StatGenet_April16

Probability; binomial distribution

- N trials
- Two outcomes {0,1}
- Probability $(p_1, p_2); p_1 + p_2 = 1$

We don't know how many 0s and 1s, (k, N-k), will be observed.

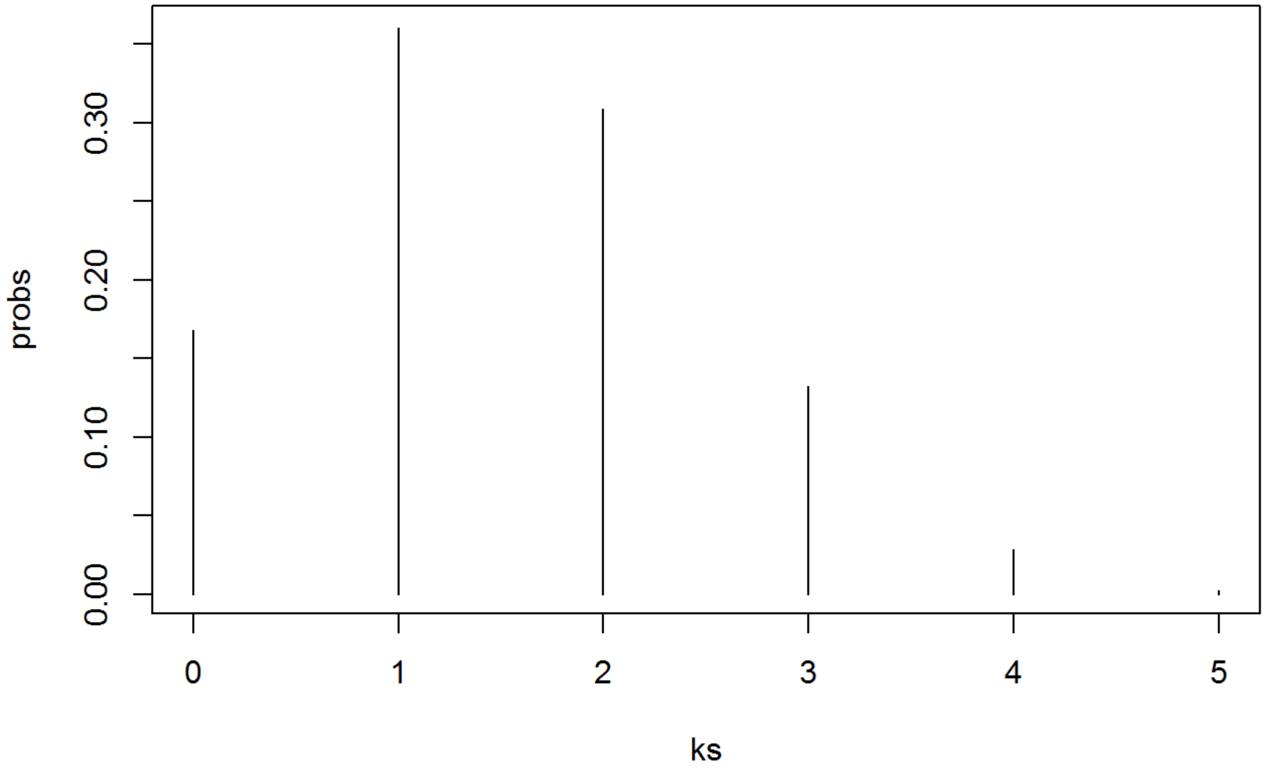
$$P((k, N-k)|(p_1, p_2)) = \binom{N}{k} p_1^k \times p_2^{N-k}$$
$$= \frac{N!}{k!(N-k)!} p_1^k \times p_2^{N-k}$$

```
N <- 5
k <- 3
ps <- c(0.3,0.7)
my.prob <- function(N,k,ps){
  factorial(N)/(factorial(k)*factorial(N-k)) * ps[1]^k * ps[2]^(N-k)
}
my.prob(N,k,ps)
```

```
## [1] 0.1323
```

- Calculate all cases and draw them.

```
ks <- 0:N
probs <- rep(0,length(ks))
for(i in 1:length(ks)){
  probs[i] <- my.prob(N,ks[i],ps)
}
plot(ks,probs,type="h")
```




```
sum(probs)
```

```
## [1] 1
```

