

# Module 3: Binomial model

*Ege Rubak and Jesper Møller*  
*Based on material by Søren Højsgaard*

## Rolling a six

Consider a simple experiment: Roll a die  $n$  times; record the number of times six comes up, and denote it  $y$ .

Suppose e.g.  $n = 10$  and  $y = 3$ .

We let  $\theta$  denote the true (but to us unknown) probability of rolling a six (success) and  $1 - \theta$  is the probability of a roll less than six (failure):

$$Pr(S) = \theta, \quad Pr(F) = 1 - \theta.$$

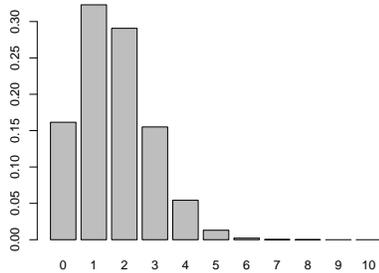
## The binomial model

The binomial distribution could be a model for these data:  $y = 3$  is a realization of a binomial random variable  $Y \sim bin(n, \theta)$ .

The density for  $y$  is

$$Pr(Y = y; \theta) = \frac{n!}{y!(n-y)!} \theta^y (1-\theta)^{n-y}.$$

```
yval <- 0:10  
barplot(dbinom(yval, size=10, prob=1/6), names.arg=yval)
```



---

Thus, if we know  $\theta$ , then we can make all sorts of interesting computations based on the binomial model. E.g. the mean and variance which are given by

$$E(Y) = n\theta, \quad \text{Var}(Y) = n\theta(1 - \theta).$$

Or we can calculate the probability of seeing 0 sixes or the probability of seeing 5 or more sixes when rolling a die 10 times.

For example, if the die is fair and  $\theta = 1/6$  we get

```
dbinom(0, size=10, prob=1/6) # Pr(0 sixes)
```

```
## [1] 0.1615056
```

```
1-pbinom(4, size=10, prob=1/6) # Pr(5 or more sixes)
```

```
## [1] 0.01546197
```

## A moment estimate

In practice  $\theta$  is unknown and must be estimated from data. Intuition says that  $\theta$  should be estimated as

$$\hat{\theta} = y/n = 3/10 = 0.3.$$

It is useful to write  $\hat{\theta}(y) = y/n$  to emphasize the dependence on data.

For the corresponding random variable  $\hat{\theta}(Y) = Y/n$ ,

$$E(Y/n) = \theta, \quad \text{Var}(Y/n) = \frac{1}{n^2}n\theta(1 - \theta) = \theta(1 - \theta)/n.$$

Hence  $\hat{\theta}(Y)$  has the correct mean value (unbiased) and the variance of  $\hat{\theta}(Y)$  goes to 0 when  $n \rightarrow \infty$  (consistent).

To calculate the variance, we plug in the estimate and find

$$\text{Var}(Y/n) \approx \frac{y}{n}(1 - \frac{y}{n})/n = 0.3 \times 0.7/10 = 0.021$$

I.e. the estimated standard deviation of the estimate (called std. error) is approximately  $\sqrt{0.021} = 0.14$ .

## The likelihood

When  $y = 3$  is observed, then the binomial density is a function of  $\theta$ , and it is called the likelihood function:

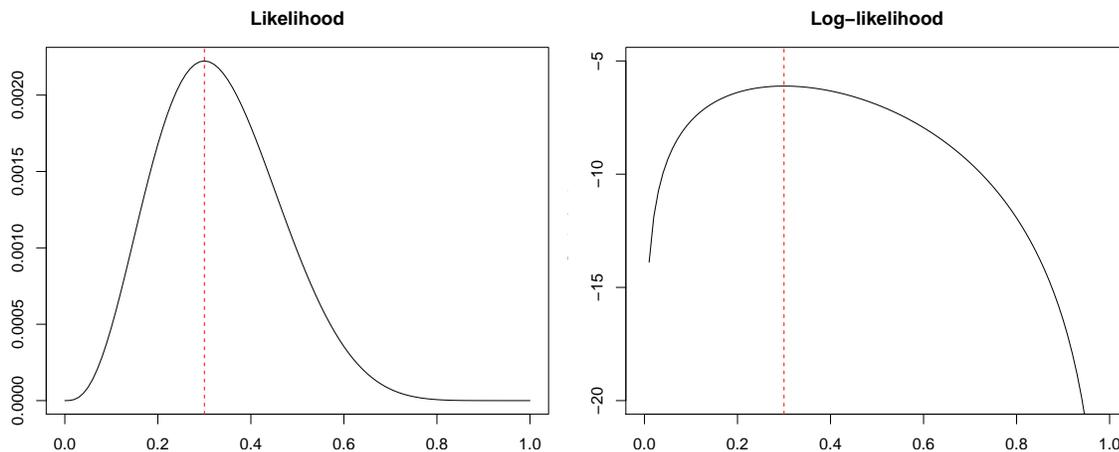
$$L(\theta) = Pr_{\theta}(Y = y) = \frac{n!}{y!(n - y)!}\theta^y(1 - \theta)^{n - y} \propto \theta^y(1 - \theta)^{n - y}.$$

