Math 3081 (Probability and Statistics)

Lecture #14 of 27 $\sim$ July 28th, 2021

Maximum Likelihood Estimates

- One More Modeling Example
- Parameter Estimation
- Maximum Likelihood Estimates
- Estimators

This material represents §3.1.1-3.1.2 from the course notes, and problems 1-3 from WeBWorK 5.

More Models, I

On average, a certain Math 3081 instructor makes 3 typos per week (there are 4 lectures per week).

1. Describe the distribution of the random variable X measuring the total number of typos in one week.
2. What is the probability there are exactly 5 typos this week?
3. What is the probability there are no typos in today's lecture?
4. Describe the distribution of the random variable Y measuring the total amount of time before the next typo is made.
5. After one typo, what is the probability that at least 4 full lectures pass before another typo is made?
6. Estimate the probability of obtaining more than 180 typos if the course runs for 52 weeks.

More Models, II

On average, a certain Math 3081 instructor makes 3 typos per week (there are 4 lectures per week).

1. Describe the distribution of the random variable X measuring the total number of typos in one week.

More Models, II

On average, a certain Math 3081 instructor makes 3 typos per week (there are 4 lectures per week).

1. Describe the distribution of the random variable X measuring the total number of typos in one week.

- From the description, typos are a relatively rare event, and it is reasonable to assume that they are independent.
- So, the distribution of typos will be Poisson, and the parameter will be the average number of typos per week, which is $\lambda = 3$.

2. What is the probability there are exactly 5 typos this week?

More Models, II

On average, a certain Math 3081 instructor makes 3 typos per week (there are 4 lectures per week).

1. Describe the distribution of the random variable X measuring the total number of typos in one week.

- From the description, typos are a relatively rare event, and it is reasonable to assume that they are independent.
- So, the distribution of typos will be Poisson, and the parameter will be the average number of typos per week, which is $\lambda = 3$.

2. What is the probability there are exactly 5 typos this week?

- This is $P(P_\lambda = 5) = \dfrac{3^5 e^{-3}}{5!} \approx 0.1008$.

More Models, III

On average, a certain Math 3081 instructor makes 3 typos per week (there are 4 lectures per week).

3. What is the probability there are no typos in today's lecture?

More Models, III

On average, a certain Math 3081 instructor makes 3 typos per week (there are 4 lectures per week).

3. What is the probability there are no typos in today's lecture?

- The average number of typos in one lecture is $3/4$, so the number of typos in today's lecture will be Poisson-distributed with parameter $\lambda = 3/4$.
- Then the probability of no typos is $e^{-3/4} \approx 0.4724$.[1]

4. Describe the distribution of the random variable Y measuring the total time, in weeks, before the next typo is made.

[1] Since their arr sevral typoes in thiss footnorte, the actua; probability is 0.

More Models, III

On average, a certain Math 3081 instructor makes 3 typos per week (there are 4 lectures per week).

3. What is the probability there are no typos in today's lecture?

- The average number of typos in one lecture is $3/4$, so the number of typos in today's lecture will be Poisson-distributed with parameter $\lambda = 3/4$.
- Then the probability of no typos is $e^{-3/4} \approx 0.4724$.[1]

4. Describe the distribution of the random variable Y measuring the total time, in weeks, before the next typo is made.

- The waiting time will be exponential because the probability of obtaining a typo is independent of the amount of time since the last typo.
- The average time between typos is $3/4$ of a week, so $\lambda = 3/4$.

[1] Since their arr sevral typoes in thiss footnorte, the actua; probability is 0.

More Models, III

On average, a certain Math 3081 instructor makes 3 typos per week (there are 4 lectures per week).

5. After one typo, what is the probability that at least 4 full lectures pass before another typo is made?

More Models, III

On average, a certain Math 3081 instructor makes 3 typos per week (there are 4 lectures per week).

5. After one typo, what is the probability that at least 4 full lectures pass before another typo is made?

- Since the waiting time is exponential, this is $P(E_\lambda \geq 4) = e^{-4 \cdot 3/4} = e^{-3} \approx 0.0498$.

6. Estimate the probability of obtaining more than 180 typos if the course runs for 52 weeks.

More Models, III

On average, a certain Math 3081 instructor makes 3 typos per week (there are 4 lectures per week).

5. After one typo, what is the probability that at least 4 full lectures pass before another typo is made?

- Since the waiting time is exponential, this is $P(E_\lambda \geq 4) = e^{-4 \cdot 3/4} = e^{-3} \approx 0.0498$.

6. Estimate the probability of obtaining more than 180 typos if the course runs for 52 weeks.

- The exact distribution is Poisson with $\lambda = 52 \cdot 3 = 156$, but it is very cumbersome to evaluate the exact probability this way.
- Instead, by the central limit theorem, we can observe that the distribution is approximately normal with mean $\mu = 52\lambda = 156$ and standard deviation $\sqrt{52\lambda} \approx 12.4900$.
- Including a continuity correction, the approximate probability is $P(N_{156,12.4900} \geq 179.5) = P(N_{0,1} \geq 1.8815) = 0.0300$.

Overview of §3, I

We now move into the third chapter of the course, on parameter and interval estimation.

- So far, we have discussed probability and random variables and focused on useful distributions such as the normal, exponential, and Poisson distributions.
- In most cases, we have always started by being given a distribution and then proceeding to use it to answer various questions (e.g., finding probabilities, computing expected values, describing the behavior of sample averages, etc.).
- Our goal in this chapter is, in a fairly direct sense, to invert this analysis: instead, we start with data obtained by sampling a distribution with certain unknown parameters, and our goal is to extract information about the most reasonable values for these parameters given the observed data.

