

Math 4397/6397, Fall 2009  
Problem Set 5, due Thursday, Oct 1

Solutions

Problem 1. Suppose we study the number of times a student sits in a classroom with a TB infected, coughing neighbor until the student contracts the disease. Assume that each classroom encounter is an independent Bernoulli trial with probability  $p$  that the student becomes infected. This leads to the so-called geometric distribution  $P(\text{Person is infected on encounter } x) = p(1-p)^{x-1}$  for  $x = 1, 2, \dots$

a. The likelihood function for an observed  $x$  is

$$\mathcal{L}(p, x) = p(1-p)^{x-1}$$

so minimizing the log-likelihood

$$\ell(p, x) = \log p + (x-1) \log(1-p)$$

gives the (only) critical point

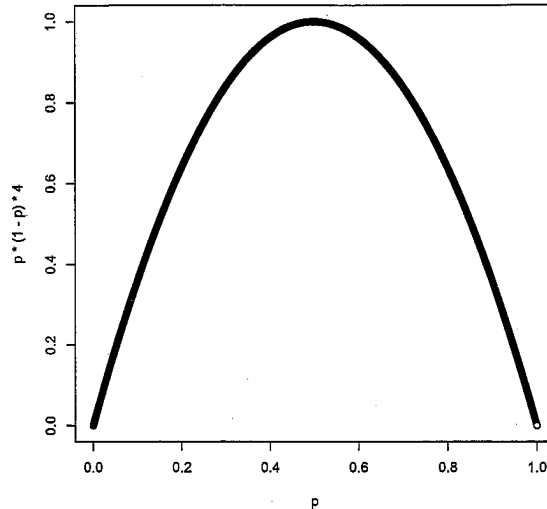
$$p = \frac{1}{x}.$$

At  $p = 0$  or  $p = 1$ , the likelihood is zero, but at  $p = 1/x$  it is positive, so this is the MLE.

b. Given  $x = 2$ . We use

```
p<-seq(0,1,0.001)
plot(p,p*(1-p)*4)
```

to plot the (normalized) likelihood. It seems that many  $p$  values are comparable (likelihood ratios, say, below 8) with the MLE.



- c. We compute  $P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = p + p(1-p) + p(1-p)^2 = 0.0297$ , so this would be a rare event.
- d. Assuming all individuals are infected in an iid manner, then the likelihood function is the product

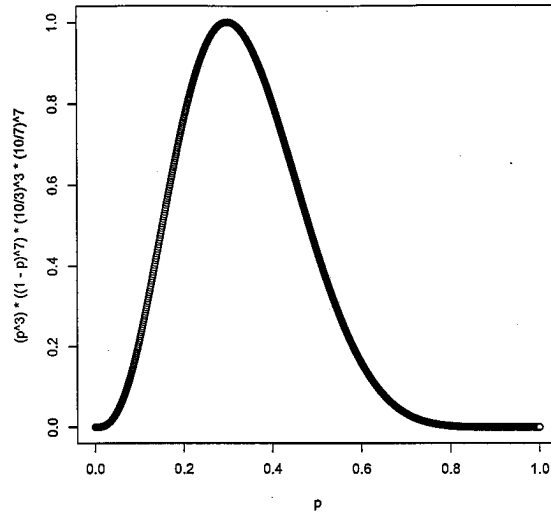
$$\mathcal{L}(p, x_1, x_2, \dots, x_n) = p(1-p)^{x_1-1} p(1-p)^{x_2-1} \dots p(1-p)^{x_n-1} = p^n (1-p)^{\sum_i x_i - n}$$

and the log likelihood

$$\ell(p, x_1, x_2, \dots, x_n) = n \log p + \left( \sum_{i=1}^n x_i - n \right) \log(1-p)$$

Again searching for a critical point gives  $p = \frac{n}{\sum x_i} = \frac{1}{\bar{X}_n}$ , with a positive likelihood, while at  $p = 0$  or  $p = 1$  the likelihood vanishes. Therefore,  $1/\bar{X}_n$  is the MLE.

