TECHNISCHE UNIVERSITÄT WIEN

Introduction to Statistics
Statistical Inference
Point Estimation

LV Nr. 105.692
Summer Semester 2021

Probability and Statistics

- Probability theory: start from one probability model, also called a probability distribution, and derive expectations, probabilities, quantiles, etc. for that distribution. In short, given a probability model, describe data from that model.
- Statistics: start from a statistical model, which is a family of probability distributions, collect data assumed to have one of the distributions in the model and derive which distribution that is. In short, given a statistical model and data, infer which distribution in the model is the one for the data.

Statistical Models

- A statistical model is a family of probability distributions.
- A parametric statistical model is a family of probability distributions specified by a finite set of parameters. E.g.: Bernoulli(p), $\mathcal{N}(\mu, \sigma^2)$, etc.
- A nonparametric statistical model is a family of probability distributions too big to be specified by a finite set of parameters.
  - all probability distributions on $\mathbb{R}$
  - all continuous symmetric probability distributions on $\mathbb{R}$
  - all probability distributions on $\mathbb{R}$ having second moments
  - etc.

Statistical Models and submodels

- If $\mathcal{M}$ is a statistical model, it is a family of probability distributions.
- A submodel of a statistical model $\mathcal{M}M$ is a family of probability distributions that is a subset of M.
- If $\mathcal{M}M$ is parametric, then we often specify it by giving the pmf (if the distributions are discrete) or pdf (if the distributions are continuous)

$$\{f_\theta : \theta \in \Theta\}$$

where Θ is the parameter space of the model.

  - We can have models and submodels for nonparametric families as well:
  - All probability distributions on $\mathbb{R}$ is a statistical model.
  - all continuous symmetric probability distributions on $\mathbb{R}$ is a submodel of that.

Statistical Models and Submodels

- Submodels of parametric families are often specified by fixing the values of some parameters.
    - All univariate normal distributions is a statistical model.
    - All univariate normal distributions with known variance is a submodel of that. Its only unknown parameter is the mean. Its parameter space is $\mathbb{R}$.
    - All univariate normal distributions with known mean is a different submodel. Its only unknown parameter is the variance. Its parameter space is $(0, \infty)$.

Parameters

- The word parameter has two closely related meanings in statistics.
  - One of a finite set of variables that specifies a probability distribution within a family. Examples: p for Bernoulli(p), and σ^2 for $\mathcal{N}(\mu, \sigma^2)$.
  - A numerical quantity that can be specified for all probability distributions in the family. Examples: mean, median, variance, 25th quartile.
- The first applies only to parametric statistical models. The parameters are the parameters of the model. The second also applies to nonparametric statistical models.
- The word true has a technical meaning in statistics. In the phrase "true unknown parameter" or "true unknown distribution" it refers to the probability distribution of the data, which is assumed (perhaps incorrectly) to be one of the distributions in the statistical model under discussion.

Statistics

- The word statistic (singular) has a technical meaning in statistics (plural, meaning the subject).
  - A statistic is a function of data only, not parameters. Hence a statistic can be calculated from the data for a problem, even though the true parameter values are unknown.
  - The sample mean $\bar{X}_n$, the sample variance S_n^2, the sample median are all statistics.
  - The random variable
    $$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$
    is NOT a statistic because it contains a parameter (μ).

Estimates

- A statistic X is an estimate of the parameter θ if we declare it as such.
  - The term estimate does not indicate that X has any particular properties. It only indicates our intention to use X to say something about the true unknown value of the parameter θ.
  - There can be many different estimates of a parameter θ
  - The sample mean is an obvious estimate of the population mean μ
  - The sample median can also be used but it may be less representative
  - the sample variance is a nonsensical estimate of μ
- We denote the statistic $\hat{\theta}$, or $\hat{\theta}_n$, as a sample (of size n) based estimate of a parameter θ

Frequentist and Bayesian Statistics

- Frequentist Statistics is based on the empirical distribution of an infinite sequence of iid random variables $X_1, X_2, \ldots$
  - It defines the probability $\mathbb{P}(X_i \in A)$ as what the corresponding expectation for the empirical distribution converges to using the LLN:

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(X_i) \xrightarrow{P} \mathbb{P}(X_i \in A)$$

  - The frequentist theory of statistics uses sampling distributions.
  - If $\hat{\theta}_n$ is an estimate of a parameter θ, then the frequentist theory says that
    - The true value of the parameter θ is an unknown constant. It is not random.
    - An estimate $\hat{\theta}_n$ of this parameter is a random variable and the correct way to describe its randomness is its sampling distribution.

Frequentist Statistics

- Frequentist Statistics is sampling distribution statistics and frequently suffers from the fact that sampling distributions depend on unknown parameters:
  - Suppose we want to estimate the population mean μ using the sample mean $\bar{X}_n$ as an estimate.
  - For large n, we know that

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

    but we don't know μ or σ^2
  - We can estimate both, but now we need to know about the variability of the random vector $\bar{X}_n, \hat{\sigma}_n^2$
  - Frequentist statistics is all about how to deal with this infinite regress. This will consume the rest of the semester.

Bayesian Statistics

Bayesian Statistics is named after its originator Thomas Bayes, whose work was published posthumously in 1764.

- It makes conditional probability the fundamental tool of inference. It takes probability theory as the correct description of all uncertainty.
- If the true value of a parameter θ is unknown, then we are uncertain about it, and the correct description of our knowledge about it or lack thereof is a probability distribution.
  - What the frequentist calls a statistical model, the Bayesian calls a conditional distribution.
  - The frequentist writes $f_\theta(x)$ for the relevant pdf.
  - The Bayesian writes $f(x \mid \theta)$ for the relevant pdf.
  - Both the data x and the parameter θ are random for the Bayesian.

