**Chapter 2**

**“Probability” vs “Statistics”**

**fitting** We used different visualizations and showed how to run simulation experiments to test whether our data could be fit by a fair four-box multinomial model. We encountered the chi-square statistic and saw how to compare simulation and theory using a qq-plot.

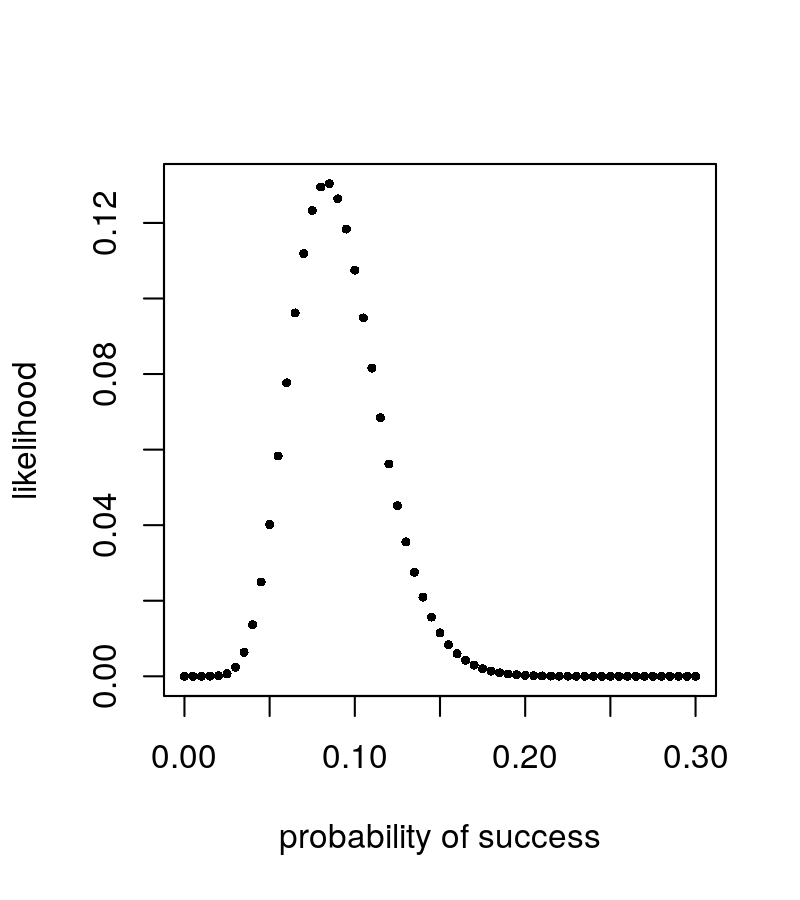
**Estimation** We explained maximum likelihood and Bayesian estimation procedures. These approaches were illustrated on examples involving nucleotide pattern discovery and haplotype estimations.

**Prior and posterior distributions** When assessing data of a type that has been previously studied, such as haplotypes, it can be beneficial to compute the posterior distribution of the data. This enables us to incorporate uncertainty in the decision-making, by way of a simple computation. The choice of the prior has little effect on the result as long as there is sufficient data.

**CpG islands and Markov chains** We saw how dependencies along DNA sequences can be modeled by Markov chain transitions. We used this to build scores based on likelihood ratios that enable us to see whether long DNA sequences come from CpG islands or not. When we made the histogram of scores, we saw in Figure [2.25](https://web.stanford.edu/class/bios221/book/Chap-Models.html#fig:chap2-r-ScoreMixture-1) a noticeable feature: it seemed to be made of two pieces. This **bimodality** was our first encounter with mixtures, they are the subject of Chapter [4](https://web.stanford.edu/class/bios221/book/Chap-Mixtures.html#Chap:Mixtures).

**Maximum likelihood** choose the parameter that makes the observed data the most likely

Example of binomial distribution



1. We want to know: how to calculate the likelihood of a certain parameter

likelihood = **dbinom**(**success\_counts** , prob=success.rate, size= num of trial)

loglikelihood = **function**(theta, n = 300, k = 40) {

115 **+** k **\*** **log**(theta) **+** (n **-** k) **\*** **log**(1 **-** theta)

}

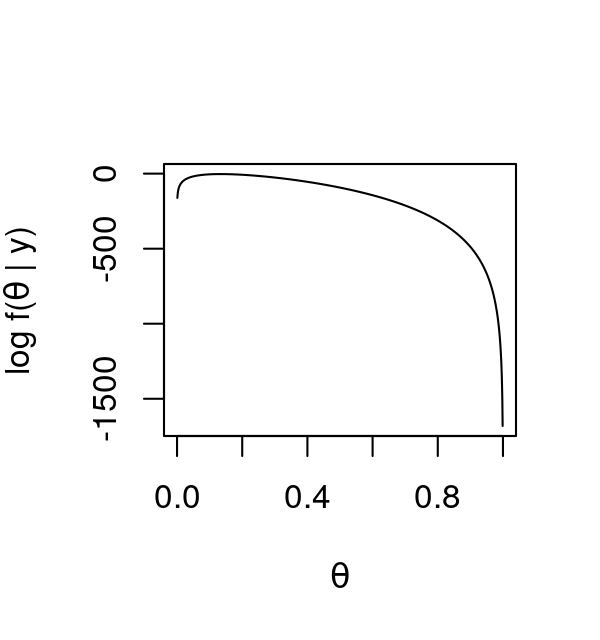
1. Generate the “candidate” parameter

thetas = **seq**(0, 1, by = 0.001)

1. Calculate the likelihood for all candidate parameter and plot

**plot**(thetas, **loglikelihood**(thetas), xlab = **expression**(theta),

ylab = **expression**(**paste**("log f(", theta, " | y)")),type = "l")

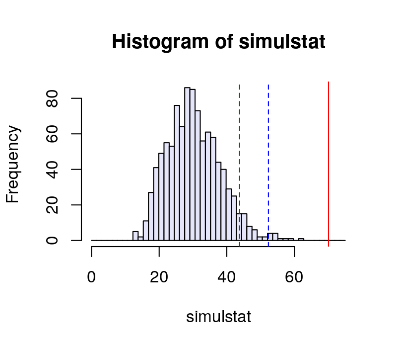


test whether the nucleotides are equally distributed across the four nucleotides for this first gene

stat = **function**(obsvd, exptd = 20 **\*** pvec) {

**sum**((obsvd **-** exptd)**^**2 **/** exptd)

}



**Bayesian Thinking**

The parameters of our distributions are –at least conceptually– definite, knowable, fixed numbers.

we use probability distributions to express our knowledge about the parameters, and use data to *update* this knowledge, for instance by shifting those distributions or making them more narrow; this is provided by the Bayesian paradigm

We have to start with an initial guess for the labels, estimate the parameters and go through several iterations of the algorithm, updating at each step our current best guess of the group labels and the parameters until we see no substantial improvement in our optimizations.