Data Analysis for the Life Sciences
Chapter 7: Statistical Models

Jesica S. Rodriguez-Lopez

CUNY The City University of New York

BROAD INSTITUTE

Statistical Models: Introduction

"All models are wrong, but some are useful" -George E. P. Box

- Why fit Models?
  - Explanation:
    - To know (explore?) the relationship between the explanatory variables and the dependent variable
    - Examples?
      - Galileo in the 16th century trying to describe the velocity of a falling object
    - Objective:
      - minimal number of explanatory variables
  - Prediction:
    - To create predictions for new cases.
    - Examples?
      - Father & son heights
    - Objective:
      - minimal number of explanatory variables

Statistical Models: Introduction

- p-value: the **<u>probability</u>** of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true.
- Probability distribution of some sort was used to quantify the null hypothesis
- Most p-values in the scientific literature are based on:
    - Sample averages or
    - least squares estimates from a linear model and
    - make use of the CLT

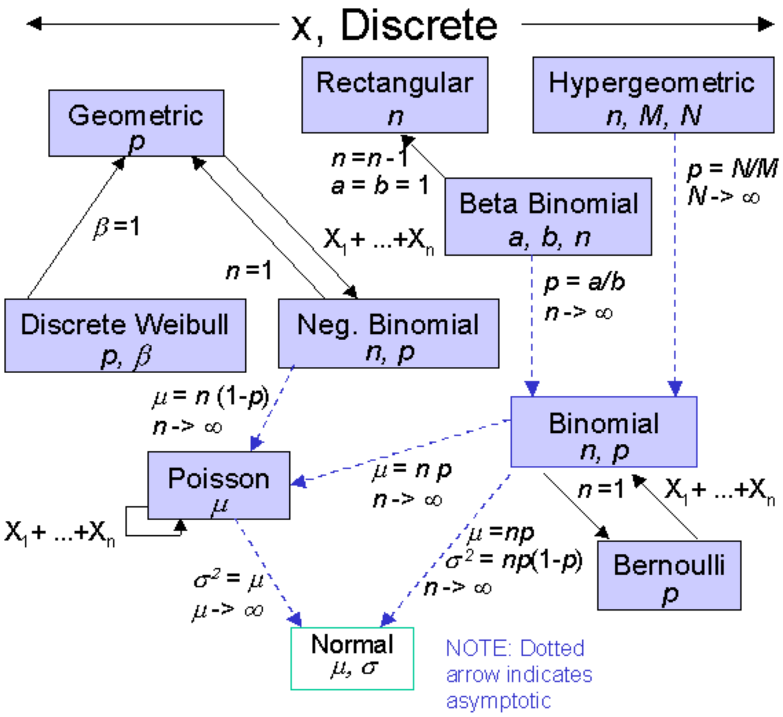
Statistical Models: Introduction

- Central Limit Theorem
  - Regardless of the population distribution model, as the sample size increases, the **sample mean** tends to be normally distributed around the population mean, and its standard deviation shrinks as n increases.
  - Valid when
    - Samples are independent
    - Sample size is “big enough”
- t-distribution approximation
  - ***When population data is approximately normal***, **sample mean** can be approximated as t-distribution