Overview of §3, II

- We will start by discussing "pointwise" estimation methods, which will allow us to find good predictions for the value of an unknown parameter.
- Next, we discuss various properties that we would like our estimators to have, and briefly explain how in some cases it is possible to establish that a particular parameter is actually optimal with respect to certain reasonable conditions.
- We finish by broadening our focus to "interval" estimation methods, which allow us to give a measurement of the expected precision of our estimate (thereby quantifying "how good" we think the estimate is).

Parameter Estimation, I

To motivate our formal development of estimation methods, we first outline a few scenarios in which we would like to use parameter estimation.

For consistency, we will call our unknown parameter θ throughout our discussion.

Parameter Estimation, II

Example: Suppose we have an unfair coin with an unknown probability θ of coming up heads. We would like to estimate θ.

- Suppose we flip the coin 10 times, and the results are TTTTT THTTH.
- What is the most reasonable estimate for θ given these results?

Parameter Estimation, II

Example: Suppose we have an unfair coin with an unknown probability θ of coming up heads. We would like to estimate θ.

- Suppose we flip the coin 10 times, and the results are TTTTT THTTH.
- What is the most reasonable estimate for θ given these results?
- In this case, it seems reasonable to say that since 8 of the flips are tails and 2 of the flips are heads, the most reasonable estimate for θ would be $2/10 = 0.2$.
- It seems far more likely that we would obtain the results above with a coin that has a $1/5$ chance of landing heads (since then the expected number of heads in 10 flips is 2, exactly what we observed) than, say, if the coin had a $1/2$ chance of landing heads (since then the expected number of heads would be 5, far more than we observed).

Parameter Estimation, III

Example: We expect the number of calls received by an emergency help line per hour at night should have a Poisson distribution with parameter $\lambda = \theta$. We want to estimate θ.

- Suppose we count the number of calls in five consecutive hours, and the totals are 4 calls, 2 calls, 0 calls, 3 calls, and 4 calls.
- What is the most reasonable estimate for θ given these results?
- To ponder: what seems like a plausible range of possible values for θ? How would you try to decide which value is "most reasonable"?

Parameter Estimation, IV

Example: Suppose that we are waiting for a package delivery from an unreliable service. We model the wait time by an exponential distribution with some parameter $\lambda = \theta$. We want to estimate θ.

- Suppose we order 4 packages, and the delivery times are 1.25 days, 0.02 days, 0.18 days, and 0.63 days.
- What is the most reasonable estimate for θ given these results?
- To ponder: what seems like a plausible range of possible values for θ? How would you try to decide which value is "most reasonable"?

Parameter Estimation, V

Here is an outline of one possible way to identify a plausible value of θ in the first two examples:

- First, we compute the probability of obtaining the sampling data we received in terms of θ.
- Then we search among the possible values of θ for the one that makes our observed outcomes most likely to have occurred.
- If θ is far away from this estimate, it will be unlikely for us to have observed the data sample we got.
- However, if θ is close to this estimate, it is much more plausible for us to see the data we did.

We can take a similar approach for the third example, provided we use the probability density function in place of the actual probabilities of obtaining the observed values (since they will always be zero).

Likelihood, I

To formalize the procedure we just described, we first define the likelihood of obtaining the observed data as a function of the parameter θ:

Definition

Suppose the values $x_1, x_2, \ldots, x_n$ are observed by sampling a discrete or continuous random variable X with probability density function $f_X(x;\theta)$ that depends upon an unknown parameter θ. Then the <u>likelihood function</u> $L(\theta) = \prod_i f_X(x_i;\theta)$ represents the probability associated to the observed values x_i.

Likelihood, II

- In the situation where X is a discrete random variable, then under the assumption that all of the samples are independent, the product $f_X(x_1;\theta) \cdot f_X(x_2;\theta) \cdot \cdots \cdot f_X(x_n;\theta)$ represents the probability of obtaining the outcomes $x_1, x_2, \dots, x_n$ from sampling X a total of n times in a row.
- In the situation where X is a continuous random variable, the product represents the probability density of obtaining that sequence of outcomes.
- In either scenario, we think of the likelihood function $L(\theta)$ as measuring the overall probability that we would obtain the observed data by sampling the distribution with parameter θ.

Likelihood, III

Example: If an unfair coin with an unknown probability θ of coming up heads is flipped 10 times, and the results are TTTTT THTTH, find the associated likelihood function $L(\theta)$.

Likelihood, III

Example: If an unfair coin with an unknown probability θ of coming up heads is flipped 10 times, and the results are TTTTT THTTH, find the associated likelihood function $L(\theta)$.

- The values of the probability density function on the 10 flips, respectively, are $1-\theta$, $1-\theta$, $1-\theta$, $1-\theta$, $1-\theta$ and $1-\theta$, θ, $1-\theta$, $1-\theta$, θ.
- Thus, the likelihood function is $L(\theta) = (1-\theta)^6 \cdot \theta \cdot (1-\theta)^2 \cdot \theta = \theta^2(1-\theta)^8$.
- Note here that the likelihood function only depends on the numbers of heads and tails flipped: it would be the same for any other sequence with 8 tails and 2 heads.

Likelihood, IV

Example: If a Poisson distribution with parameter $\lambda = \theta$ is sampled five times and the results are 4, 2, 0, 3, 4, find the associated likelihood function $L(\theta)$.

Likelihood, IV

Example: If a Poisson distribution with parameter $\lambda = \theta$ is sampled five times and the results are 4, 2, 0, 3, 4, find the associated likelihood function $L(\theta)$.

