Statistics in Geophysics: Inferential Statistics

Steffen Unkel

Department of Statistics
Ludwig-Maximilians-University Munich, Germany

Parameter estimation

- We will be studying problems of statistical inference.
- Many problems of inference have been dichotomized into two areas: estimation of parameters and tests of hypotheses.
- Parameter estimation: Let X be a random variable, whose density is $f_X(x;\theta)$, where the form of the density is assumed known except that it contains an unknown parameter θ.
- The problem is then to use the observed values $x_1, \ldots, x_n$ of a random sample $X_1, \ldots, X_n$ to estimate the value of θ or the value of some function of θ, say $\tau(\theta)$.

Estimator and estimate

- Any statistic $T = g(X_1, \ldots, X_n)$ whose values are used to estimate θ is defined to be an estimator of θ.

- That is, T is a known function of observable random variables that is itself a random variable.

- An estimate is the realized value $t = g(x_1, \ldots, x_n)$ of an estimator, which is a function of the realized values $x_1, \ldots, x_n$.

- Example: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator of a mean μ and $\bar{x}_n$ is an estimate of μ. Here, T is $\bar{X}_n$, t is $\bar{x}_n$ and $g(\cdot)$ is the function defined by summing the arguments and then dividing by n.

Background

- In 1921, R. A. Fisher pointed out an attractive rationale, called maximum likelihood (ML), for estimating parameters.
- This procedure says one should examine the likelihood function of the sample values and take as the estimates of the unknown parameters those values that maximize this likelihood function.
- ML is unifying concept to cover a broad range of problems.
- It is generally accepted as the best rationale to apply in estimating parameters, when one is willing to assume the form of the population probability law is known.

Likelihood function

- If $X_1, \ldots, X_n$ are an i.i.d. sample from a population with pdf or pmf $f(x|\theta)$, the likelihood function is defined by

$$L(\theta) = L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta) \ .$$

- Maximum likelihood principle: Given $x_1, \ldots, x_n$ take as the estimate of θ the value $\hat{\theta}$ that maximizes the likelihood, that is,

$$L(\hat{\theta}) = \max_{\theta} L(\theta) \ .$$

- The value $\hat{\theta}$ that maximizes the likelihood is called the maximum likelihood estimate (MLE) for θ.

Log-likelihood and score function

- It is often more convenient to work with the logarithm of the likelihood function, called the log-likelihood:

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta) \ .$$

- If the log-likelihood is differentiable (in θ), possible candidates for the MLE are the values that solve

$$s(\theta) = \frac{\partial}{\partial\theta} l(\theta) = 0 \ .$$

- The first derivative of the log-likelihood is called the score function.

Example

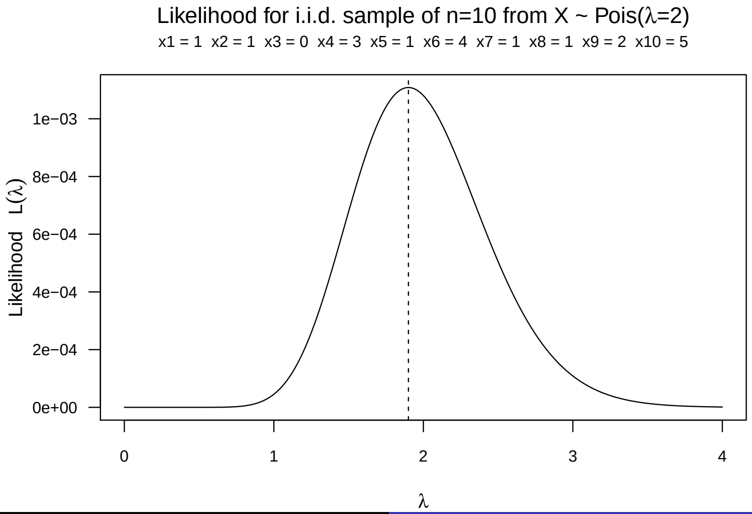
- Let $x_1, \ldots, x_n$ be realizations from $X_i \overset{i.i.d.}{\sim} \mathcal{P}(\lambda)$ $(i = 1, \ldots, n)$ with unknown parameter λ.

