Chapter 8

Revised exercises

8.1 Statistics: Estimating Parameters

MATHEMATICAL TECHNIQUES

♠ Suppose we wish to calculate the proportion of days that the temperature rises above 20°C. Evaluate the following sampling schemes.

- EXERCISE **8.1.1**
  Sample 100 consecutive days beginning on January 1.
- EXERCISE **8.1.2**
  Sample 100 consecutive days beginning on June 1.
- EXERCISE **8.1.3**
  Sample the temperature on March 15 for 100 years.
- EXERCISE **8.1.4**
  What might be a good way method if only 100 days could be sampled?

♠ A clever pollster decides to find the average income of people by calling random individuals up on their cellular phones. Suppose, however, that only 20% of people have cellular phones, and that their incomes are normally distributed according to $N(40000, 4.0\times10^6)$ (measured in 1999 U.S. dollars). The distribution for people without cellular phones is $N(20000, 1.0\times10^6)$.

- EXERCISE **8.1.5**
  What distribution describes the result of sampling 20 people with cellular phones? Use the rule of thumb that 95% of the distribution lies within two standard deviations of the mean to give a probable range. Compare this with the true average of the population.
- EXERCISE **8.1.6**
  What would be the results of sampling 20 people without cellular phones? Use the rule of thumb that 95% of the distribution lies within two standard deviations of the mean to give a probable range. Compare this with the true average of the population.

♠ Find the likelihood as a function of the binomial proportion p for each of the following, and evaluate at the expected value of p (use 1/2 for a fair coin and 1/6 for a fair die).

- EXERCISE **8.1.7**
  Flipping 2 out of 4 heads with a fair coin.
- EXERCISE **8.1.8**
  Rolling 2 out of 4 6's with a fair die. Why is the likelihood smaller than the answer to previous problem?
- EXERCISE **8.1.9**
  Flipping 2 out of 12 heads with a fair coin.
- EXERCISE **8.1.10**
  Rolling 2 out of 12 6's with a fair die. Why is the likelihood larger than the answer to previous problem?

♠ Find the probability distribution associated with the following situations, and say which part corresponds to the likelihood found in the earlier problem.

- EXERCISE **8.1.11**

  The number of heads in four flips of a fair coin. Compare with exercise 8.1.7.

- EXERCISE **8.1.12**

  The number of 6's rolled in four rolls of a fair die. Compare with exercise 8.1.8.

♠ Find the likelihood as a function of the binomial proportion p in each of the following cases, and find the maximum likelihood.

- EXERCISE **8.1.13**

  Team A wins 5 out of 6 games in a series against team B. Find the maximum likelihood estimator of the probability that team A wins a game against team B. If you were willing to gamble, would it make sense to enter a bet about the next game in the series where you win \$1 if team A wins, but lose \$6 if team A loses?

- EXERCISE **8.1.14**

  1 out of 150 people you know win \$500 in a raffle that costs \$5 to enter. Find the maximum likelihood estimator of the probability of winning the raffle. What is your best guess of the average payoff?

♠ Find the likelihood as a function of the Poisson parameter Λ, find the maximum likelihood, and evaluate the likelihood at the maximum and at the other given value of Λ.

- EXERCISE **8.1.15**

  20 events occur in one minute. Compare the likelihood with the maximum likelihood estimator of Λ with the likelihood if $\Lambda = 10.0$.

- EXERCISE **8.1.16**

  10 high energy cosmic rays hit in an expensive detector over the course of 1 year. Compare the likelihood with the maximum likelihood estimator of Λ with the likelihood if $\Lambda = 8.0$.

- EXERCISE **8.1.17**

  The number of events that occur are counted for 3 minutes. 20 events occur the first minute, 16 events occur the second minute, and 21 events occur the third minute. Compare the likelihood with the maximum likelihood estimator of Λ with the likelihood if $\Lambda = 20.0$.