- The value of the pdf on each result are, respectively, $\left[\frac{\theta^4 e^{-\theta}}{4!}\right], \left[\frac{\theta^2 e^{-\theta}}{2!}\right], \left[\frac{\theta^0 e^{-\theta}}{0!}\right], \left[\frac{\theta^3 e^{-\theta}}{3!}\right], \left[\frac{\theta^4 e^{-\theta}}{4!}\right]$.
- Thus, multiplying these together yields

$$\begin{aligned} L(\theta) &= \left[\frac{\theta^4 e^{-\theta}}{4!}\right] \cdot \left[\frac{\theta^2 e^{-\theta}}{2!}\right] \cdot \left[\frac{\theta^0 e^{-\theta}}{0!}\right] \cdot \left[\frac{\theta^3 e^{-\theta}}{3!}\right] \cdot \left[\frac{\theta^4 e^{-\theta}}{4!}\right] \\ &= \frac{\theta^{13} e^{-5\theta}}{6912}. \end{aligned}$$

Likelihood, V

Example: If an exponential distribution with parameter $\lambda = \theta$ is sampled four times and the results are 1.25, 0.02, 0.18, 0.63, find the associated likelihood function $L(\theta)$.

Likelihood, V

Example: If an exponential distribution with parameter $\lambda = \theta$ is sampled four times and the results are 1.25, 0.02, 0.18, 0.63, find the associated likelihood function $L(\theta)$.

- The value of the pdf on each result are, respectively, $[\theta e^{-1.25\theta}]$, $[\theta e^{-0.02\theta}]$, $[\theta e^{-0.18\theta}]$, $[\theta e^{-0.63\theta}]$.
- Thus, multiplying these together yields $L(\theta) = [\theta e^{-1.25\theta}] \cdot [\theta e^{-0.02\theta}] \cdot [\theta e^{-0.02\theta}] \cdot [\theta e^{-0.63\theta}] = \theta^4 e^{-2.08\theta}$.

Maximum Likelihood, I

Our approach now is to compute the value of the unknown parameter that maximizes the likelihood of obtaining the observed data; this is known as the method of maximum likelihood. More explicitly:

Method (Maximum Likelihood)

Suppose the values $x_1, x_2, \ldots, x_n$ are observed by sampling a random variable X with probability density function $f_X(x;\theta)$ that depends upon an unknown parameter θ.
Then a maximum likelihood estimate (MLE) for θ, often written as $\hat{\theta}$ or θ_e, is a value of θ that maximizes the likelihood function $L(\theta) = \prod_i f_X(x_i;\theta)$.

Maximum Likelihood, II

A few comments:

- In principle, there could be more than one value of θ maximizing the function $L(\theta)$. In practice, there is usually a unique maximum, which we refer to as *the* maximum likelihood estimate of θ.
- If $f_X(x_i; \theta)$ is a differentiable function of θ (which is usually the case) then $L(\theta)$ will also be a differentiable function of θ.
- Then, by the usual principle from calculus, any maximum likelihood estimate will be a global maximum of L hence be a root of the derivative $L'(\theta)$.
- Since $L(\theta)$ is a product, to compute the roots of its derivative it is much easier instead to use logarithmic differentation, which amounts to computing the roots of the derivative of its logarithm $\ln L(\theta) = \sum_i \ln[f_X(x_i; \theta)]$, which is called the <u>log-likelihood</u>.

Maximum Likelihood, III

Example: An unfair coin with an unknown probability θ of coming up heads is flipped 10 times, and the results are TTTTT THTTH. Find the maximum likelihood estimate for θ.

Maximum Likelihood, III

Example: An unfair coin with an unknown probability θ of coming up heads is flipped 10 times, and the results are TTTTT THTTH. Find the maximum likelihood estimate for θ.

- Earlier, we computed the likelihood function $L(\theta) = (1-\theta)^6 \cdot \theta \cdot (1-\theta)^2 \cdot \theta = \theta^2(1-\theta)^8$.
- The log-likelihood is $\ln L(\theta) = 2\ln(\theta) + 8\ln(1-\theta)$ with derivative $\frac{d}{d\theta}[\ln L(\theta)] = \frac{2}{\theta} - \frac{8}{1-\theta}$.
- Setting the derivative equal to zero yields $\frac{2}{\theta} - \frac{8}{1-\theta} = 0$ so that $2(1-\theta) = 8\theta$, whence $\theta = 1/5$.
- This yields an estimate $\hat{\theta} = 1/5$. Note that this agrees with our intuitive argument earlier that the most reasonable value of θ is the actual proportion of heads obtained in the sample.

Maximum Likelihood, IV

Example: An unfair coin with an unknown probability θ of coming up heads is flipped 10 times, and the results are TTTTT THTTH. Find the maximum likelihood estimate for θ.

Maximum Likelihood, IV

Example: An unfair coin with an unknown probability θ of coming up heads is flipped 10 times, and the results are TTTTT THTTH. Find the maximum likelihood estimate for θ.

- We could have differentiated $L(\theta)$ directly: $L'(\theta) = 2\theta(1-\theta)^8 - 8\theta^2(1-\theta)^7 = \theta(1-\theta)^7(2-10\theta)$.
- Setting $L'(\theta) = 0$ and solving then yields $\theta = 0, 1, 1/5$.
- Notice that although $\theta = 0$ and $\theta = 1$ are roots of $L'(\theta) = 0$, and hence are critical numbers for $L(\theta)$, they are in fact local minima, whereas $\theta = 1/5$ is a local maximum.
- We implicitly ignored the two values $\theta = 0$ and $\theta = 1$ when analyzing the log-likelihood because these make $\frac{d}{d\theta} \ln L(\theta)$ undefined rather than zero.
- In principle, we should always check that the candidate value actually does yield the *maximum* likelihood, but we will omit such verifications when there is only one possible candidate.