- The aim is to estimate λ by maximum likelihood.

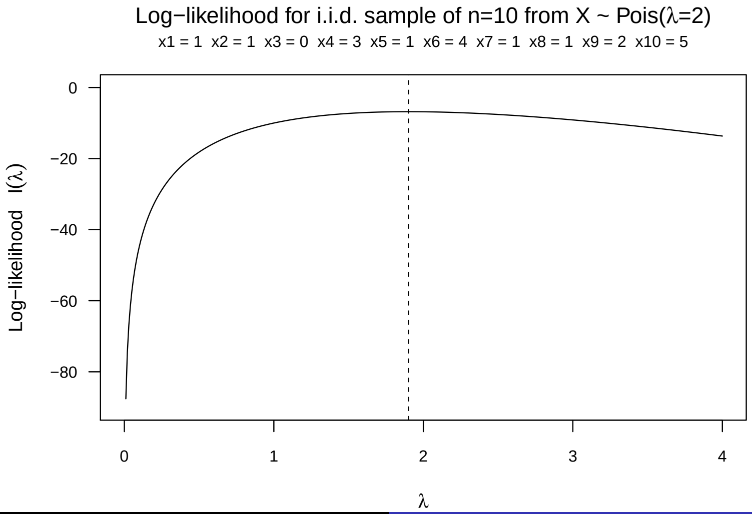
- Likelihood function:

$$\begin{aligned} L(\lambda) &= f(x_1, , \ldots, x_n|\lambda) \\ &= f(x_1|\lambda) \cdots f(x_n|\lambda) \\ &= \prod_{i=1}^{n} f(x_i|\lambda) \\ &= \prod_{i=1}^{n} \left(\frac{\lambda^{x_i}}{x_i!} \exp(-\lambda) \right) \quad . \end{aligned}$$

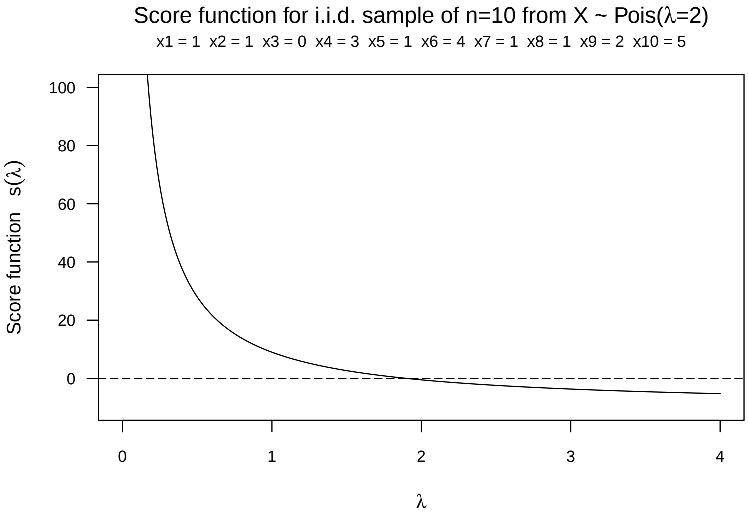
Example


Likelihood for i.i.d. sample of n=10 from X ~ Pois(λ=2)
x1 = 1 x2 = 1 x3 = 0 x4 = 3 x5 = 1 x6 = 4 x7 = 1 x8 = 1 x9 = 2 x10 = 5
Likelihood L(λ)
1e−03
8e−04
6e−04
4e−04
2e−04
0e+00
0
1
2
3
4
λ

Example


Log−likelihood for i.i.d. sample of n=10 from X ~ Pois(λ=2)
x1 = 1 x2 = 1 x3 = 0 x4 = 3 x5 = 1 x6 = 4 x7 = 1 x8 = 1 x9 = 2 x10 = 5
Log−likelihood l(λ)
0
−20
−40
−60
−80
0
1
2
3
4
λ

Example


Score function for i.i.d. sample of n=10 from X ~ Pois(λ=2)
x1 = 1 x2 = 1 x3 = 0 x4 = 3 x5 = 1 x6 = 4 x7 = 1 x8 = 1 x9 = 2 x10 = 5
Score function s(λ)
100
80
60
40
20
0
0
1
2
3
4
λ