- EXERCISE **8.1.18**

  10 high energy cosmic rays hit in an expensive detector in its first year, 7 in the second year, 11 in the third and 8 in the fourth and final year. Compare the likelihood with the maximum likelihood estimator of Λ with the likelihood if $\Lambda = 10.0$.

♠ Find the likelihood as a function of the parameter q of a geometric distribution, find the maximum likelihood, and evaluate the likelihood at the maximum and at the other given value of q.

- EXERCISE **8.1.19**

  Flies are tested for the ability to learn to fly toward the smell of potato, and the first one to succeed is the 13th. It had been predicted that the probability of success is 0.1.

- EXERCISE **8.1.20**

  Random compounds are tested for the ability to suppress a particular type of tumor, and first to succeed is the 94th. It had been predicted that the probability of success is 0.005.

- EXERCISE **8.1.21**

  The experiment testing flies for the ability to learn to fly toward the smell of potato is repeated three times. In the first experiment the first fly to succeed is the 13th, in the second experiment it is the 8th, and in the third experiment it is the 12th. It had been predicted that the probability of success is 0.1.

- EXERCISE **8.1.22**

  The experiment testing compounds for the ability to suppress tumors is repeated twice times. In the first experiment the first compound to succeed is the 94th, and in the second experiment it is the 406th. It had been predicted that the probability of success is 0.005.

♠ Write down the equations that would express the fact that the following estimators are unbiased.

- EXERCISE **8.1.23**

  The estimator of q in exercise 8.1.19.

• EXERCISE **8.1.24**
The estimator of Λ in exercise 8.1.15. If you think about the definition of the expectation, you might be able to demonstrate that this estimator is unbiased.

APPLICATIONS

♠ In each of the following cases where a very small number of individuals is tested for a gene, find and graph the likelihood function for the proportion p of individuals in the whole population with this gene, find the maximum likelihood, and make sense of the likelihood at $p = 0$ and $p = 1$.

• EXERCISE **8.1.25**
Two individuals are tested for a particular gene and one has it.

• EXERCISE **8.1.26**
Three individuals are tested for a particular gene, and all three have it.

♠ 30 out of 100 individuals are found to be infected with a disease. Estimate the proportion of infected women and infected men in the following circumstances. Assuming that the whole population is composed of 50% women, estimate the infected proportion in the whole population.

• EXERCISE **8.1.27**
The sample consists of 20 out of 50 infected women and 10 out of 50 infected men.

• EXERCISE **8.1.28**
The sample consists of 20 out of 40 infected women and 10 out of 60 infected men.

• EXERCISE **8.1.29**
The sample consists of 20 out of 20 infected women and 10 out of 80 infected men.

• EXERCISE **8.1.30**
The sample consists of 0 out of 50 infected women and 30 out of 50 infected men.

♠ Two couples are trying to have more girl babies. For each, find the likelihood function for the fraction q of female sperm and the maximum likelihood, and compare with the likelihood of $q = 0.5$.

• EXERCISE **8.1.31**
The first couple has 7 boys before having a girl. Use the geometric distribution to build the likelihood as a function of q.

• EXERCISE **8.1.32**
Another couple has 4 boys, then 1 girl, then 3 more boys. Find the likelihood as a function of q.

♠ Use the method of maximum likelihood to estimate the rate λ from the following exponential data.

event	waiting time 1	waiting time 2
1	1.565	0.47279
2	0.888	1.69516
3	0.874	3.67104
4	5.156	0.97018
5	0.018	0.37539
6	0.048	0.44228
7	1.496	3.79148
8	0.422	1.19057
9	0.721	0.14163
10	1.119	0.19354

In each case, find the likelihood function, find the maximum likelihood, and say whether it seems possible that the true rate is 1.0?

• EXERCISE **8.1.33**
For waiting time 1.

• EXERCISE **8.1.34**
For waiting time 2.

♠ Mutations are counted in 4 pieces of DNA that are one million base pairs long. There are 14 mutations in the first piece, 17 in the second piece, 8 in the third piece, and 5 in the fourth.