Bayes Rule

- Before we see data, we have a distribution for θ that reflects our knowledge, if any, about it. Suppose the pdf is $f(\theta)$, which is called the prior distribution.
- After we see data, we have a joint distribution:

$$f(x \mid \theta) f(\theta)$$

so that

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{\int f(x \mid \theta) f(\theta) d\theta}$$

that reflects our knowledge about θ after we see x. This is the posterior distribution.

- What is called Bayes rule is the process of finding the posterior.

Bayesian vs Frequentist Statistics

Bayesian Statistics:

- The true value of the parameter θ is unknown so it is assumed to be random
- An estimate $\hat{\theta}_n$ of this parameter is not a random variable after it is seen. The only randomness remaining is in the posterior distribution.

Frequentist Statistics:

- The true value of the parameter θ is an unknown constant feature of a population.
- An estimate $\hat{\theta}_n$ of this parameter is a function of the random data, so it is random.

We **focus on Frequentist Statistics!**

Nuisance Parameters

- Some parameters are more important than others. Which parameters are more important depends on the context.
- The most important parameter or parameters is called parameter of interest.
- The other parameter or parameters is called nuisance parameter.
- Example: If $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2, but we are interested in μ, then σ^2 is considered a nuisance parameter.

Theory of Estimation

- A probability space (also called a *probability model*) is a triple $(\Omega, \mathcal{F}, P)$, where Ω is a sample space, $\mathcal{F}$ is a sigma algebra of events (subsets of Ω) and P is a probability function defined on $\mathcal{F}$.
- A triplet $(\Omega, \mathcal{F}, \mathcal{P})$, where $\mathcal{P}$ is a collection of probability functions defined on $\mathcal{F}$ is called a *statistical experiment* or a *statistical model*.
- Namely, $\mathcal{P}$ is a collection of probability functions such that $(\Omega, \mathcal{F}, P)$ is a probability space for each $P \in \mathcal{P}$.
- We are going to consider only *parametric* statistical models, i.e. those with $\mathcal{P}$ of the form $\{P_\theta,\ \theta \in \Theta \subseteq \mathbb{R}^k\}$, where θ is a parameter, which takes values in the *parameter space* Θ, a subset of $\mathbb{R}^k$.

Theory of Estimation

- We assume that the elementary outcomes are vectors of real numbers and that these outcomes are realizations of a collection of i.i.d. random variables, i.e. each outcome is of the form $x = (x_1, \ldots, x_n)$, where each x_i, $i = 1, \ldots, n$ is a realization of a random variable X_i and the random variables $X_1, \ldots, X_n$ are a random sample from some distribution with cdf $F_\theta(x)$, where $\theta \in \Theta$ is unknown and constant.
- Under these assumptions, each of the probability functions P_θ appearing in the definition of a parametric statistical model is uniquely determined by the corresponding cdf $F_\theta(x)$, i.e. there is a one-to-one correspondence between the collection $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and the associated family $\{F_\theta(\cdot) : \theta \in \Theta\}$ of marginal cdfs.

Point estimators

Definition

Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F_\theta(x)$, where θ is an unknown parameter. A *point estimator* is any statistic $T(X_1, \ldots, X_n)$.

- A point estimator of a parameter θ is a random variable and usually it is denoted with a hat $\hat{\theta}$.
- A realized value of a (point) estimator is called a (*point*) *estimate*. The quantity we are trying to estimate (typically θ) is called the *estimand*.
- It is not required that the range of the estimator coincides with the range of the estimand, i.e. $T(X_1, \ldots, X_n) \in \Theta$ is not required.
- On the other hand, an estimator $T(X_1, \ldots, X_n)$ of θ is a *good* estimator (only) if it is *close* to θ in some probabilistic sense and this will typically require $T(X_1, \ldots, X_n) \in \Theta$.

Theory of Estimation

- Some estimates are better than others. One of the main themes of frequentist statistics is the properties of estimates that make one better than another.
- We need theory to help choose among estimates:
- Suppose the statistical model under consideration is the family of all probability distributions on $\mathbb{R}$ that are symmetric and have first moments. The parameter of interest is the center of symmetry, which is both the mean and the median.
- Both the sample mean and sample median are obvious estimators of this parameter. Which is better? According to what criteria? Under what conditions?

Bias and Unbiasedness

- If T is an estimator of a parameter θ, then we say T is unbiased if

$$\mathbb{E}_\theta(T) = \theta), \quad \text{for all } \theta \in \Theta$$

  where Θ is the parameter space of the statistical model under consideration.
- The notation $\mathbb{E}_\theta$ denotes the expectation operator for the distribution with parameter value θ.
- If an estimator is not unbiased, then it is biased.
- The bias of an estimator T of a parameter θ is

$$b(\theta) = \mathbb{E}_\theta(T) - \theta$$

Mean Square Error

The mean square error (MSE) of an estimator T of a parameter θ is

$$MSE_\theta(T) = \mathbb{E}_\theta\left((T-\theta)^2\right)$$

We have seen that

$$MSE_\theta(T) = \mathbb{V}\mathrm{ar}_\theta(T) + b^2(\theta)$$

i.e.,

$$\mathrm{MSE} = \mathrm{variance} + \mathrm{bias}^2$$

Bias-Variance Trade-off

- **MSE** is a sensible measure of goodness of an estimator.
- MSE provides a justification why unbiasedness in not necessarily good.
- It says that there is a bias-variance trade-off: You can make bias small only by increasing variance and vice versa.
- The only way you can make bias zero is to make the variance of the estimator very large or even infinite, which is not a good trade.