Maximum Likelihood, V

Example: An exponential distribution with parameter θ is sampled five times and the results are 1.25, 0.02, 0.18, 0.63. Find the maximum likelihood estimate for θ.

- We computed the likelihood function $L(\theta) = [\theta e^{-1.25\theta}] \cdot [\theta e^{-0.02\theta}] \cdot [\theta e^{-0.18\theta}] \cdot [\theta e^{-0.63\theta}] = \theta^4 e^{-2.08\theta}$.

Maximum Likelihood, V

<u>Example</u>: An exponential distribution with parameter θ is sampled five times and the results are 1.25, 0.02, 0.18, 0.63. Find the maximum likelihood estimate for θ.

- We computed the likelihood function $L(\theta) = [\theta e^{-1.25\theta}] \cdot [\theta e^{-0.02\theta}] \cdot [\theta e^{-0.18\theta}] \cdot [\theta e^{-0.63\theta}] = \theta^4 e^{-2.08\theta}$.
- The log-likelihood is $\ln L(\theta) = 4 \ln \theta - 2.08\theta$.
- Then $\frac{d}{d\theta}[\ln L(\theta)] = \frac{4}{\theta} - 2.08$.
- This is equal to zero for $\theta = 4/2.08 \approx 1.9231$, so this value is our maximum likelihood estimate.
- Notice that the expected value of the exponential distribution is $1/\theta$. If we set this equal to the observed expected value $2.08/4$, we obtain the maximum likelihood estimate for θ.

Maximum Likelihood, VI

Example: A Poisson distribution with parameter $\lambda = \theta$ representing the number of calls to an emergency help line is sampled five times and the results are 4, 2, 0, 3, 4. Find the maximum likelihood estimate for θ.

- We computed the likelihood function $L(\theta) = \dfrac{\theta^{13} e^{-5\theta}}{6912}$.

Maximum Likelihood, VI

Example: A Poisson distribution with parameter $\lambda = \theta$ representing the number of calls to an emergency help line is sampled five times and the results are 4, 2, 0, 3, 4. Find the maximum likelihood estimate for θ.

- We computed the likelihood function $L(\theta) = \dfrac{\theta^{13} e^{-5\theta}}{6912}$.
- The log-likelihood is $\ln L(\theta) = 13 \ln(\theta) - 5\theta - \ln(6912)$.
- Then $\dfrac{d}{d\theta}[\ln L(\theta)] = \dfrac{13}{\theta} - 5$.
- This is equal to zero for $\theta = 13/5$, so this value is our maximum likelihood estimate.
- Notice that this value $\theta = 13/5$ represents the average number of calls to the help line per hour in the data sample. (Think for yourself why this is a sensible parameter estimate.)

Maximum Likelihood, VII

Example: A normal distribution with mean 0 and standard deviation θ is sampled four times and the results are 2.08, 0.34, -2.65, and 2.28. Find the maximum likelihood estimate for θ.

- The probability density function for this normal distribution is $f_X(x;\theta) = \dfrac{1}{\theta\sqrt{2\pi}} e^{-x^2/(2\theta^2)}$.

Maximum Likelihood, VII

Example: A normal distribution with mean 0 and standard deviation θ is sampled four times and the results are 2.08, 0.34, -2.65, and 2.28. Find the maximum likelihood estimate for θ.

- The probability density function for this normal distribution is $f_X(x;\theta) = \dfrac{1}{\theta\sqrt{2\pi}} e^{-x^2/(2\theta^2)}$.
- To find the log-likelihood it is easier to take the logarithm of $f_X(x;\theta)$ and then sum afterwards.
- We see $\ln f_X(x;\theta) = -\ln(\sqrt{2\pi}) - \ln(\theta) - \dfrac{x^2}{2\theta^2}$.

Maximum Likelihood, VII

Example: A normal distribution with mean 0 and standard deviation θ is sampled four times and the results are 2.08, 0.34, -2.65, and 2.28. Find the maximum likelihood estimate for θ.

- The probability density function for this normal distribution is $f_X(x;\theta) = \dfrac{1}{\theta\sqrt{2\pi}} e^{-x^2/(2\theta^2)}$.
- To find the log-likelihood it is easier to take the logarithm of $f_X(x;\theta)$ and then sum afterwards.
- We see $\ln f_X(x;\theta) = -\ln(\sqrt{2\pi}) - \ln(\theta) - \dfrac{x^2}{2\theta^2}$.
- Now we sum to obtain the log-likelihood: $\ln L(\theta)$

$$= -4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \frac{2.08^2}{2\theta^2} - \frac{0.34^2}{2\theta^2} - \frac{(-2.65)^2}{2\theta^2} - \frac{2.28^2}{2\theta^2}$$

$$= -4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \frac{16.6629}{2\theta^2}.$$

Maximum Likelihood, VIII

<u>Example</u>: A normal distribution with mean 0 and standard deviation θ is sampled four times and the results are 2.08, 0.34, -2.65, and 2.28. Find the maximum likelihood estimate for θ.

- Log-likelihood: $\ln L(\theta) = -4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \dfrac{16.6629}{2\theta^2}$.