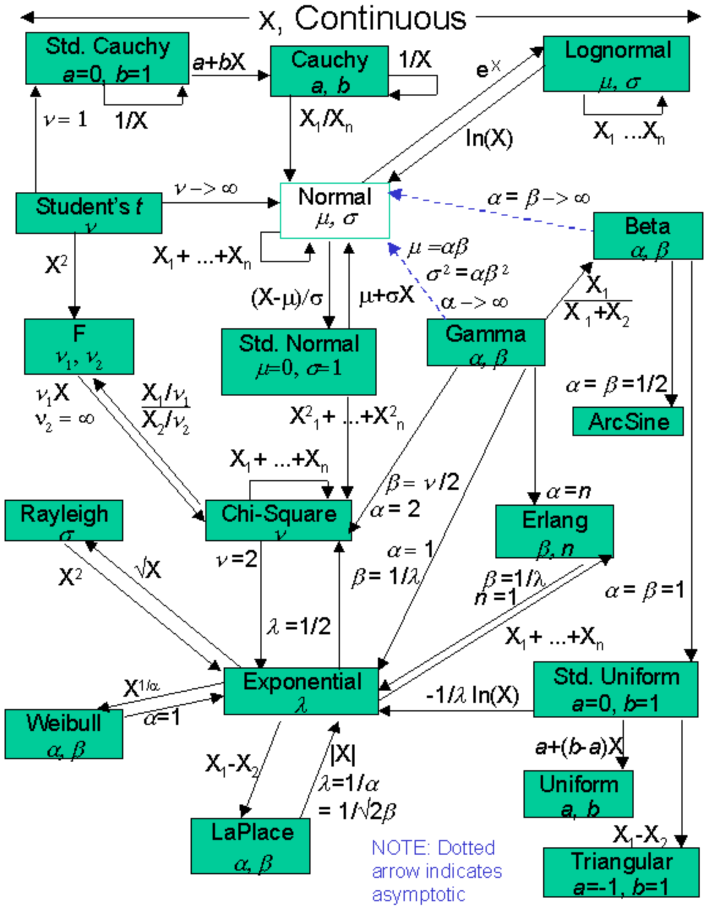
Statistical Models: Introduction

- What if:
  - Sample size is not “big enough”?
  - Samples are not independent?
  - Population distribution is not normal?
- Normal distribution is not the only **parametric** probability distribution!
  - coin tosses
  - the number of people who win the lottery, and
  - US incomes

Statistical Models: Introduction



NOTE: Dotted arrow indicates asymptotic



NOTE: Dotted arrow indicates asymptotic

Source: http://www.statisticalengineering.com/distributions.html

The Binomial Distribution

- It reports the probability of observing $S = k$ successes in N trails as:

  $$Pr(S \ = \ k) = \binom{N}{k} p^k (1-p)^{N-k}$$

  When p is the probability of success
- The best known example is coin tosses with S the number of heads when tossing N coins. In this example $p = 0.5$, if coins are balanced
- When N is large enough the proportion of successes $^S/_N \sim N\big(Np, Np(1-p)\big)$
- R functions:
  - $dbinom(k, N, p)$: returns binomial point probabilities: $Pr(S \ = \ k)$
  - $pbinom(k, N, p)$: returns binomial cumulative probabilities: $Pr(S \ \leq \ k)$
  - $qbinom(percentil, N, p)$: returns quantiles from binomial probabilities

The Binomial Distribution - Exercises

1. The probability of conceiving a girl is 0.49. What is the probability that a family with 4 children has 2 girls and 2 boys (you can assume no twins)?

$Pr(S\ = 2)$, with $N = 4$, and $p = 0.49$

$Pr(S\ = 2)$:

```
dbinom(2,4,0.49)
[1] 0.3747001
```

The Binomial Distribution - Exercises

2. What is the probability that a family with 10 children has 4 girls and 6 boys (you can assume no twins)?

$Pr(S \ = 4)$, with $N = 10$, and $p = 0.49$

$Pr(S \ = 4)$:

```
dbinom(4,10,0.49)
[1] 0.1966421
```

The Binomial Distribution - Exercises

3. The genome has 3 billion bases. About 20% are C, 20% are G, 30% are T and 30% are A. Suppose you take a random interval of 20 bases, what is the probability that the GC-content (proportion of Gs or Cs) is strictly above 0.5 in this interval (you can assume independence)?

$Pr(S > 10)$ since 10=0.5*20, with $N = 20$, and $p = 0.4 = 20\% \text{ from C} + 20\% \text{ from G}$

$Pr(S > 10) = 1 - Pr(S \leq 10)$:

```
1-pbinom(10,20,0.4)
[1] 0.1275212
```

The Binomial Distribution - Exercises

4. The probability of winning the lottery is 1 in 175,223,510. If 20,000,000 people buy a ticket, what is the probability that more than one person wins?

   $Pr(S > 1)$, with $N = 20{,}000{,}000$, and $p = 1/175{,}223{,}510$

   $Pr(S > 1) = 1 - Pr(S \leq 1)$:

```
1-pbinom(1, 20000000, 1/175223510)
[1] 0.006038878
```

The Binomial Distribution - Exercises

5. We can show that the binomial approximation is approximately normal when N is large and p is not too close to 0 or 1. This means that:

$$\frac{S_N - E(S_N)}{\sqrt{Var(S_N)}} \sim N(0,1)$$

   Using the results for sums of independent random variables, we can show that $E(S_N) = Np$ and $Var(S_N) = Np(1-p)$.

The Binomial Distribution - Exercises

5. (Cont…) The genome has 3 billion bases. About 20% are C, 20% are G, 30% are T, and 30% are A. Suppose you take a random interval of 20 bases, what is the exact probability that the GC-content (proportion of Gs of Cs) is greater than 0.35 and smaller or equal to 0.45 in this interval?