- EXERCISE **8.1.35**
  Write the likelihood function for the expected number of mutations per million bases in the first piece and find the maximum likelihood.
- EXERCISE **8.1.36**
  Write the likelihood function for the expected number of mutations per million bases in the second piece and find the maximum likelihood.
- EXERCISE **8.1.37**
  Write the likelihood function for the expected number of mutations per million bases in the first two pieces and find the maximum likelihood. Compare this with the estimated expected number for each of the two pieces separately.
- EXERCISE **8.1.38**
  Write the likelihood function for the expected number of mutations per million bases in the first four pieces and find the maximum likelihood. Compare this with the estimated expected number for each of the four pieces separately.

♠ Mutation rates differ depending on whether changing the nucleotide base changes the amino acid (non-synonymous sites) or not (synonymous sites). Suppose that mutation rates for synonymous sites are three times those in non-synonymous sites. A piece of DNA has 200 non-synonymous sites with 12 mutations and 100 synonymous sites with 15 mutations. Our goal is the estimate λ, the basic mutation rate per hundred non-synonymous sites.

- EXERCISE **8.1.39**
  Estimate λ using just the non-synonymous sites and just the synonymous sites.
- EXERCISE **8.1.40**
  Estimate λ using the non-synonymous and synonymous sites simultaneously (multiply the likelihood functions and find the maximum).

♠ Color-blindness is due to a recessive allele that appears on the X chromosome. If the color-blindness allele has frequency p, a fraction p of males will show the phenotype, and a fraction p^2 of females will show the phenotype (because they require two copies). We wish to estimate the fraction p from a sample of 1000 males, 90 of whom are color blind, and 1000 females, 13 of whom are color blind.

- EXERCISE **8.1.41**
  Estimate p using just the males and just the females.
- EXERCISE **8.1.42**
  How would you estimate p using the males and the females together? If you are feeling very determined, it is possible to solve the equations.

Chapter 9

Answers

8.1.1. This is terrible, because it samples mainly the winter (or the summer in Australia).
8.1.3. This is terrible, because March 15 is not representative of the whole year.
8.1.5. $N(40000, 2.0\times10^5)$ by the central limit theorem for averages (the variance is divided by $n = 20$). The standard deviation is \$447, so the range is from about \$39,050 to \$40,950. If 0.2 of the people average 40,000 and 0.8 of the people average 20,000, the overall average is $0.2 \cdot 40,000 + 0.8 \cdot 20,000 = 24,000$, which is far outside the range resulting from people with cellular phones.
8.1.7. $L(p) = b(2; 4, p) = 6p^2(1-p)^2$. For a fair coin, $p = 0.5$ and $L(0.5) = 0.375$.
8.1.9. $L(p) = b(2; 12, p) = 66p^2(1-p)^{10}$. For a fair coin, $p = 0.5$ and $L(0.5) = 0.016$.
8.1.11. The probability distribution follows the binomial distribution, with $b(0; 4, 0.5) = 0.0625$, $b(1; 4, 0.5) = 0.25$, $b(2; 4, 0.5) = 0.375$, $b(3; 4, 0.5) = 0.25$, $b(4; 4, 0.5) = 0.0625$. The probability of two heads with a fair coin matches the likelihood of $p = 0.5$ when there are two heads.
8.1.13. $L(p) = b(5; 6, p) = 6p^5(1-p)$. The derivative is

$$L'(p) = 30p^4(1-p) - 6p^5 = 6p^4(5(1-p) + p)$$

which is 0 at $p = 0$ and $5(1-p) + p = 6p - 5 = 0$ or $p = 5/6$. The maximum cannot occur at $p = 0$ because the likelihood there is 0 (or at the other endpoint $p = 1$ for the same reason), so the maximum likelihood estimator of the probability is $5/6$. If this is the real probability that team A wins, you win \$1 with probability $5/6$ and lose \$6 with probability $1/6$, for expected winnings of $1 \cdot 5/6 - 6 \cdot 1/6 = -1/6$. The bet would lose money on average.
8.1.15. $L(\Lambda) = p(20; \Lambda) = \frac{e^{-\Lambda}\Lambda^{20}}{20!}$. Then