Bias-Variance Trade-off: Example

Consider the following estimators of a population variance:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

$$\widetilde{S}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

$$S_n^\star = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

- S_n^2 is unbiased
- $\widetilde{S}_n$ is the variance of the empirical distribution and has nice asymptotic properties because of that
- $S_n^\star$ minimizes the MSE if the data are iid normal (next)
- Thus, the unbiased estimator is not best if MSE is the criterion. This is because of the bias-variance trade-off: you need some bias to reduce the variance sufficiently to get the smallest MSE.

Bias-Variance Trade-off: Example

Let $X_i \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. Consider the two estimators of σ^2,

$$\tilde{S}_n^2 = \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

and the sample variance

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

The sample variance satisfies

$$\frac{(n-1)S_n^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi_{n-1}^2$$

Bias-Variance Trade-off: Example

implying

$$\begin{aligned}
\mathbb{E}(S_n^2) &= \frac{\sigma^2}{n-1} \cdot \mathbb{E}\Big(\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2\Big) \\
&= \frac{\sigma^2}{n-1} \cdot (n-1) = \sigma^2 \qquad \text{and} \\
\mathbb{V}\text{ar}(S_n^2) &= \Big(\frac{\sigma^2}{n-1}\Big)^2 \cdot \mathbb{V}\text{ar}\Big(\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2\Big) \\
&= \Big(\frac{\sigma^2}{n-1}\Big)^2 \cdot 2(n-1) = \frac{2}{n-1}\,\sigma^4.
\end{aligned}$$

Bias-Variance Trade-off: Example

Thus,

$$MSE_{(\mu,\sigma^2)}(S_n^2) = \frac{2}{n-1}\,\sigma^4.$$

Similarly,

$$\begin{aligned}
\mathbb{E}(\tilde{S}_n^2) &= \mathbb{E}\left(\frac{n}{n-1}\cdot S_n^2\right) = \frac{n}{n-1}\cdot\sigma^2 \qquad \text{and} \\
\mathbb{V}\mathrm{ar}\,(\tilde{S}_n^2) &= \mathbb{V}\mathrm{ar}\left(\frac{n}{n-1}\cdot S_n^2\right) = \left(\frac{n}{n-1}\right)^2\cdot\mathbb{V}\mathrm{ar}\,(S_n^2) \\
&= \frac{n^2}{(n-1)^2}\cdot\frac{2}{n-1}\,\sigma^4 = \frac{2(n-1)}{n^2}\,\sigma^4
\end{aligned}$$

Bias-Variance Trade-off: Example

Then,

$$MSE_{(\mu,\sigma^2)}(\tilde{S}_n^2) = \frac{2(n-1)}{n^2}\,\sigma^4 + \left(\frac{\sigma^2}{n}\right)^2 = \frac{2n-1}{n^2}\sigma^2$$
$$= \frac{1-\frac{1}{n}}{n}\,\sigma^4 \;<\; \frac{2}{n-1}\,\sigma^4 = MSE_{(\mu,\sigma^2)}(S_n^2).$$

1. The estimator S_n^2 is unbiased, while $\tilde{S}_n^2$ is biased.
2. However, the variance of $\tilde{S}_n^2$ is much smaller that the variance S_n^2, i.e. $MSE(\tilde{S}_n^2) < MSE(S_n^2)$ for all values μ and σ^2.
3. Hence, $\tilde{S}_n^2$ is the better estimate in the sense that it has smaller deviation from the unknown parameter σ^2.

Consistent and Asymptotically Normal

A statistic $\hat{\theta}_n$ is a consistent and asymptotically normal (CAN) estimator of a parameter θ if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \tau^2).$$

for some constant τ^2 that may have nothing to do with the corresponding population variance.

The constant τ^2 is called the asymptotic variance of the CAN estimator $\hat{\theta}_n$.

Examples

- $\bar{X}_n$ is a CAN estimate of the population mean μ with asymptotic variance equal to the population variance σ^2.
- The sample median is a CAN estimate of the population median m. Its asymptotic variance is $1/(4f(m)^2)$, where f is the population pdf.
- S_n^2 is a CAN estimate of the population variance σ^2. Its asymptotic variance is $\mu_4 - \sigma^4$, where μ_4 is the population fourth central moment.

Asymptotic Relative Efficiency

- The asymptotic relative efficiency (ARE) of two CAN estimators of the same parameter is the ratio of their asymptotic variances.
- It is a sensible measure of goodness of an estimator, because if τ^2 is the asymptotic variance of $\hat{\theta}_n$, this means

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta, \frac{\tau^2}{n}\right)$$

- If $\hat{\theta}_{n1}$ and $\hat{\theta}_{n2}$ are two CAN estimators of the same parameter but different asymptotic variances τ_1^2 and τ_2^2 and different sample sizes n_1 and n_2 and the actual variances are approximately equal

$$\frac{\tau_1^2}{n_1} \approx \frac{\tau_2^2}{n_2}$$

then

$$\frac{\tau_1^2}{\tau_2^2} \approx \frac{n_1}{n_2}$$

Asymptotic Relative Efficiency

- The ARE is approximately the ratio of sample sizes needed to get the same accuracy, because variance measures the spread of the sampling distribution of an estimator.
- If the data cost is proportional to sample size, then ARE is the correct measure of relative cost to get the same accuracy.
- If $\hat{\theta}_{n1}$ and $\hat{\theta}_{n2}$ are two CAN estimators of the same parameter, then the one with the smaller asymptotic variance is better.
- The ARE depends on the population distribution. You have to calculate the ARE for each population distribution you are interested in.