Likelihood

Probability

- N trials
- Two outcomes {0,1}
- Probability $(p_1, p_2); p_1 + p_2 = 1$

We don't know how many 0s and 1s will be observed.

$$P((k, N-k)|(p_1, p_2)) = \binom{N}{k} p_1^k \times p_2^{N-k}$$
$$= \frac{N!}{k!(N-k)!} p_1^k \times p_2^{N-k}$$

Likelihood

- N trials
- Two outcomes {0,1}
- Observations (k,N-k)

We don't know probability $(p_1, p_2); p_1 + p_2 = 1$

$$L((p_1,p_2)|(k,N-k)) = \begin{pmatrix} N \\ k \end{pmatrix} p_1^k \times p_2^{N-k}$$
$$= \frac{N!}{k!(N-k)!} p_1^k \times p_2^{N-k}$$

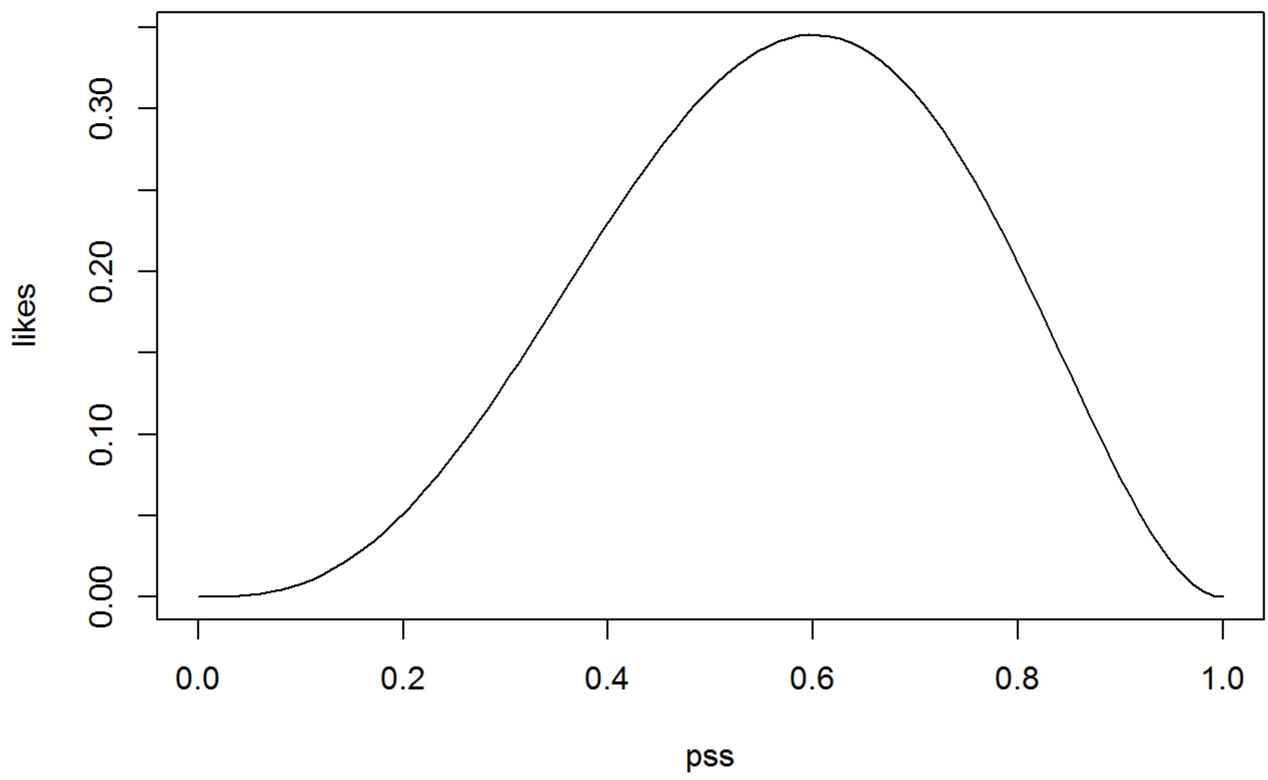
Do we have to change the function “my.prob()”?

```
N <- 5
k <- 3
ps <- c(0.3,0.7)
my.prob <- function(N,k,ps){
  factorial(N)/(factorial(k)*factorial(N-k)) * ps[1]^k * ps[2]^(N-k)
}
my.prob(N,k,ps)
```

```
## [1] 0.1323
```