Numerical optimization

Newton-Raphson method

- Suppose that we want to approximate the solution to $s(\theta) = 0$.
- Let us also suppose that we have somehow found an initial approximation to this solution, say $\theta^{(0)}$.
- If $\theta^{(k)}$ is an approximation to $s(\theta) = 0$ and if $s'\left(\theta^{(k)}\right) \neq 0$, the next approximation is given by

$$\theta^{(k+1)} = \theta^{(k)} - \frac{1}{s'\left(\theta^{(k)}\right)} \cdot s\left(\theta^{(k)}\right) \quad .$$

- This iterative scheme continues until a prespecified convergence criterion is met.

Other estimation methods

- The method of moments uses sample moments to estimate the parameters of an assumed probability law.
- Least squares estimation minimizes the sum of the squares of the deviations of the observed values and the fitted values.
- Bayesian estimation is based on combining the evidence contained in the data with prior knowledge, based on subjective probabilities, of the values of unknown parameters.

Evaluating estimators

- We have outlined reasonable techniques for finding out estimators of parameters.
- Are some of many possible estimators better in some sense, than others?
- When we are faced with the choice of two or more estimators for the same parameter, it becomes important to develop criteria for comparing them.
- We will now define certain properties, which an estimator may or may not possess, that will help us in deciding whether one estimator is better than another.

Unbiasedness

Definition:

- An estimator $T = g(X_1, \ldots, X_n)$ is defined to be an unbiased estimator of an unknown parameter θ if and only if

$$E(T) = \theta \text{ for all values of } \theta.$$

- The difference $E(T) - \theta$ is called the bias of T and can be either positive, negative, or zero.

- An estimator T of θ is said to be asymptotically unbiased if

$$\lim_{n \to \infty} E(T) = \theta \ .$$

Precision of estimation

- For observations $x_1, \ldots, x_n$ an estimator T yields an estimate $t = g(x_1, \ldots, x_n)$.
- In general, the estimate will not be equal to θ.
- For unbiased estimators the precision of the estimation method is captured by the variance of the estimator, $\mathrm{Var}(T)$.
- The square root of $\mathrm{Var}(T)$ (the standard deviation of T) is called the standard error, which in general has to be estimated itself.

Lower bound for variance

- Let X be a random variable with density $f(x, \theta)$. Under certain regularity conditions:

$$\mathrm{Var}(T) \geq \frac{1}{n\mathrm{E}\left[\left(\frac{\partial}{\partial\theta} \ln f(x, \theta)\right)^2\right]} \quad ,$$

where T is an unbiased estimator of θ.

- The equation above is called the Cramér-Rao inequality, and the right-hand side is called the Cramér-Rao lower bound for the variance of unbiased estimators of θ.

Mean-squared error

Definition:

- The mean-squared error (MSE) of $T = g(X_1, \ldots, X_n)$ (as an estimator for θ) is

$$\mathrm{MSE}(T) = \mathrm{E}[(T-\theta)^2] = \mathrm{Var}(T) + (\mathrm{E}(T) - \theta)^2 \ .$$

- Suppose T is an unbiased estimator of θ, then $\mathrm{MSE}(T) = \mathrm{Var}(T)$.

Consistency

Definition:

- Let $T = g(X_1, \ldots, X_n)$ be an estimator for θ. Then, T is a consistent estimator for θ if

$$\lim_{n\to\infty} \mathrm{P}(|T-\theta| \geq \epsilon) = 0 \text{ for any } \epsilon > 0 \ .$$

- From the Chebyshev inequality we know that

$$\begin{aligned} \mathrm{P}(|T-\theta| \geq \epsilon) &\leq \frac{1}{\epsilon^2}\mathrm{E}[(T-\theta)^2] \\ &= \frac{1}{\epsilon^2}\mathrm{MSE}(T) \ . \end{aligned}$$

- It follows that if $\mathrm{MSE}(T) \to 0$ as $n \to \infty$, then T is consistent.