Maximum Likelihood, VIII

Example: A normal distribution with mean 0 and standard deviation θ is sampled four times and the results are 2.08, 0.34, -2.65, and 2.28. Find the maximum likelihood estimate for θ.

- Log-likelihood: $\ln L(\theta) = -4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \dfrac{16.6629}{2\theta^2}$.
- Now take the derivative: this gives $\dfrac{d}{d\theta}[\ln L(\theta)] = -\dfrac{4}{\theta} + \dfrac{16.6629}{\theta^3}$.
- Setting this equal to zero and solving yields $\theta = \pm\sqrt{\dfrac{16.6629}{4}} \approx \pm 2.0410$.
- Since the standard deviation is always nonnegative, the maximum likelihood estimate is $\hat{\theta} \approx 2.0410$.

Maximum Likelihood, IX

Example: A normal distribution with mean θ and standard deviation θ is sampled four times and the results are 14, -3, 12, and 8. Find the maximum likelihood estimate for θ.

- The probability density function for this normal distribution is $f_X(x;\theta) = \dfrac{1}{\theta\sqrt{2\pi}} e^{-(x-\theta)^2/(2\theta^2)}$.

Maximum Likelihood, IX

Example: A normal distribution with mean θ and standard deviation θ is sampled four times and the results are 14, -3, 12, and 8. Find the maximum likelihood estimate for θ.

- The probability density function for this normal distribution is $f_X(x;\theta) = \dfrac{1}{\theta\sqrt{2\pi}} e^{-(x-\theta)^2/(2\theta^2)}$.
- Thus, $\ln f_X(x;\theta) = -\ln(\sqrt{2\pi}) - \ln(\theta) - \dfrac{(x-\theta)^2}{2\theta^2}$.
- We then sum the appropriate values to obtain the log-likelihood: $\ln L(\theta) = -4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \dfrac{(14-\theta)^2}{2\theta^2} - \dfrac{(-3-\theta)^2}{2\theta^2} - \dfrac{(12-\theta)^2}{2\theta^2} - \dfrac{(8-\theta)^2}{2\theta^2} = -4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \dfrac{413-62\theta+4\theta^2}{2\theta^2}$.

Maximum Likelihood, X

Example: A normal distribution with mean θ and standard deviation θ is sampled four times and the results are 14, -3, 12, and 8. Find the maximum likelihood estimate for θ.

- We have $\ln L(\theta) = 4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \dfrac{413 - 62\theta + 4\theta^2}{2\theta^2}$.
- Thus, $\dfrac{d}{d\theta}[\ln L(\theta)] = -\dfrac{4}{\theta} + \dfrac{413}{\theta^3} - \dfrac{31}{\theta^2}$. Setting this equal to zero and clearing denominators yields $-4\theta^2 - 31\theta + 413 = 0$ which has roots $\theta = -\frac{59}{4}, 7$.
- Since the standard deviation is always nonnegative, the maximum likelihood estimate is $\theta = 7$.

Maximum Likelihood, X

In some cases, after taking the derivative of the log-likelihood we may be left with an equation that cannot be solved analytically for θ (unlike the examples we just did, where we could always solve explicitly for θ).

In such cases, we must resort to numerical approximation procedures, such as Newton's method, to find the desired root.

Maximum Likelihood, XI

Example: The continuous random variable with probability density function $f_X(x;\theta) = \dfrac{2(\theta - x)}{\theta^2}$ for $0 \le x \le \theta$ is sampled five times, and the results are 1.31, 0.83, 1.19, 0.20, and 0.06. Find the maximum likelihood estimate for θ.

Maximum Likelihood, XI

Example: The continuous random variable with probability density function $f_X(x;\theta) = \dfrac{2(\theta - x)}{\theta^2}$ for $0 \le x \le \theta$ is sampled five times, and the results are 1.31, 0.83, 1.19, 0.20, and 0.06. Find the maximum likelihood estimate for θ.

- We have $\ln f_X(x;\theta) = \ln 2 + \ln(\theta - x) - 2\ln(\theta)$, so now we sum the appropriate values to obtain the log-likelihood: $\ln L(\theta) = 4\ln 2 + \ln(\theta - 1.31) + \ln(\theta - 0.83) + \ln(\theta - 1.19) + \ln(\theta - 0.2) + \ln(\theta - 0.06) - 10\ln(\theta)$.
- The derivative is then $\frac{d}{d\theta}[\ln L(\theta)] = \dfrac{1}{\theta - 1.31} + \dfrac{1}{\theta - 0.83} + \dfrac{1}{\theta - 1.19} + \dfrac{1}{\theta - 0.20} + \dfrac{1}{\theta - 0.06} - \dfrac{10}{\theta}$.
- Setting the derivative equal to zero and clearing denominators yields a polynomial equation of degree 5 in θ, whose roots cannot be easily evaluated.

Maximum Likelihood, XII

Example: The continuous random variable with probability density function $f_X(x;\theta) = \dfrac{2(\theta - x)}{\theta^2}$ for $0 \le x \le \theta$ is sampled five times, and the results are 1.31, 0.83, 1.19, 0.20, and 0.06. Find the maximum likelihood estimate for θ.

Maximum Likelihood, XII

Example: The continuous random variable with probability density function $f_X(x;\theta) = \dfrac{2(\theta - x)}{\theta^2}$ for $0 \le x \le \theta$ is sampled five times, and the results are 1.31, 0.83, 1.19, 0.20, and 0.06. Find the maximum likelihood estimate for θ.