- Calculate all cases and draw them.

```
pss <- seq(from=0,to=1,length=100)
likes <- rep(0,length(pss))
for(i in 1:length(pss)){
  this.ps <- c(pss[i],1-pss[i])
  likes[i] <- my.prob(N,k,this.ps)
}
plot(pss,likes,type="l")
```



```
sum(likes) # ??
```

```
## [1] 16.5
```

Most likeliness and differentiation

- Which (p_1, p_2) did give you the highest value?
- The likelihood curve peaked at the maximum likelihood estimate (MLE).
- $\frac{dL((p_1,p_2)|(k,N-k))}{dp_1}(p_{1,MLE}) = 0$

?? The 1st derivative of $L((p_1, p_2)|(k, N-k))$

The area under the curve and integration

```
sum(likes)
```

```
## [1] 16.5
```

Numeric integration

```
pss[2]-pss[1]
```

```
## [1] 0.01010101
```

```
diff(pss)
```

```
##  [1] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
##  [7] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [13] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [19] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [25] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [31] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [37] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [43] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [49] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [55] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [61] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [67] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [73] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [79] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [85] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [91] 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101 0.01010101
## [97] 0.01010101 0.01010101 0.01010101
```

The area under the curve is

```
sum(likes * (pss[2]-pss[1]))
```

```
## [1] 0.1666667
```

```
1/(1:10)
```

```
## [1] 1.0000000 0.5000000 0.3333333 0.2500000 0.2000000 0.1666667 0.1428571
## [8] 0.1250000 0.1111111 0.1000000
```

- What does this value mean?

$$\int_0^1 L((p_1, p_2)|(k, N-k))dp_1 = \frac{1}{6}$$

or

$$\int_0^1 6 \times L((p_1, p_2)|(k, N-k))dp_1 = 1$$
$$\int_0^1 6 \times \frac{N!}{k!(N-k)!} p_1^k \times p_2^{N-k} dp_1 = 1$$
$$\int_0^1 6 \times \frac{5!}{3!(5-3)!} p_1^3 \times p_2^{5-3} dp_1 = 1$$
$$\int_0^1 \frac{(5+1)!}{3!(5-3)!} p_1^3 \times p_2^{5-3} dp_1 = 1$$
$$\int_0^1 \frac{(N+1)!}{k!(N-k)!} p_1^k \times p_2^{N-k} dp_1 = 1$$

$$\Gamma(\alpha) = (\alpha - 1)!$$
$$\Gamma(\beta + 1) = \beta!$$

$$\int_0^1 \frac{(N+1)!}{k!(N-k)!} p_1^k \times p_2^{N-k} dp_1 = 1$$
$$\int_0^1 \frac{\gamma(N+2)}{\Gamma(k+1)\Gamma((N-k)+1)!} p_1^k \times p_2^{N-k} dp_1 = 1$$
$$\int_0^1 \frac{\gamma((k+1)+((N-k)+1))}{\Gamma(k+1)\Gamma((N-k)+1)!} p_1^k \times p_2^{N-k} dp_1 = 1$$
$$\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)!} p_1^{\alpha-1} \times p_2^{\beta-1} dp_1 = 1$$

Beta distribution

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_1^{\alpha-1} \times p_2^{\beta-1}$$

- Beta distribution is a “distribution function” because its area under the curve is 1.
- Beta distribution is in a good relation with binomial distribution.
- This “good relation” is described as “beta distribution is the conjugate prior of bionmial distribution”.

See the “Example” section of “Conjugate prior” in Wikipedia. https://en.wikipedia.org/wiki/Conjugate_prior (https://en.wikipedia.org/wiki/Conjugate_prior)

Information table of distribution articles of Wikipedia

https://en.wikipedia.org/wiki/Binomial_distribution (https://en.wikipedia.org/wiki/Binomial_distribution)

Multinomial distribution and Dirichlet distribution

https://en.wikipedia.org/wiki/Multinomial_distribution (https://en.wikipedia.org/wiki/Multinomial_distribution)

https://en.wikipedia.org/wiki/Dirichlet_distribution (https://en.wikipedia.org/wiki/Dirichlet_distribution)