Example: Mean vs Median

Let X_i be iid $\mathcal{N}(\mu, \sigma^2)$. Then, the sample mean

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

with asymptotic variance σ^2
Also, the sample median is a CAN estimator of μ. Its asymptotic variance is

$$\frac{1}{4f(\mu)^2} = \frac{1}{4\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^2} = \frac{\pi\sigma^2}{2}$$

Since ,

$$\sigma^2 < \frac{\pi\sigma^2}{2}$$

the sample mean is the better estimator, with

$$ARE(\overline{X}_n, \widetilde{X}_n) = \frac{2}{\pi}$$

Method of moments

Method of moments estimators are found by equating the first k sample moments to the corresponding k population moments, and solving the resulting system of equations.

Let $X_1, \ldots, X_n$ be a sample from a population with cdf $F_\theta(x)$, i.e. with pdf $f_\theta(x)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$ is a vector of unknown parameters. For any $j = 1, \ldots, k$ let $\mu_j : \Theta \to \mathbb{R}$ be defined by

$$\mu_j(\theta) = \int_{\mathbb{R}} x^j \, f_\theta(x) \, dx, \qquad \theta \in \Theta.$$

A *method of moments estimator* $\hat{\theta}$ of θ solves the *estimating equations*

$$\frac{1}{n} \sum_{i=1}^{n} X_i^j = \mu_j(\hat{\theta}), \qquad j = 1, \ldots, k. \tag{1}$$

Why Method of Moments?

- Every empirical ordinary or central moment of order k is a consistent estimator of the corresponding population moment, assuming that population moments of order k exist
- Every empirical ordinary or central moment of order k is a CAN estimator of the corresponding population moment, assuming that population moments of order $2k$ exist
- These empirical moments are jointly consistent or jointly CAN, although we have not proven this.

Method of Moments (cont.)

1. If there are p unknown parameters, choose p moments and evaluate them
2. This gives p equations giving moments as a function of parameters.
3. Solve these equations for the parameters to obtain p equations giving parameters as a function of moments.
4. Plug in empirical moments for population moments. This gives estimates of the parameters as a function of empirical moments.
5. Derive the asymptotic distribution of the estimators using the delta method.

Examples

1. Let $X_i \sim Bernoulli(p)$, where $p \in [0,1]$ is an unknown parameter. In this case, $\theta = p$, $\Theta = [0,1]$ and

$$\mu(p) = p, \qquad 0 \leqslant p \leqslant 1.$$

Therefore, the method of moments (MM) estimator of p is

$$\hat{p} = \bar{X}.$$

And we already know its asymptotic distribution

$$\hat{p} \approx \mathcal{N}\left((p, \frac{p(1-p)}{n}\right)$$

Examples

1. Let X_i be i.i.d. uniform $(0, a)$, where $a > 0$ is unknown parameter. We have

$$\mu(a) = \frac{a}{2}, \qquad a > 0.$$

The MM estimator $\hat{a}$ is found by solving the equation $\mu(\hat{a}) = \bar{X}$, i.e. $\frac{\hat{a}}{2} = \bar{X}$. Thus,

$$\hat{a} = 2\bar{X}.$$

Notice that the method of moments estimator is not a good estimator. Even though a is unknown, we know that $X_i > a$ is impossible when $X_i \sim$ uniform $(0, a)$. It is possible to have $X_i > \hat{a}$ for some i, (e.g. if $n = 3$ and $X_1 = X_2 = 2$ and $X_3 = 8$), so it seems that a *better* estimator can be constructed.

Examples (ctd.)

Let $X_i \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. In this case, $\theta = (\mu, \sigma^2)$, $\Theta = \mathbb{R} \times (0, \infty)$ and the functions μ_1 and μ_2 are given by

$$\mu_1(\mu, \sigma^2) = \mu$$
$$\mu_2(\mu, \sigma^2) = \sigma^2 + \mu^2.$$

The method of moments estimator $(\hat{\mu}, \hat{\sigma}^2)$ solves the following system

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu_1(\hat{\mu}, \hat{\sigma}^2) = \hat{\mu}$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu_2(\hat{\mu}, \hat{\sigma}^2) = \hat{\sigma}^2 + \hat{\mu}^2$$

From the first equation we obtain $\hat{\mu} = \bar{X}$. Using this relation, $\hat{\sigma}^2$ is obtained from the second equation, i.e.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2 = \tilde{S}^2.$$

Failure of MME

In examples considered so far, there is a unique solution $\hat{\theta} \in \Theta$ of the system (1) of estimating equations. The following example shows that the method of moments can break. In such cases, some variant of the method of moments may work.

Consider $X_i \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 > 0$ is an unknown parameter. In this case, $\theta = \sigma^2$ and $\Theta = (0, +\infty)$. The function μ_1 is given by $\mu_1(\sigma^2) = 0$. The equation

$$\bar{X} = \mu_1(\hat{\sigma}^2) = 0$$

has infinitely many solutions when $\bar{X} = 0$ and no solution when $\bar{X} \neq 0$.
On the other hand, the sample counterpart of the equation

$$\mathbb{E}(X^2) = \mu_2(\sigma^2) = \sigma^2$$

has a unique solution

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$

Generalizing this example, let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F_\theta(x)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$ is a vector of unknown parameters.
Even if we cannot find a unique solution $\hat{\theta}$ of the estimating equations (1), we may be able to find functions $g_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \ldots, k$ such that the system of equations

$$\frac{1}{n} \sum_{i=1}^{n} g_j(X_i) = \int_{\mathbb{R}} g_j(x)\, f(x|\hat{\theta})\, dx, \qquad j = 1, \ldots, k, \tag{2}$$

has a unique solution $\hat{\theta} \in \Theta$.
Estimators $\hat{\theta}$ constructed in this way are also called *method of moments* estimators.