Efficiency

Definition:

- If T_1 and T_2 are two estimators of θ, then T_1 is more efficient than T_2 if

$$\text{MSE}(T_1) \leq \text{MSE}(T_2) \text{ for any value of } \theta$$

with strict inequality holding somewhere.

- For two unbiased estimators T_1 and T_2 of θ, T_1 is more efficient than T_2 if

$$\text{Var}(T_1) \leq \text{Var}(T_2) \text{ for any value of } \theta$$

with strict inequality holding somewhere.

Interval estimation

- So far, we have dealt with the point estimation of a parameter.
- It seems desirable that a point estimate should be accompanied by some measure of the possible error of the estimate.
- We might make the inference of estimating that the true value of the parameter is contained in some interval.
- Interval estimation: Define two statistics $T_1 = g_1(X_1, \ldots, X_n)$ and $T_2 = g_2(X_1, \ldots, X_n)$, where $T_1 \leq T_2$, so that $[T_1, T_2]$ constitutes an interval for which the probability can be determined that it contains the unknown θ.

Confidence interval

Definition:

- Given a random sample $X_1, \ldots, X_n$ let $T_1 = g_1(X_1, \ldots, X_n)$ and $T_2 = g_2(X_1, \ldots, X_n)$ be two statistics satisfying $T_1 \leq T_2$ for which
$$\mathrm{P}(T_1 \leq \theta \leq T_2) = 1 - \alpha \ .$$
- Then the random interval $[T_1, T_2]$ is called a $(1-\alpha)$-confidence interval for θ.
- $1 - \alpha$ is called the confidence coefficient and T_1 and T_2 are called the lower and upper confidence limits, respectively.
- A value $[t_1, t_2]$, where $t_j = g_j(x_1, \ldots, x_n)$ $(j = 1, 2)$ is an observed $(1-\alpha)$-confidence interval for θ.

One-sided confidence interval

Definition:

- Let $T_1 = -\infty$ and $T_2 = g_2(X_1, \ldots, X_n)$ be a statistic for which

$$P(\theta \leq T_2) = 1 - \alpha \ .$$

- Then T_2 is called a one-sided upper confidence limit for θ.

- Similarly, let $T_2 = \infty$ and $T_1 = g_1(X_1, \ldots, X_n)$ be a statistic for which

$$P(T_1 \leq \theta) = 1 - \alpha \ .$$

- Then T_1 is called a one-sided lower confidence limit for θ.

Confidence intervals for the mean (with known variance)

$100(1-\alpha)$ %-confidence interval for μ (scenario σ^2 known)

- For a normally distributed random variable X:

$$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \quad .$$

- For an arbitrarily distributed random variable X and $n > 30$,

$$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

  is an approximate confidence interval for μ.

- For $0 < p < 1$, z_p is the p-quantile of the standard normal distribution, that is, it is the value for which $F(z_p) = \Phi(z_p) = p$. Hence, $z_p = \Phi^{-1}(p)$.

Confidence intervals for the mean (with unknown variance)

$100(1-\alpha)$ %-confidence interval for μ (scenario σ^2 unknown)

- For a normally distributed random variable X:

$$\left[\bar{X} - t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right] \quad ,$$

  where $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ and $t_{1-\alpha/2}(n-1)$ being the $(1-\alpha/2)$-quantile of the t-distribution with $n-1$ degrees of freedom.

- For an arbitrarily distributed random variable X and $n > 30$,

$$\left[\bar{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

  is an approximate confidence interval for μ.

Confidence intervals for the variance

$100(1-\alpha)$ %-confidence interval for σ^2

- For a normally distributed random variable X:

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\right] \quad ,$$

where $\chi^2_{1-\alpha/2}(n-1)$ and $\chi^2_{\alpha/2}(n-1)$ denote the $(1-\alpha/2)$-quantile and $(\alpha/2)$-quantile, respectively, of the chi-square distribution with $n-1$ degrees of freedom.

Confidence interval for a proportion

$100(1-\alpha)$ %-confidence interval for π

- In dichotomous populations and for $n > 30$, an approximate confidence interval for $\pi = \mathrm{P}(X = 1)$ is given by

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right] \quad ,$$

where $\hat{\pi} = \bar{X}$ denotes the relative frequency.