would be violated in an epidemic situation, because if a new case of disease occurs, then subsequent deaths are likely to build up over a short period until after the epidemic subsides.



Section 4.10 – The Poisson Distribution

- Eq. 4.8 – The probability of k events occurring in a time period t for a Poisson random variable with parameter λ is given by

$$\Pr(X = k) = \frac{e^{-\mu} \mu^k}{k!}, \quad k = 0, 1, 2, \dots$$

where $\mu = \lambda t$ and e is approximately 2.71828

- N. B.
 - λ is the expected number of events over time period t .
 - For $X \sim \text{Binom}(n, p)$ n is finite and X is bounded by n
 - For $X \sim \text{Poiss}(l)$ X is infinite and number of deaths is infinite but large k implied very small probability



Poisson Distribution in R

Description:

Density, distribution function, quantile function
and random
generation for the Poisson distribution with
parameter 'lambda'.

Usage:

```
dpois(x, lambda, log = FALSE)
ppois(q, lambda, lower.tail = TRUE, log.p = FALSE)
qpois(p, lambda, lower.tail = TRUE, log.p = FALSE)
rpois(n, lambda)
```



```
dpois(x, lambda, log = FALSE)
```

- Ex. 4.33

```
> dpois(x = 0, lambda=2.3)
```

```
[1] 0.1002588
```

```
> dpois(x = 1, lambda=2.3)
```

```
[1] 0.2305953
```

Is there a slick way to compute all of these probabilities from (0,1, ..., 5)?



Section 4.12 Expected Value and Variance of the Poisson Distribution

- Eq. 4.9 – For a Poisson distribution with parameter μ , the mean and variance are both equal to μ .



Section 4.13 – Poisson Approximation to the Binomial Distribution

- Eq. 4.10 (Poisson Approximation to the Binomial Distribution) – The binomial distribution with large n and small p can be accurately approximated by a Poisson distribution with parameter $\mu = np$.
- Ex. 4.40
- Can you solve this in R?



```
qpois(p, lambda, lower.tail = TRUE,  
log.p = FALSE)
```

- qpois finds q such that

$$p = \Pr(X \leq q \mid X \sim \text{Poiss}(\lambda))$$



rpois(n, lambda)

- Draws a sample of size n where $X \sim \text{Pois}(\lambda)$

```
> rpois(n=100, lambda=2.3)
```

```
[1] 3 5 2 1 1 1 1 2 6 1 1 1 1 1 5 3 4 3 2 5 5 1  
2 1 3 6 1 3 2 4 1 5 1 3 1 2 1 2 4 1 2 2 2 1 1 5  
4 3 2 0 3 3  
[53] 4 3 3 1 0 4 2 3 3 2 4 3 3 3 6 2 1 0 3 3 4 3  
1 3 3 4 3 8 2 4 0 2 3 7 1 3 2 1 2 4 1 1 3 6 4 1  
2 1
```



Infectious Disease

- Ex. 4.25 pg. 113
- Let X = number of gonorrhea cases reported over a three month period. How is X distributed?
- $X \sim \text{Poiss}(\mu)$
- What is μ ?

- What do we wish to compute?



Infectious Disease (cont.)

- How do we compute this in R?



Hypertension

- Ex. 4.32 pg. 114
 - Let X = number of hypertensives in a sibship.
 - How is X distributed?
-
- How can we make the desired calculation in R?

Hypertension

- 4.33 pg. 114
- The probability of 2+ affected siblings under the independence model is $0.079704 + 0.005832 = .086$
- Let X = number of sibships with two or more affected siblings.
- How is X distributed?

- How can we assess the independence assumption?

- How is this computed in R?



Homework Chapter 4

- 4.1, 4.2, 4.3, 4.4, 4.15, 4.17, 4.23, 4.26, 4.37, 4.42, 4.43, 4.63, 4.64, 4.65, 4.95