MME

[HW]

1. Let $X_1, \ldots, X_n$ be i.i.d. $Bin(k, p)$, where both k and p are unknnown. Find the method of moments estimators of k and p.
2. Let $X_1, \ldots, X_n$ be a random sample from $Gamma(\alpha, \beta)$. Find the method of moments estimators of α and β and derive their joint asymptotic distribution.

Maximum Likelihood Estimation

The method of moments is a very general method of estimation
Another estimation method is maximum likelihood:

1. Suppose we have a parametric statistical model specified by a pmf or pdf, f_θ.
2. The pmf or pdf considered as a function of the unknown parameter or parameters rather than of the data is called the likelihood function

$$L(\theta) = f_\theta(x)$$

3. Although $L(\theta)$ also depends on the data x, we suppress this in the notation.
4. If the data are considered random, then $L(\theta)$ is a random variable, and the function L is a random function.
5. If the data are considered nonrandom, as when the observed value of the data is plugged in, then $L(\theta)$ is a number, and L is an ordinary mathematical function.

Likelihood

Definition

Let $X = (X_1, \ldots, X_n)$ be a discrete (continuous) n-dimensional random vector with joint pmf (pdf) $f_\theta(x) : \mathbb{R}^n \to [0, +\infty)$, where $\theta \in \Theta$ is an unknown parameter vector. For any $x = (x_1, \ldots, x_n)$, the *likelihood function* given x is the function $L(\cdot|x) : \Theta \to [0, +\infty)$ of the form

$$L(\theta \mid x) = L(\theta \mid x_1, \ldots, x_n) = f_\theta(x), \qquad \theta \in \Theta. \tag{3}$$

The *log likelihood function* given x is the function $\ell(\cdot|x) : \Theta \to \mathbb{R}$ given by

$$\ell(\theta|x) = \ell(\theta|x_1, \ldots, x_n) = \log L(\theta|x), \qquad \theta \in \Theta.$$

We may drop multiplicative terms not containing unknown parameters from the likelihood function. If

$$L(\theta) = h(x)g(x, \theta)$$

we can drop $h(x)$ and write

$$L(\theta) = g(x, \theta)$$

instead.

Likelihood

When $X_1, \ldots, X_n$ is a random sample from a discrete (continuous) distribution with pmf (pdf) $f_\theta(x) = f(x \mid \theta)$ the likelihood function given $x = (x_1, \ldots, x_n)$ is

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta), \qquad \theta \in \Theta,$$

while the log likelihood function given x is

$$\ell(\theta|x) = \log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i \mid \theta), \qquad \theta \in \Theta.$$

Examples

Suppose $X \sim Bin(n, p)$. Then the likelihood is

$$L_n(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

but we may, if we like, drop the term that does not contain the parameter, so

$$L_n(p) = p^x (1-p)^{n-x}$$

The log likelihood is

$$\ell_n(p) = x \log(p) + (n-x) \log(1-p)$$

Examples

Suppose X_i are iid $\mathcal{N}(\mu, \nu)$. Then the likelihood is

$$\begin{aligned} L_n(p) &= \prod_{i=1}^{n} f(x_i \mid \theta) \\ &= (2\pi)^{-n/2} \nu^{-n/2} \exp\left(-\frac{1}{2\nu} \sum_{i=1}^{n} (x_i - \mu)^2 \right) \end{aligned}$$

but we may, if we like, drop the term that does not contain the parameter, so

$$L_n(p) = \nu^{-n/2} \exp\left(-\frac{1}{2\nu} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

The log likelihood is

$$\ell_n(\mu, \nu) = -\frac{n}{2} \log(\nu) - \frac{1}{2\nu} \sum_{i=1}^{n} (x_i - \mu)^2$$

Examples

We can further simplify this using

$$\sum_{i=1}^{n}(x_i-\mu)^2=\tilde{s}_n^2+(\bar{x}_n-\mu)^2$$

where $\tilde{s}_n^2=\sum_{i=1}^{n}(x_i-\bar{x}_n)/n$. Then,

$$\ell_n(\mu,\nu)=-\frac{n}{2}\log(\nu)-\frac{n\tilde{s}_n}{2\nu}-\frac{n(\bar{x}_n-\mu)^2}{2\nu}$$

Maximum Likelihood Estimation

The maximum likelihood estimate (MLE) of an unknown parameter θ (which may be a vector) is the value of θ that maximizes the likelihood in some sense. Formally,

Definition

Let $X_1, \ldots, X_n$ be a random sample from a discrete (continuous) distribution with pmf (pdf) $f(\cdot|\theta)$, where $\theta \in \Theta$ is an unknown parameter vector. When $X = (X_1, \ldots, X_n) = x$, a *maximum likelihood estimate* $\hat{\theta}(x)$ of θ satisfies

$$L(\hat{\theta}(x)|x) = \max_{\theta \in \Theta} L(\theta|x),$$

where $L(\cdot|x)$ is the likelihood function given x. The estimator $\hat{\theta}(X)$ is a *maximum likelihood estimator* (MLE) of θ.