$Pr(7 < S \leq 9)$ since 7=0.35*20 and 9=0.45*20, with $N = 20$, and $p = 0.4 = 20\% \text{ from C} + 20\% \text{ from G}$

$Pr(7 < S \leq 9) = Pr(S \leq 9) - Pr(S \leq 7)$:

```
pbinom(9,20,0.4)-pbinom(7,20,0.4)
[1] 0.3394443
```

The Binomial Distribution - Exercises

6. For the question above, what is the normal approximation to the probability?

First, we need to standardize:

$$\frac{S_N - E(S_N)}{\sqrt{Var(S_N)}} \sim N(0,1)$$

$E(S_N) = Np = 20 * 0.4$ and

$Var(S_N) = Np(1-p) = 20 * 0.4 * 0.6$

```
b <- (9 - 20*.4)/sqrt(20*.4*.6)
a <-(7 - 20*.4)/sqrt(20*.4*.6)
pnorm(b)-pnorm(a)
[1] 0.3519231
```

The Binomial Distribution - Exercises

7. Repeat exercise **5**, but using an interval of 1000 bases. What is the difference (in absolute value) between the normal approximation and the exact distribution of the GC-content being greater than 0.35 and lesser or equal to 0.45?

   $Pr(350 < S \leq 450)$ since 350=0.35*1000 and 450=0.45*1000, with $N = 1000$, and $p = 0.4 = 20\% \text{ from C} + 20\% \text{ from G}$
   $Pr(350 < S \leq 450) = Pr(S \leq 450) - Pr(S \leq 350)$

The Binomial Distribution - Exercises

7. (Cont.)

Exact:

```
Exact= pbinom(450,1000,0.4)-pbinom(350,1000,0.4)
Exact
[1] 0.9987609
```

Approximation:

```
b <- (450 - 1000*0.4)/sqrt(1000*0.4*0.6)
a <- (350 - 1000*0.4)/sqrt(1000*0.4*0.6)
approx <- pnorm(b)-pnorm(a)
approx
[1] 0.9987512
```

Difference:

```
abs(exact-approx)
[1] 9.728752e-06
```

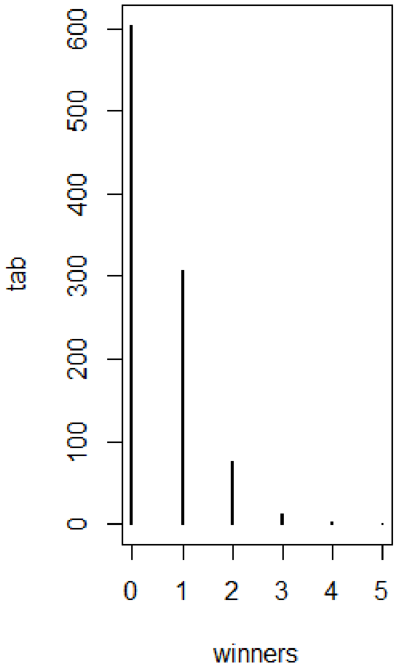
The Poisson Distribution

- The number of people that win the lottery follows a binomial distribution (we assume each person buys one ticket), where N: number of people who buy tickets is very large, and the p: probability of winning is very small, aprox. 1 to 3 people win. **Here CTL cannot/should not be apply**

The Poisson Distribution

```
p=10^-7 ##1 in 10,000,0000 chances of winning

N=5*10^6 ##5,000,000 tickets bought

winners=rbinom(1000,N,p) ##1000 is the
number of different lotto draws

tab=table(winners)

plot(tab)

prop.table(tab)
```

```
winners
    0     1     2     3     4     5
0.604 0.307 0.075 0.011 0.002 0.001
```



The Poisson Distribution

- For cases like this, where N is very large, but *p* is small enough to make $N * p$ (call it λ) a number between 0 and, for example, 10, then S can be shown to follow a Poisson distribution, which has a simple parametric form:

$$Pr(S \ = \ k) = \frac{\lambda^k exp(-\lambda)}{k!}$$

- Discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or area

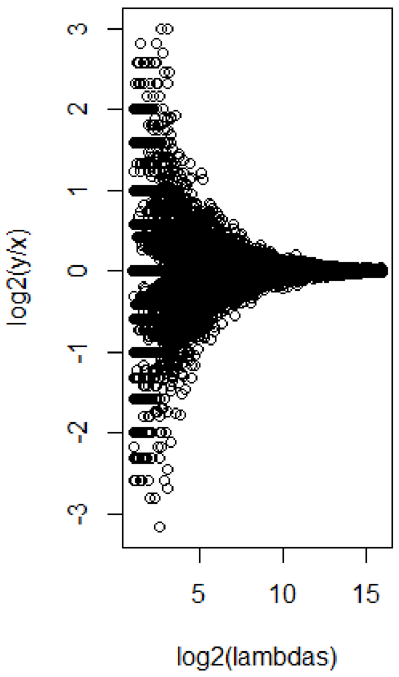
The Poisson Distribution

- The Poisson distribution is commonly used in RNAseq analyses. Because we are sampling thousands of molecules and most genes represent a very small proportion of the totality of molecules.
- Let’s say we want to report the genes with larges fold-changes. If the null hypothesis is that there is not differences, the statistical variability of this quantity depends on the total **abundance** of the gene
- **Abundance**: Number of events occurring in a fixed interval of time or area

The Poisson Distribution

```
N=10000 ##number of genes
lambdas=2^seq(1,16,len=N) ##these are the true abundances of genes
y=rpois(N,lambdas)##note that the null hypothesis is true for all genes
x=rpois(N,lambdas)
ind=which(y>0 & x>0) ##make sure no 0s due to ratio and log
library(rafalib)
splot(log2(lambdas),log2(y/x),subset=ind)
```

- For lower values of lambda there is much more variability and, if we were to report anything with a fold change of 2 or more, the number of false positives would be quite high for low abundance genes.



The Poisson Distribution - Exercises

8. The Cs in our genomes can be *methylated*□ or *unmethylated*. Suppose we have a large (millions) group of cells in which a proportion *p* of the Cs of interest are methylated. We break up the DNA of these cells and randomly select pieces and end up with *N* pieces that contain the C we care about. This means that the probability of seeing *k* methylated Cs is binomial:

   exact = dbinom(k,N,p)

   We can approximate this with the normal distribution:

   a <- (k+0.5 - N*p)/sqrt(N*p*(1-p))

   b <- (k-0.5 - N*p)/sqrt(N*p*(1-p))

   approx = pnorm(a) - pnorm(b)

   Compute the difference approx - exact for:

   N <- c(5,10,50,100,500)

   p <- seq(0,1,0.25)

The Poisson Distribution - Exercises

8. (Cont.) Compare the approximation and exact probability of the proportion of Cs being $p, k = 1, \dots, N-1$ plotting the exact versus the approximation for each p and N combination.

Study the plots and tell which one of the following is NOT true

A) The normal approximation works well when p is close to 0.5 even for small $N = 10$

B) The normal approximation breaks down when p is close to 0 or 1 even for large N

C) When N is 100 all approximations are spot on.

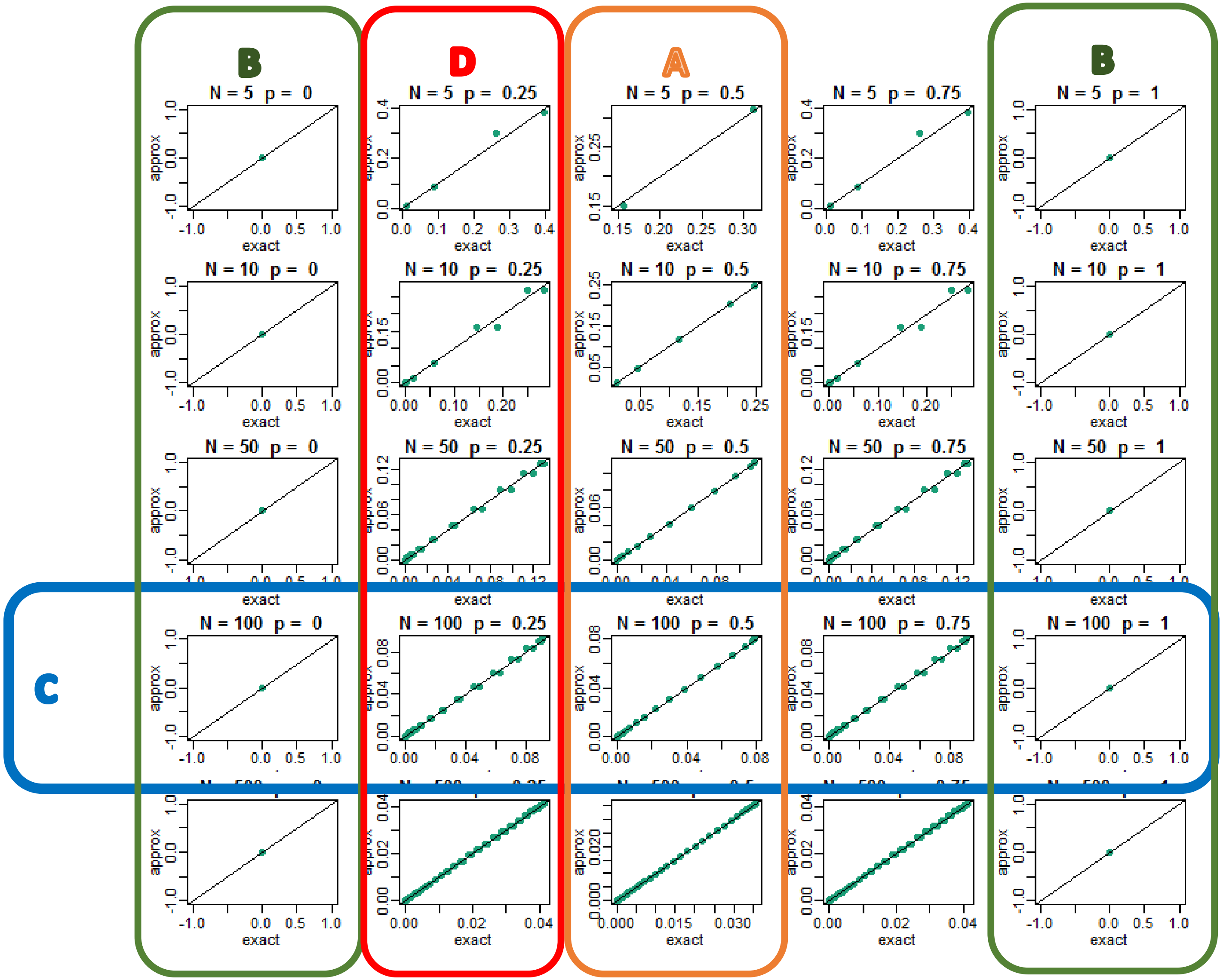
D) When $\underline{p = 0.\mathbf{25}}$ the approximation are terrible for $\underline{N =}$ $\mathbf{5, 10, 50, 100}$ and only OK for $N = \mathbf{500}$