- e. The likelihood function is now proportional to  $p^3(1-p)^7$ . The plot shows that more values for  $p$  can now be excluded by taking the likelihood ratio with the MLE.



Problem 2. Assuming that the results for each egg are independent and identically distributed:

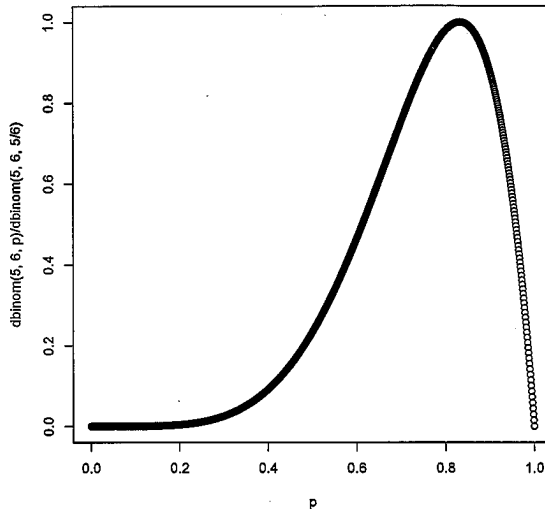
a. By the binomial distribution,

$$P(X \geq 5) = P(X = 5) + P(X = 6) = 6 * (0.1)^5 * (0.9) + (0.1)^6 = 0.000055$$

Observing this event is extremely rare under normal conditions.

b. We have the MLE estimate given by the proportion  $p = x/n = 5/6$ .

c. The plot gives



so comparing the likelihood at  $p = 0.1$  with that at  $p = 5/6$  gives the ratio

$$\frac{(5/6)^5(1/6)}{(0.1)^5(0.9)} = 7442.$$

This means there is 7442 times the evidence supporting the hypothesis  $p = 5/6$  compared to the hypothesis  $p = 0.1$ .

Problem 3. Suppose that IQs in a particular population are normally distributed with a mean of 110 and a standard deviation of 10.

- We have  $P(95 < IQ < 115) = P\left(\frac{95-110}{10} < Z < \frac{115-110}{10}\right) = \int_{-3/2}^{1/2} \phi(x)dx = \Phi(1/2) - \Phi(-3/2) = .62465$ .
- The 65th percentile for a std normal is .385, so here it is  $10*(.385)+110 = 113.85$ .
- For a single person,  $P(IQ > 130) = P\left(Z > \frac{130-110}{10}\right) = 1 - \Phi(2) = 0.0228$ . Now using this as the success probability  $p = 0.0228$  for a binomial r.v.  $X$  gives  $P(X \geq 4) = 5(0.0228)^4(1 - 0.0228) + 0.0228^5 = 1.33 \times 10^{-6}$ .
- Suppose that 500 people are sampled from this distribution. What is the probability of 400 (80%) or more having IQs above 130? We use the normal approximation for the binomial distribution. The expected number of people among

the 500 above IQ 130 is  $np = 500 * 0.0228 = 11.4$ , the standard deviation is  $\sqrt{np(1-p)} = 3.34$ . Thus we want to know  $P(Z > \frac{400-11.4}{3.34}) = P(Z > 116) = 0$ .