Maximum Likelihood Estimation (cont.)

- It is hard to find the global maximizer of the likelihood.
- A local maximizer is often used and also called an MLE. The global maximizer can behave badly or fail to exist when the right choice of local maximizer can behave well.
- $\hat{\theta}_n$ is a global maximizer of L_n if and only if it is a global maximizer of ℓ_n. Same with local replacing global.

Information inequality

Theorem

Let X be a discrete (continuous) random variable with pmf (pdf) f_0 and let f_1 be any other pmf (pdf). Then

$$\mathbb{E}\left(\log f_0(X)\right) \geqslant \mathbb{E}\left(\log f_1(X)\right).$$

The information inequality is strict unless $P(f_0(X) = f_1(X)) = 1$.

Proof: We have to show that $\mathbb{E}(\log Y) \leqslant 0$, for

$$Y = \begin{cases} \frac{f_1(X)}{f_0(X)}, & \text{for } X \in \mathcal{X} \\ 0, & \text{for } X \notin \mathcal{X} \end{cases},$$

where $\mathcal{X} = \{x : f_0(x) > 0\}$ is the support of X.

Information inequality

If Y is a random variable with $\mathbb{P}(Y \geqslant 0) = 1$, then by Jensen's inequality

$$\mathbb{E}\left(\log Y\right) \leqslant \log\left(\mathbb{E}(Y)\right).$$

Then, if X is discrete we obtain

$$\mathbb{E}(Y) = \sum_{x \in \mathcal{X}} \frac{f_1(x)}{f_0(x)} \cdot f_0(x) = \sum_{x \in \mathcal{X}} f_1(x) \leqslant \sum_{x \in \mathbb{R}} f_1(x) = 1,$$

while if X is continuous similarly we obtain

$$\mathbb{E}(Y) = \int_{\mathcal{X}} \frac{f_1(x)}{f_0(x)} \cdot f_0(x)\, dx = \int_{\mathcal{X}} \cdot f_1(x)\, dx \leqslant \int_{\mathbb{R}} \cdot f_1(x)\, dx = 1.$$

In both cases, $\mathbb{E}(Y) \leqslant 1$ and from Jensen's inequality we obtain

$$\mathbb{E}\left(\log Y\right) \leqslant \log\left(\mathbb{E}(Y)\right) \leqslant \log 1 = 0,$$

which completes the proof.

Let $X_1, \ldots, X_n$ be a random sample from a discrete (continuous) distribution with pmf (pdf) $f(\cdot|\theta)$, where $\theta \in \Theta$ is an unknown parameter vector. From the information inequality it follows that

$$\mathbb{E}\left(\log f(X|\theta)\right) \geqslant \mathbb{E}\left(\log f(X|\theta^*)\right)$$

for any $\theta^* \in \Theta$, where $\mathbb{E}$ is the expectation computed using the true (unknown) cdf $F(\cdot|\theta)$ of the random variable X. As a consequence, the true parameter value θ solves the problem of maximizing

$$\mathbb{E}\left(\log f(X|\theta^*)\right)$$

with respect to $\theta^* \in \Theta$, i.e.

$$\mathbb{E}\left(\log f(X|\theta)\right) = \max_{\theta^* \in \Theta} \mathbb{E}\left(\log f(X|\theta^*)\right) \tag{4}$$

The sample analogue of this problem is that of maximizing the *average log likelihood*

$$\frac{1}{n}\sum_{i=1}^{n}\log f(X_i|\theta^*)$$

with respect to $\theta^* \in \Theta$. The average log likelihood is a strictly increasing function of $L(\theta^*|X_1,\ldots,X_n)$ and particularly

$$\frac{1}{n}\sum_{i=1}^{n}\log f(X_i|\theta^*) = \log\left(\frac{1}{n}L(\theta^*|\,X_1,\ldots,X_n)\right).$$

Thus, a *maximum likelihood estimator* $\hat{\theta}(X_1,\ldots,X_n)$ maximizes the average log likelihood with respect to θ^*

$$\frac{1}{n}\sum_{i=1}^{n}\log f(X_i|\hat{\theta}(X_1,\ldots,X_n)) = \max_{\theta^*\in\Theta}\left(\frac{1}{n}\sum_{i=1}^{n}\log f(X_i|\theta^*)\right).$$

Local and Global Maxima

Suppose W is an open interval of $\mathbb{R}$ and $f : W \to \mathbb{R}$ is a differentiable function. From calculus, a necessary condition for a point $x \in W$ to be a local maximum of f is

$$f'(x) = 0$$

Also, if f is twice differentiable and $f'(x) = 0$ holds, and also

$$f''(x) \leqslant 0$$

then x is a local maximum.
If, instead

$$f''(x) \geqslant 0$$

then x is a local minimum.

Conditions for global maxima are, in general, very difficult. Every known procedure requires exhaustive search over many possible solutions.
One exception: Concavity.

Concavity

Suppose W is an open interval of $\mathbb{R}$ and $f : W \to \mathbb{R}$ is a twice differentiable function. Then f is *concave* if

$$f''(x) \leqslant 0, \quad \forall x \in W$$

and *strictly concave* if

$$f''(x) < 0, \quad \forall x \in W$$

Many commonly used likelihoods are concave which guarantees global maximizers!