$$\begin{aligned} L'(\Lambda) &= \frac{1}{20!}\left(-e^{-\Lambda}\Lambda^{20} + 20e^{-\Lambda}\Lambda^{19}\right) \\ &= \frac{e^{-\Lambda}\Lambda^{19}}{20!}(-\Lambda + 20) \end{aligned}$$

which is equal to 0 when $\Lambda = 0$ and when $\Lambda = 20.0$. Because $L(0) = 0$, $\Lambda = 20.0$ is the maximum likelihood estimator. Then $L(20.0) = 0.089$ and $L(10.0) = 0.0058$. The likelihood of $\Lambda = 20.0$ is higher.
8.1.17. The likelihood of all three measurements is the product of the likelihoods of each, so

$$\begin{aligned} L(\Lambda) &= p(20; \Lambda)p(16; \Lambda)p(21; \Lambda) \\ &= \frac{e^{-\Lambda}\Lambda^{20}}{20!}\frac{e^{-\Lambda}\Lambda^{16}}{16!}\frac{e^{-\Lambda}\Lambda^{21}}{21!} \\ &= \frac{e^{-3\Lambda}\Lambda^{57}}{20!16!21!} \end{aligned}$$

Then

$$\begin{aligned} L'(\Lambda) &= \frac{1}{20!16!21!}\left(-3e^{-3\Lambda}\Lambda^{57} + 57e^{-3\Lambda}\Lambda^{56}\right) \\ &= \frac{e^{-3\Lambda}\Lambda^{56}}{20!16!21!}(-3\Lambda + 57) \end{aligned}$$

which is equal to 0 when $\Lambda = 0$ and when $\Lambda = 19.0$. Because $L(0) = 0$, $\Lambda = 19.0$ is the maximum likelihood estimator. Then $L(19.0) = 0.000523$ and $L(20.0) = 0.000485$. Although both are tiny, the likelihood of $\Lambda = 19.0$ is higher.

8.1.19. The likelihood function is $L(q) = q(1-q)^{12}$, with derivative

$$L'(q) = (1-q)^{12} - 12q(1-q)^{11} = (1-q)^{11}(1-q-12q).$$

This is zero at $q = 1$, where the likelihood is 0, and at $q = 1/13$. $L(1/13) = 0.0294$ and $L(0.1) = 0.0282$, which is slightly smaller.

8.1.21. The likelihood function is the product of the likelihoods for each experiment, so

$$L(q) = q(1-q)^{12}q(1-q)^{7}q(1-q)^{11} = q^3(1-q)^{30}$$

with derivative

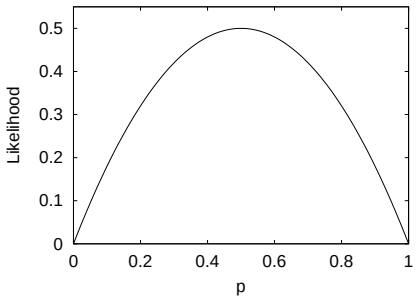
$$L'(q) = 3q^2(1-q)^{30} - 30q^3(1-q)^{29} = q^2(1-q)^{29}(3-33q).$$

This is zero at $q = 0$ and $q = 1$, where the likelihood is 0, and at $q = 1/11$. $L(1/11) = 0.0000431$ and $L(0.1) = 0.0000424$, which is slightly smaller.

8.1.23. If the true proportion were q, the waiting time would be n with probability $q(1-q)^n$. In this case, we estimate $\hat{q} = \frac{1}{n}$. We would need to show

$$q = \sum_{n=1}^{\infty} \frac{1}{n} q(1-q)^n.$$

8.1.25. The likelihood function is $L(p) = 2p(1-p)$.



The maximum is at $p = 0.5$. The likelihood is zero if $p = 0$ or $p = 1$ because it is impossible to get 1 out of 2 if either nobody has the gene ($p = 0$) or everybody has the gene ($p = 1$).