e. We have  $\bar{X}_{100} \sim N(110, \frac{10}{\sqrt{100}})$  This means the variance of  $\bar{X}_{100}$  is one and thus

$$P(\bar{X}_{100} > 112.5) = P(Z > 112.5 - 110) = 0.0062.$$

Problem 4. a. Using `x<-rexp(1000)` gives a mean

```
> mean(x)
[1] 0.9789478
```

and sample variance

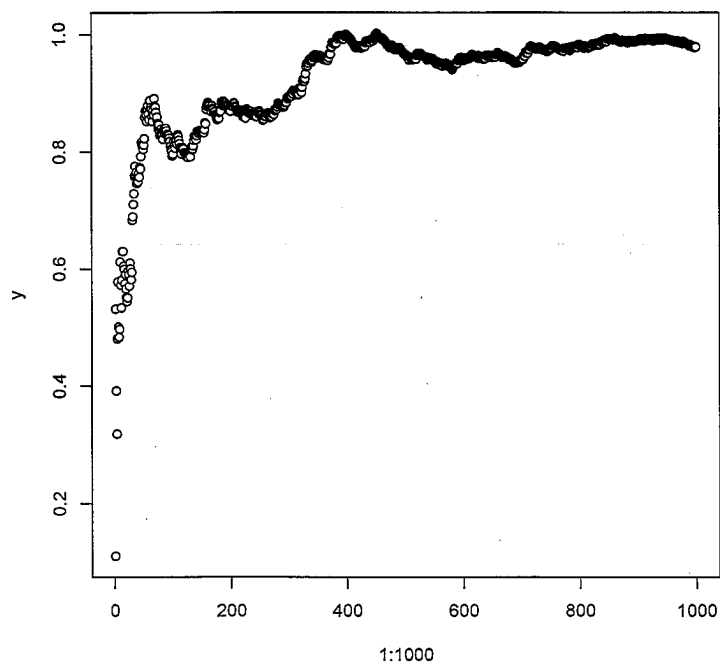
```
> var(x)
[1] 0.9337963
```

Both of these numbers should be close to one (their expected value) because of the law of large numbers.

b. We use the code

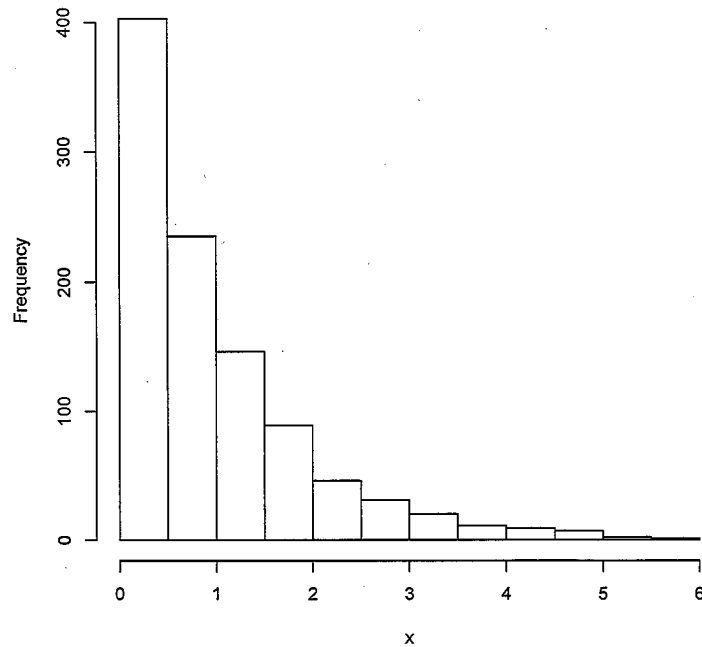
```
y <- cumsum(x) / (1 : length(x))
```

to create a vector of the sequential sample means and `plot(1:1000,y)` to obtain



c. The histogram

Histogram of x



looks roughly like an exponential density because of the law of large numbers (relative frequencies of occurrences for each "bin" tend to probabilities).

d. We use

```
> x<-rexp(1000*100)
> m<-matrix(x,1000,100)
> means<-apply(m,1,mean)
> hist(means)
```

The sample means should have an average close to  $x = 1$ , and a standard deviation (std error) of  $1/\sqrt{100} = \frac{1}{10}$ .

e. The histogram looks like that of a normally distributed random variable. This is a consequence of the central limit theorem. (Note that R has already rescaled the axis.)

Histogram of means