MLE

When a unique maximum likelihood estimator $\hat{\theta}$ of $\theta = (\theta_1, \ldots, \theta_k)$ exists, it can usually be constructed by solving the *likelihood equations*

$$\frac{\partial}{\partial \theta_j} \ell(\theta | X_1, \ldots, X_n) \Big|_{\theta = \hat{\theta}} = 0, \qquad j = 1, \ldots, k, \tag{5}$$

and verifying that a second-order condition holds.

Example: Binomial

The log likelihood is

$$\ell_n(p) = x \log(p) + (n - x) \log(1 - p)$$

with derivatives

$$\begin{aligned}
\ell_n'(p) &= \frac{x}{p} - \frac{n - x}{1 - p} \\
&= \frac{x - np}{p(1 - p)} \\
\ell_n''(p) &= -\frac{x}{p^2} - \frac{n - x}{(1 - p)^2} < 0 \quad \forall p \in (0, 1)
\end{aligned}$$

Example: Binomial

Setting $\ell'(p) = 0$ obtains $p = x/n$. Since $\ell_n''(p) < 0$, for all p, the likelihood is strictly concave and

$$\hat{p}_n = \frac{x}{n}$$

is the unique global maximizer of the log likelihood.

The preceding analysis doesn't work when $\hat{p}_n = 0$ or $\hat{p}_n = 1$ because the log likelihood and its derivatives are undefined when $p = 0$ or $p = 1$.

Binomial (contd.)

Let $X_i \sim Bernoulli(p)$, where $p \in [0, 1]$ is an unknown parameter. Each X_i is discrete with pmf

$$f(x \mid p) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} p^x (1-p)^{1-x}, & x \in \{0, 1\} \\ 0, & \text{otherwise} \end{cases} = p^x (1-p)^{1-x} \cdot \mathbb{1}_{x \in \{0,1\}},$$

where $0^0 = 1$ and $\mathbb{1}_{(\cdot)}$ is the indicator function.
It is sufficient to consider the case where $x_i \in \{0, 1\}$ for $i = 1, \ldots, n$, as the likelihood is zero for all other values of $x = (x_1, \ldots, x_n)$.

Binomial (contd.)

Thus, the likelihood given x is

$$L(p \mid x) = \prod_{i=1}^{n} f(x_i|p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i} \cdot (1-p)^{n-\sum_{i=1}^{n} x_i}, \qquad p \in [0,1].$$

Then the log likelihood given x is

$$\begin{aligned} \ell(p \mid x) = \log L(p \mid x) &= \log \left(p^{\sum_{i=1}^{n} x_i} \cdot (1-p)^{n-\sum_{i=1}^{n} x_i} \right) \\ &= \Big(\sum_{i=1}^{n} x_i \Big) \cdot \log p + \Big(n - \sum_{i=1}^{n} x_i \Big) \cdot \log (1-p), \quad p \in [0,1]. \end{aligned}$$

Binomial (contd.)

1. If $\sum_{i=1}^{n} x_i = 0$ then $L(p \mid x) = (1-p)^n$ is a decreasing function of p, and $p = 0$ maximizes $L(p \mid x)$ with respect to $p \in [0, 1]$.
2. If $\sum_{i=1}^{n} x_i = n$ then $L(p \mid x) = p^n$ and $p = 1$ maximizes $L(p \mid x)$ with respect to $p \in [0, 1]$.
3. If $0 < \sum_{i=1}^{n} x_i < 1$, the maximum likelihood estimate can be found by solving the first-order condition for an interior maximum,

$$\frac{d}{dp}\, \ell(p \mid x) \Big|_{p=\hat{p}} = 0.$$

as done already, to obtain that in all cases

$$\hat{p}_n = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Normal

Let $X_i \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. The marginal pdf of X_i is $f(\cdot|\mu, \sigma^2)$, where

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\,\sigma^2}}\; e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The likelihood given $x = (x_1, \ldots, x_n)$ is

$$\begin{aligned} L(\mu, \sigma^2|x) &= \prod_{i=1}^{n} f(x_i|\mu, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \cdot e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \end{aligned}$$

Normal (ctd.)

The log likelihood given x is of the form

$$\ell(\mu, \sigma^2|x) = \log L(\mu, \sigma^2|x) = \log \left((2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2} \right)$$

$$= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2.$$

The likelihood equations are

$$\frac{\partial}{\partial \mu} l(\mu, \sigma^2 | X_1, \ldots, X_n) \Big|_{(\mu, \sigma^2) = (\hat{\mu}, \hat{\sigma}^2)} = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (X_i - \hat{\mu}) = 0$$

$$\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2 | X_1, \ldots, X_n) \Big|_{(\mu, \sigma^2) = (\hat{\mu}, \hat{\sigma}^2)} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^{n} (X_i - \hat{\mu}) = 0.$$

Normal (ctd.)

The unique solution to these equations is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{X}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2 = \tilde{S}_n^2.$$

The matrix of second partial derivatives at $(\mu, \sigma^2) = (\hat{\mu}, \hat{\sigma}^2)$

$$\begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix}$$

is negative definite, and thus $(\hat{\mu}, \hat{\sigma}^2) = (\bar{x}, \tilde{s}_n^2)$ is a local minimizer of $l(\mu, \sigma^2 \mid x)$.

Normal (ctd.)

Since, $\lim_{|\mu|\to\infty} L(\mu, \sigma^2|x) = 0$ for any $\sigma^2 > 0$ and $\lim_{\sigma^2\to 0} L(\mu, \sigma^2|x) = 0$ for any $\mu \in \mathbb{R}$, we conclude that

$$(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, \tilde{S}_n^2)$$

is the likelihood estimator of (μ, σ^2).