8.1.27. We estimate that 40% of women and 20% of men are infected and estimate a proportion of $0.4 \cdot 0.5 + 0.2 \cdot 0.5 = 0.3$ in the total population (by the law of total probability). This is what we would get if we ignored the sex of individuals in the sample.

8.1.29. We estimate that 100% of women and 12.5% of men are infected and estimate a proportion of $1.0 \cdot 0.5 + 0.125 \cdot 0.5 = 0.5625$ in the total population.

8.1.31. The likelihood function is $L(q) = q(1-q)^7$, with maximum at $q = 1/8$. Then $L(1/8) = 0.049$ and $L(0.5) = 0.004$, which looks a lot smaller.

8.1.33. The likelihood function is $L(\lambda) = \lambda^{10}e^{-12.307\lambda}$. The maximum is at $\lambda = 0.812$. The likelihood is 5.69×10^{-6}, which looks pretty small. The likelihood of $\lambda = 1.0$ is 4.51×10^{-6}, which is pretty close. This seems perfectly plausible.

8.1.35. $L(\Lambda) = \frac{\Lambda^{14}e^{-\Lambda}}{14!}$. Taking the derivative, we get

$$L'(\Lambda) = \frac{14\Lambda^{13}e^{-\Lambda} - \Lambda^{14}e^{-\Lambda}}{14!} = \frac{\Lambda^{13}e^{-\Lambda}}{14!}(14-\Lambda).$$

This is 0 when $\Lambda = 0$ or $\Lambda = 14$. The maximum is at $\Lambda = 14$.

8.1.37. The likelihood is the product, or

$$\begin{aligned} L(\Lambda) &= \frac{\Lambda^{14}e^{-\Lambda}}{14!}\frac{\Lambda^{17}e^{-\Lambda}}{17!} \\ &= \frac{\Lambda^{31}e^{-2\Lambda}}{14!17!} \end{aligned}$$

Taking the derivative, we get

$$L'(\Lambda) = \frac{31\Lambda^{30}e^{-2\Lambda} - 2\Lambda^{31}e^{-2\Lambda}}{14!17!} = \frac{\Lambda^{30}e^{-2\Lambda}}{14!17!}(31 - 2\Lambda).$$

This is 0 when $\Lambda = 0$ or $\Lambda = 15.5$. The maximum is at $\Lambda = 15.5$ which is the average of the estimated expected numbers in the first and second pieces.

8.1.39. There are 2 sets of 100 non-synonymous sites, so the number of mutations follows a Poisson distribution with parameter $\Lambda = 2\lambda$. The likelihood function is

$$L_n(\lambda) = \frac{e^{-2\lambda}(2\lambda)^{12}}{12!} = \frac{2^1 2}{12!}e^{-2\lambda}lambda^{12}$$

which has a maximum at $\lambda = 6$. In the one set of 100 synonymous sites the number of mutations follows a Poisson distribution with parameter $\Lambda = 3\lambda$ (because the rate is three times as high). The likelihood function is

$$L_s(\lambda) = \frac{e^{-3\lambda}(3\lambda)^{15}}{15!} = \frac{3^1 5}{15!}e^{-3\lambda}lambda^{15}$$

which has a maximum at $\lambda = 5$.

8.1.41. The likelihood function for the males is

$$L_m(p) = \left(\begin{array}{c} 1000 \\ 90 \end{array} \right) p^{90}(1-p)^{910}.$$

We know that the maximum occurs at $p = 90/1000 = 0.09$. The likelihood function for the females is

$$L_f(p) = \left(\begin{array}{c} 1000 \\ 13 \end{array} \right) (p^2)^{13}(1-p^2)^{987}.$$

The right hand side is the likelihood function describing 13 successes out of 1000, but is a function of p^2 rather than p. Then the maximum occurs at $p^2 = 13/1000 = 0.013$ and $p = 0.114$.