- Using Newton's method or another numerical approximation technique, we can find approximations to the solutions of $$\frac{1}{\theta - 1.31} + \frac{1}{\theta - 0.83} + \frac{1}{\theta - 1.19} + \frac{1}{\theta - 0.20} + \frac{1}{\theta - 0.06} - \frac{10}{\theta} = 0$$ as $\theta \approx 0.0677, 0.2308, 0.9113, 1.2460$, and 2.3307.
- Since one of the observed values was 1.31, and the density function is only nonzero when $0 \le x \le \theta$, we must have $\theta \ge 1.31$. Therefore, the maximum likelihood estimate can only be the largest root, so $\hat{\theta} \approx 2.3307$.

Maximum Likelihood, XIII

It is also possible to perform maximum likelihood estimates for more than one unknown parameter simultaneously.

- The idea is the same as in the single-parameter case: we write down the likelihood function and then attempt to maximize it.
- For a differentiable function of several variables, any local maximum must occur at a point where all partial derivatives of the function are zero.
- As before, since the likelihood function is a product, we usually work instead with the log-likelihood function.
- What this means is that we may find a multi-parameter maximum likelihood estimate by setting all of the partial derivatives of the log-likelihood function equal to zero, and then solving the resulting system of equations for the unknown parameters.

Maximum Likelihood, XIV

Example: A normal distribution with unknown mean μ and standard deviation θ is sampled three times and the results are 1, 5, and -3. Find the maximum likelihood estimate for (μ, θ).

Maximum Likelihood, XIV

Example: A normal distribution with unknown mean μ and standard deviation θ is sampled three times and the results are 1, 5, and -3. Find the maximum likelihood estimate for (μ, θ).

- The probability density function for this normal distribution is $f_X(x; \theta, \mu) = \dfrac{1}{\theta\sqrt{2\pi}} e^{-(x-\mu)^2/(2\theta^2)}$.
- Thus, the log-likelihood function is $\ln L(\theta, \mu) = -3\ln(\sqrt{2\pi}) - 3\ln(\theta) - \dfrac{(1-\mu)^2}{2\theta^2} - \dfrac{(5-\mu)^2}{2\theta^2} - \dfrac{(-3-\mu)^2}{2\theta^2} = -3\ln(\sqrt{2\pi}) - 3\ln(\theta) - \dfrac{35 - 6\mu + 3\mu^2}{2\theta^2}$.

Maximum Likelihood, XIV

Example: A normal distribution with unknown mean μ and standard deviation θ is sampled three times and the results are 1, 5, and -3. Find the maximum likelihood estimate for (μ, θ).

Maximum Likelihood, XIV

Example: A normal distribution with unknown mean μ and standard deviation θ is sampled three times and the results are 1, 5, and -3. Find the maximum likelihood estimate for (μ, θ).

- We have $\ln L(\theta, \mu) = -4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \frac{35-6\mu+3\mu^2}{2\theta^2}$.
- Thus, the two partial derivatives are $\frac{\partial}{\partial\mu}[\ln L(\theta,\mu)] = \frac{6-6\mu}{2\theta^2}$ and $\frac{\partial}{\partial\theta}[\ln L(\theta,\mu)] = -\frac{4}{\theta} + \frac{35-6\mu+3\mu^2}{\theta^3}$.
- Setting the partial derivatives equal to zero and solving yields, respectively, $6 - 6\mu = 0$ so that $\mu = 1$, and $\theta^2 = \frac{35-6\mu+3\mu^2}{4} = 8$ so that $\theta = \pm\sqrt{8}$.
- Since the standard deviation is positive, we obtain the maximum likelihood estimates $\mu = 1$ and $\theta = \sqrt{8}$.

Maximum Likelihood, XV

With more complicated functions of several parameters, the resulting system of equations can be very difficult to solve, even with numerical methods.

- For this reason, certain other methods are used in lieu of maximum likelihood estimates.
- One such method is known as the method of moments.
- This method involves computing the so-called moments $E(X^k)$ for integers $k = 1, 2, \ldots, n$ where n is the total number of unknown parameters, and then setting them equal to the corresponding moments of the sample data.
- The resulting system of equations is often much easier to solve than the system arising from a maximum likelihood estimate.

Maximum Likelihood, XVI

- In many cases, the estimates yielded by the method of moments are a good approximation to those arising from maximum likelihood estimates (and for many common distributions, they are often the same), and can be used as a starting point for approximation methods.
- In the one-parameter case, the method of moments is the same as requiring that the estimate's expected value agrees with the sample's expected value.
- In the two-parameter case, since $\mathrm{var}(X) = E(X^2) - E(X)^2$, it is the same as requiring that the estimate's expected value and variance agree with the sample's expected value and variance.

Estimators, I

Instead of performing a maximum likelihood estimate for each set of sample data, we can instead try to write down a general formula for the estimate in terms of the data values we observe.

- Such a function is an estimator for our parameter of interest θ.
- We typically denote an estimator using a hat: $\hat{\theta}$.
- The estimator $\hat{\theta}$ will be a function of the sample data values $x_1, x_2, \ldots, x_n$.
- We draw a distinction between an estimate and an estimator: an estimate is a numerical value for a specific collection of sample data, while an estimator is a function that provides an estimate for any input collection of sample data.

Estimators, II

Example: A Poisson distribution with parameter θ is sampled n times yielding outcomes $x_1, x_2, \dots, x_n$. Find the maximum likelihood estimator $\hat{\theta}(x_1, \dots, x_n)$ for θ in terms of $x_1, x_2, \dots, x_n$.