Hence, the log-likelihood is

$$l(\theta) = \log L(\theta) = y \log \theta + (n - y) \log(1 - \theta).$$

---

```
lik <- function(parm, y, n){parm^y * (1 - parm)^(n - y)}
loglik <- function(parm, y, n){y * log(parm) + (n - y) * log(1 - parm)}
n <- 10; y <- 3
curve(lik(x, y, n), main = "Likelihood")
abline(v = y/n, col = "red", lty = 2)
curve(loglik(x, y, n), main = "Log-likelihood", ylim = c(-20, -5))
abline(v = y/n, col = "red", lty = 2)
```



## The frequentist approach

From a frequentist perspective we want to find the “best” estimate of  $\theta$  given data, and “best” is here the value of  $\theta$  that maximizes  $L(\cdot)$ . The estimate is called the maximum likelihood estimate (MLE).

In practice it is usually easier to maximize  $l(\theta) = \log L(\theta)$  instead of  $L(\theta)$  because the log turns a product into a sum, and sums are easier to differentiate than products.

The usual approach to maximizing  $l(\theta)$  is to first differentiate  $l(\theta)$  to obtain

$$S(\theta) = l'(\theta)$$

where  $S(\theta)$  is called the **score function**. Next we solve the **score equation**  $S(\theta) = 0$  to obtain  $\hat{\theta}$ .

It is not hard to spot that the maximum of  $L(\cdot)$  (and of  $l(\cdot)$ ) is at  $\theta = y/n = 0.3$ , but it is informative to look at the plots for different choices of  $n$  and  $y$ . This is left as an exercise.

## Summary of frequentist approach

- When we have a statistical model (a probability distribution) and data, then we have the likelihood (and the log-likelihood) functions.
- The maximum likelihood estimate (MLE)  $\hat{\theta}$  is the value of  $\theta$  that maximizes the likelihood function.
- The variance of the corresponding estimator is approximately minus the inverse of the second derivative of the log likelihood evaluated at the MLE.

## The Bayesian approach

If we take a Bayesian perspective then things change as explained in the main slides for this module:

The parameter  $\theta$  is a random quantity on equal footing with  $Y$

and we have to specify the prior

$$\pi(\theta).$$

This is our belief about  $\theta$  before seeing any data, and then given data  $y$  we get the posterior from the likelihood via Bayes’ rule

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)},$$

where  $\pi(y|\theta)$  is (proportional to) the likelihood we have specified previously and  $\pi(y)$  is the marginal probability for the data  $y$ .

When data  $y$  is observed then  $\pi(y)$  above is just a number, which ensures that  $\pi(\theta|y)$  is a density, i.e. that  $\int \pi(\theta|y)d\theta = 1$ . We do often not calculate  $\pi(y)$  directly: We just use that  $\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$

## A discrete prior

Example: Assume (to ease computations) that the only valid choices for  $\theta$  are now .1, .3, .5, .7 and .9. Before rolling the die we think the die has been rigged somehow and we take the prior to be

```
theta <- c(.1, .3, .5, .7, .9)
prior <- c(0.10, 0.15, 0.25, 0.30, 0.20)
```

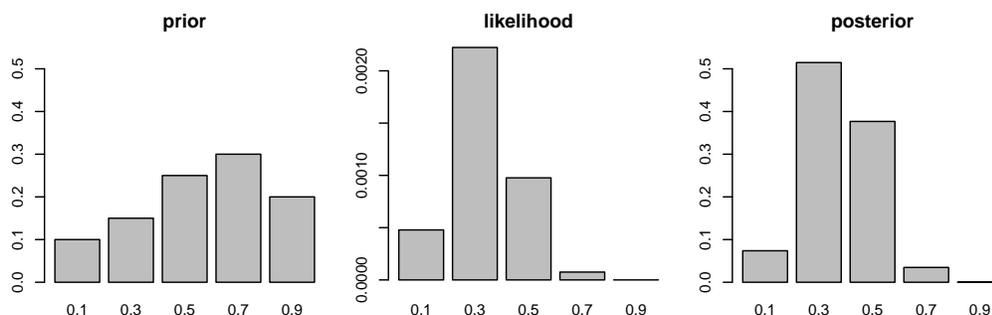
Then we can calculate the likelihood and the posterior

```
n <- 10; y <- 3
likval <- lik(theta, y, n)
posterior <- likval * prior
posterior <- posterior / sum( posterior )
round(100*posterior, 3)
```

```
## [1] 7.381 51.470 37.675 3.473 0.002
```

The plot shows it all:

```
par(mfrow=c(1,3), mar = c(3, 3, 3, 0.5))
barplot(prior, main="prior", names.arg=theta, ylim = c(0, .55))
barplot(likval, main="likelihood", names.arg=theta)
barplot(posterior, main="posterior", names.arg=theta, ylim = c(0, .55))
```



We can compute e.g. the prior and posterior means:

```
sum(theta * prior)
```

```
## [1] 0.57
```

```
sum(theta * posterior)
```

```
## [1] 0.374492
```

And corresponding variances (uncertainties before and after seeing data).

```
sum(theta^2 * prior) - sum(theta * prior)^2
```

```
## [1] 0.0611
```

```
sum(theta^2 * posterior) - sum(theta * posterior)^2
```

```
## [1] 0.01803762
```