**Chapter1**

**The *Bernoulli* distribution** a single binary trial such as a coin flip

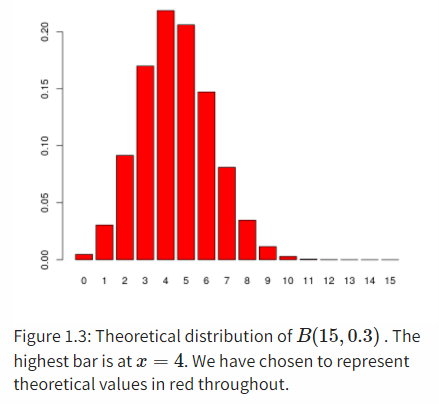
**The *binomial* distribution** n binary trial

**The *Poisson* distribution** when p is small (the 1s are rare). It has only one parameter λ, and the Poisson distribution for λ=np

**The *multinomial* distribution** discrete events that have more than two possible outcomes or **levels**

**Stimulation**

**The *binomial* distribution** n binary trial



probabilities = **dbinom**(0**:**15, prob = 0.3, size = 15)

Theoretical distribution of B(15,0.3)B(15,0.3) . The highest bar is at x=4

**The *Poisson* distribution** when p is small (the 1s are rare). It has only one parameter λ, and the Poisson distribution for λ=np

ELIZA example

the false positive rate is 1%: p(+|-)=0.01

50 patients: n=50

protein is tested at 100 different positions: 100 Independent trials

maxes = **replicate**(100000, {

**max**(**rpois**(100, 0.5))

})

**table**(maxes)

## 1 2 3 4 5 6 7 9

## 7 23028 60840 14364 1604 141 15 1

**mean**( maxes **>=** 7 )

## [1] 0.00016

P(x=7|lambda=0.5)

**Monte Carlo** method:

computer simulation

finds the probabilities of the events we’re interested in based on our generative model

**The *multinomial* distribution** discrete events that have more than two possible outcomes or **levels**

DNA example

how much data we need to collect if we want to test whether a multinomial model with equal probabilities is consistent with the data?

H0: pA=pC=pG=pT=1/4

sequence of length n=20

We’ll propose this as our critical value for testing data and will reject the hypothesis that the data come from a fair process, with equally likely nucleotides

pvec = **rep**(1**/**4, 4)

obsunder0 = **rmultinom**(1000, prob = pvec, size = 20)

exptd = 20 **\*** pvec

stat = **function**(obsvd, exptd) {

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

}

S0 = **apply**(obsunder0, 2, stat)

**quantile**(S0, probs = 0.95)-> 7.6

>95%: 1000 obsvd, sum((obsvd - exptd)^2 / exptd)<7.6

Data: pA=1/8,pC=3/8,pG=3/8,pT=1/8

pvecA = **c**(3**/**8, 1**/**4, 3**/**12, 1**/**8)

observed = **rmultinom**(1000, prob = pvecA, size = 20)

S1 = **apply**(observed, 2, stat)

power = **mean**(S1 **>** 7.6)

[1] 0.199

Run across 1000 simulations, the test identified 199 as coming from an alternative distribution. We’ve thus computed that the probability P(reject H0|HA)P(reject H0|HA) is 0.199.

With a sequence length of n=20, we have a power of about 20% to detect the difference between the fair generating process and our **alternative**.