Estimators, II

Example: A Poisson distribution with parameter θ is sampled n times yielding outcomes $x_1, x_2, \ldots, x_n$. Find the maximum likelihood estimator $\hat{\theta}(x_1, \ldots, x_n)$ for θ in terms of $x_1, x_2, \ldots, x_n$.

- Since $f_X(x;\theta) = \dfrac{\theta^x e^{-\theta}}{x!}$, taking the logarithm gives $\ln f_X(x;\theta) = x \ln \theta - \theta - \ln(x!)$.
- Then the log-likelihood function is obtained by summing this function over the values $x_1, x_2, \ldots, x_n$.
- Explicitly, $\ln L(\theta) = (x_1 + x_2 + \cdots + x_n) \ln(\theta) - n\theta - \ln(x_1! x_2! \cdots x_n!)$.
- Thus, $\dfrac{d}{d\theta}[\ln L(\theta)] = \dfrac{x_1 + x_2 + \cdots + x_n}{\theta} - n$, which is equal to zero for $\hat{\theta} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$.

Estimators, III

Example: A normal distribution with unknown mean μ and standard deviation σ is sampled sampled n times yielding outcomes $x_1, x_2, \ldots, x_n$. Find the maximum likelihood estimators $\hat{\mu}(x_1, \ldots, x_n)$ and $\hat{\sigma}(x_1, \ldots, x_n)$.

Estimators, III

Example: A normal distribution with unknown mean μ and standard deviation σ is sampled sampled n times yielding outcomes $x_1, x_2, \dots, x_n$. Find the maximum likelihood estimators $\hat{\mu}(x_1, \dots, x_n)$ and $\hat{\sigma}(x_1, \dots, x_n)$.

- Since $f_X(x; \mu, \sigma) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$, the log-likelihood function is $\ln L(\mu, \sigma) = n \ln(\sqrt{2\pi}) - n \ln(\sigma) - \dfrac{(x_1 - \mu)^2 + \cdots + (x_n - \mu)^2}{2\sigma^2}$.
- The partial derivatives are $\dfrac{\partial}{\partial \mu}[\ln L(\mu, \sigma)] = \dfrac{(x_1 - \mu) + \cdots + (x_n - \mu)}{\sigma^2}$ and $\dfrac{\partial}{\partial \sigma}[\ln L(\mu, \sigma)] = -\dfrac{n}{\sigma} + \dfrac{(x_1 - \mu)^2 + \cdots + (x_n - \mu)^2}{\theta^3}$.

Estimators, IV

Example: A normal distribution with unknown mean μ and standard deviation σ is sampled n times yielding outcomes $x_1, x_2, \ldots, x_n$. Find the maximum likelihood estimators $\hat{\mu}(x_1, \ldots, x_n)$ and $\hat{\sigma}(x_1, \ldots, x_n)$.

Estimators, IV

Example: A normal distribution with unknown mean μ and standard deviation σ is sampled n times yielding outcomes $x_1, x_2, \ldots, x_n$. Find the maximum likelihood estimators $\hat{\mu}(x_1, \ldots, x_n)$ and $\hat{\sigma}(x_1, \ldots, x_n)$.

- We have $\frac{\partial}{\partial \mu}[\ln L(\mu, \sigma)] = \frac{(x_1-\mu)+\cdots+(x_n-\mu)}{\sigma^2}$ and $\frac{\partial}{\partial \sigma}[\ln L(\mu, \sigma)] = -\frac{n}{\sigma} + \frac{(x_1-\mu)^2+\cdots+(x_n-\mu)^2}{\theta^3}$.
- Setting the partial derivatives equal to zero and solving yields, respectively, $\mu = \frac{x_1+\cdots+x_n}{n}$ and $\sigma^2 = \frac{(x_1-\mu)^2+\cdots+(x_n-\mu)^2}{n} = \frac{1}{n}(x_1^2 + \cdots + x_n^2) - \left(\frac{x_1+\cdots+x_n}{n}\right)^2$.
- The resulting values of $\hat{\mu}$ and $\hat{\sigma}$ are simply the mean and standard deviation of the outcome set $\{x_1, x_2, \ldots, x_n\}$.

Estimators, V

Example: A uniform distribution on $[0, \theta]$ is sampled n times yielding outcomes $x_1, x_2, \ldots, x_n$. Find the maximum likelihood estimator $\hat{\theta}$ for θ.

Estimators, V

<u>Example</u>: A uniform distribution on $[0, \theta]$ is sampled n times yielding outcomes $x_1, x_2, \ldots, x_n$. Find the maximum likelihood estimator $\hat{\theta}$ for θ.

- The probability density function for this distribution is
$$f_X(x;\theta) = \begin{cases} 1/\theta & \text{for } 0 \le x \le \theta \\ 0 & \text{for other } x \end{cases}.$$
- Therefore, the likelihood function is
$$L(\theta) = \begin{cases} 1/\theta^n & \text{if } x_1, x_2, \ldots x_n \le \theta \\ 0 & \text{otherwise} \end{cases}.$$
- Since $1/\theta^n$ decreases with increasing θ, we can see that the maximum value will occur for the smallest possible θ for which the first condition is satisfied, which is the maximum of $x_1, x_2, \ldots, x_n$.
- Thus, the MLE here is $\hat{\theta} = \max(x_1, x_2, \ldots, x_n)$.

Estimators, VI

The discrete analogue of this last problem is known as the German tank problem.

- During World War II, British intelligence was able to capture numerous components from German tanks, each of which was stamped with its manufacturing number.
- The labels were thus effectively drawn at random from $[0, \theta]$ where θ was the total number of German tanks.
- For obvious reasons, it was of substantial military interest to estimate as precisely as possible the total number θ of enemy tanks; the (quite surprising) result of this calculation shows the largest part number observed is actually a good estimate.
- As a historical matter, the projections obtained by the statisticians analyzing this problem were far more accurate than those obtained by other methods!