The Poisson Distribution - Exercises

8. (Cont.)

```
Ns <- c(5,10,50,100,500)
ps <- seq(0,1,0.25)
library(rafalib)
mypar(4,5)
for(N in Ns){
    ks <- 1:(N-1)
    for(p in ps){
        exact = dbinom(ks,N,p)
        a = (ks+0.5 - N*p)/sqrt(N*p*(1-p))
        b = (ks-0.5 - N*p)/sqrt(N*p*(1-p))
        approx = pnorm(a) - pnorm(b)
        LIM <- range(c(approx,exact))
        plot(exact,approx,main=paste("N =",N," p = ,p), xlim=LIM, ylim=LIM, col=1, pch=16)
        abline(0,1)
    }
}
```

The Poisson Distribution - Exercises



The Poisson Distribution - Exercises

9. We saw in the previous question that when pp is very small, the normal approximation breaks down. If NN is very large, then we can use the Poisson approximation.

   Earlier we computed 1 or more people winning the lottery when the probability of winning was 1 in 175,223,510 and 20,000,000 people bought a ticket. Using the binomial, we can compute the probability of exactly two people winning to be:

   ```
   N <- 20000000
   p <- 1/175223510
   dbinom(2,N,p)
   [1] 0.005811321
   ```

The Poisson Distribution - Exercises

9. (Cont.) If we were to use the normal approximation, we would greatly underestimate this:

```
a <- (2+0.5 - N*p)/sqrt(N*p*(1-p))
b <- (2-0.5 - N*p)/sqrt(N*p*(1-p))
pnorm(a) - pnorm(b)
[1] 2.04756e-05
```

To use the Poisson approximation here, use the rate $\lambda = Np$ representing the number of people per 20,000,000 that win the lottery. Note how much better the approximation is:

```
dpois(2,N*p)
[1] 0.005811321
```

The Poisson Distribution - Exercises

9. (Cont.) In this case. it is practically the same because N is very large and Np is not 0. These are the assumptions needed for the Poisson to work. What is the Poisson approximation for more than one person winning?

   From 3:

   Exact:

   $Pr(S > 1)$, with $N = 20{,}000{,}000$, and $p = 1/175{,}223{,}510$

   $Pr(S > 1) = 1 - Pr(S \leq 1)$:

   ```
   1-pbinom(1, 20000000, 1/175223510)
   [1] 0.006038878
   ```

   Poisson Approximation:

   ```
   N = 20000000
   p = 1/175223510
   1 - ppois(1, N*p)
   [1] 0.006038879
   ```