Estimators, VII

<u>Example</u>: Pareto's law states that if X denotes an individual's income, then $P(X \geq x) = (k/x)^\theta$ for $x \geq k$, where k is the minimum income for the population and $\theta \geq 1$ is a parameter. Assume k is known.

1. Find the probability density function $p_X(x;\theta)$.
2. If n incomes $x_1, x_2, \ldots, x_n$ are randomly sampled from the population, find the maximum likelihood estimator $\hat{\theta}$ for θ.
3. If $k = \$18\,000$ and the sampled incomes are $52 000, $19 000, $23 000, $55 000, find $\hat{\theta}$ and the expected income.

Estimators, VIII

<u>Example</u>: Pareto's law states that if X denotes an individual's income, then $P(X \geq x) = (k/x)^\theta$ for $x \geq k$, where k is the minimum income for the population and $\theta \geq 1$ is a parameter. Assume k is known.

1. Find the probability density function $p_X(x; \theta)$.

Estimators, VIII

Example: Pareto's law states that if X denotes an individual's income, then $P(X \geq x) = (k/x)^\theta$ for $x \geq k$, where k is the minimum income for the population and $\theta \geq 1$ is a parameter. Assume k is known.

1. Find the probability density function $p_X(x;\theta)$.

- The given information is equivalent to $P(X < x) = 1 - (k/x)^\theta$ for $x \geq k$.
- Since the probability on the left is precisely the one used for the definition of the cumulative distribution function, we can find the pdf by differentiating with respect to x.
- This yields

$$\begin{aligned} p_X(x;\theta) &= \tfrac{\partial}{\partial x}[1 - (k/x)^\theta] \\ &= \theta k^\theta (1/x)^{\theta+1}. \end{aligned}$$

Estimators, IX

Example: Pareto's law states that if X denotes an individual's income, then $P(X \geq x) = (k/x)^\theta$ for $x \geq k$, where k is the minimum income for the population and $\theta \geq 1$ is a parameter. Assume k is known.

2. If n incomes $x_1, x_2, \ldots, x_n$ are randomly sampled from the population, find the maximum likelihood estimator $\hat{\theta}$ for θ.

Estimators, IX

<u>Example</u>: Pareto's law states that if X denotes an individual's income, then $P(X \geq x) = (k/x)^\theta$ for $x \geq k$, where k is the minimum income for the population and $\theta \geq 1$ is a parameter. Assume k is known.

2. If n incomes $x_1, x_2, \ldots, x_n$ are randomly sampled from the population, find the maximum likelihood estimator $\hat{\theta}$ for θ.

- Since $\ln f_X(x; \theta) = \ln \theta + \theta \ln k - (\theta + 1) \ln x$, the log-likelihood function is $\ln L(\theta) = n \ln \theta + n\theta \ln k - (\theta + 1) \ln(x_1 x_2 \cdots x_n)$.
- Thus, $\frac{\partial}{\partial \theta}[\ln L(\theta)] = n/\theta + n \ln k - \ln(x_1 x_2 \cdots x_n)$.
- Setting this equal to zero and solving for θ yields the maximum likelihood estimator $\hat{\theta} = 1/[\frac{1}{n} \ln(x_1 x_2 \cdots x_n) - \ln k]$.

Estimators, X

<u>Example</u>: Pareto's law states that if X denotes an individual's income, then $P(X \geq x) = (k/x)^\theta$ for $x \geq k$, where k is the minimum income for the population and $\theta \geq 1$ is a parameter. Assume k is known.

3. If $k = \$18\,000$ and the sampled incomes are $\$52\,000$, $\$19\,000$, $\$23\,000$, $\$55\,000$, find $\hat{\theta}$ and the expected income.

Estimators, X

Example: Pareto's law states that if X denotes an individual's income, then $P(X \geq x) = (k/x)^{\theta}$ for $x \geq k$, where k is the minimum income for the population and $\theta \geq 1$ is a parameter. Assume k is known.

3. If $k = \$18\,000$ and the sampled incomes are \$52 000, \$19 000, \$23 000, \$55 000, find $\hat{\theta}$ and the expected income.

- We have $\hat{\theta} = 1/[\frac{1}{n}\ln(x_1 x_2 \cdots x_n) - \ln k]$.
- Plugging in the given values yields the estimate $\hat{\theta} = 1.6148$.
- The expected income we can compute by integrating: we have $E(X) = \int_k^{\infty} x\theta k^{\theta}(1/x)^{\theta+1}\,dx = \theta k^{\theta}\int_k^{\infty} x^{-\theta}\,dx = \dfrac{k\theta}{\theta - 1}$.
- Taking $k = \$18\,000$ and $\theta = 1.6148$ yields an expected income of \$47 277, to the nearest dollar.

Estimators, XI

We will remark that the Pareto distribution discussed in the last example has been used to model a number of different types of quantities that display a general feature of having a large number of small items and a smaller number of large ones:

- Wealth of individuals in a population (the particular case from the example).
- Sizes of settlements in a region with low population density
- Sizes of sand particles
- Dollar amounts of losses in insurance claims
- The total amount of playtime of games in a game library

Summary

We outlined the basic motivation of parameter estimation.

We discussed maximum likelihood estimates.

We discussed estimators and found explicit formulas for some maximum likelihood estimators.

Next lecture: Biased and unbiased estimators, efficiency of estimators