In this example the maximum likelihood estimator also coincides with the method of moments estimator.

Normal (ctd.)

For iid normal data with known mean μ and unknown variance $\nu = \sigma^2$, the log likelihood

$$\ell_n(\nu) = -\frac{n}{2}\log(\nu) - \frac{1}{2\nu}\sum_{i=1}^{n}(x_i - \mu)^2$$

has derivatives

$$\ell_n'(\nu) = -\frac{n}{2\nu} + \frac{1}{2\nu^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\ell_n''(\nu) = \frac{n}{2\nu^2} - \frac{1}{\nu^3}\sum_{i=1}^{n}(x_i - \mu)^2$$

Normal (ctd.)

Setting $\ell_n'(\nu) = 0$ and solving for ν we get

$$\hat{\sigma}_n^2 = \hat{\nu} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Since

$$\begin{aligned}\ell_n''(\hat{\nu}) &= \frac{n}{2\hat{\nu}^2} - \frac{1}{\hat{\nu}^3}\sum_{i=1}^{n}(x_i - \mu)^2 \\ &= -\frac{n}{2\hat{\nu}^2} < 0\end{aligned}$$

this MLE is a local maximizer of the log-likelihood

Normal (ctd.)

But

$$\begin{aligned}\ell_n''(\nu) &= \frac{n}{2\nu^2} - \frac{1}{\nu^3}\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= \frac{n}{2\nu^2} - \frac{n\hat{\nu}}{\nu^3}\end{aligned}$$

is not negative for all data and all $\nu > 0$, so we cannot say the MLE is the unique global maximizer from this analysis.

Uniform: MLE on Boundary of Parameter Space

Let X_i be i.i.d. uniform $[0, a]$, where $a > 0$ is unknown parameter. Each X_i is continuous with pdf

$$f(x|a) = \begin{cases} \frac{1}{a}, & 0 \leqslant x \leqslant a \\ 0, & \text{else} \end{cases} = \frac{1}{\theta} \cdot \mathbb{1}_{(x \in [0,a])}.$$

It suffices to consider the case where $x_i \geqslant 0$ for $i = 1, \ldots, n$, as the likelihood is zero for all other values of $x = (x_1, \ldots, x_n)$. The likelihood given x is

$$\begin{aligned} L(a \mid x) &= \prod_{i=1}^{n} f(x_i|a) = \prod_{i=1}^{n} \frac{1}{a} \mathbb{1}_{(x_i \in [0,a])} \\ &= \frac{1}{a^n} \prod_{i=1}^{n} \mathbb{1}_{(x_i \in [0,a])} = \frac{1}{a^n} \mathbb{1}_{(\max\limits_{1 \leqslant i \leqslant n} x_i \leqslant a)}, \qquad a > 0, \end{aligned}$$

Uniform (ctd.)

In this example the maximum likelihood estimator cannot be constructed by solving the likelihood equations.
We go back to the definition of the maximum likelihood estimator. The likelihood given x is zero for $a < \max_{1 \leqslant i \leqslant n} x_i$ and is a decreasing function of a for $a \geqslant \max_{1 \leqslant i \leqslant n} x_i$.
As a consequence, the maximum likelihood estimator of a is

$$\hat{a} = \max_{1 \leqslant i \leqslant n} X_i = X_{(n)},$$

the largest order statistic.
In this case, the maximum likelihood estimator is different from the method moments estimator $2\bar{X}$.

HW

1. **[HW]** Let $X_1, \ldots, X_n$ be a random sample from the pdf

$$f(x|\theta) = \begin{cases} \frac{\theta}{x^2}, & 0 < \theta \leqslant x \\ 0, & \text{otherwise} \end{cases}.$$

   Estimate θ using both the method of moments and the maximum likelihood. Calculate the means and variances of the two estimators. Which one should be preferred? Justify your answer.

2. **[HW]** Let $X_1, \ldots, X_n$ be i.i.d. with the pdf

$$f(x|\theta) = \begin{cases} \theta\, x^{\theta-1}, & 0 \leqslant x \leqslant 1,\ \theta > 0 \\ 0, & \text{otherwise} \end{cases}.$$

   1. Find the method of moments estimator of θ.
   2. Find the MLE of θ, and show that its variance converges to zero as $n \to \infty$.

HW

1. **[HW]** Let $X_1, \ldots, X_n$ be random sample from a population with pmf

$$P_\theta(X = x) = \theta^x (1-\theta)^{1-x}, \qquad x = 0 \text{ or } x = 1, \qquad 0 \leqslant \theta \leqslant \frac{1}{2}.$$

   1. Find the method of moments estimator amd MLE of θ.
   2. Find the the mean squared errors of each of the estimators.
   3. Which estimator is preferred? Justify your choice.

2. **[HW]** Let $X_1, \ldots, X_n$ be i.i.d. $\exp(\alpha)$ with the pdf

$$f(x|\alpha) = \begin{cases} \alpha e^{-\alpha x}, & x \geqslant 0 \\ 0, & \text{otherwise} \end{cases}.$$

   1. Find the method of moments estimator of α.
   2. Find the MLE of θ, and show that its variance converges to zero as $n \to \infty$.
   3. Consider another estimator $\sqrt{\frac{2}{\bar{X}^2}}$ of α. Which estimator is preferred? Explain your reasoning.

MLE in R

1. In order to do maximum likelihood estimation using the computer we need to write the likelihood function or log likelihood function (usually the latter) as a function in the computer language we are using.
2. Examples: Poisson distribution with unknown parameter λ and normal with parameters μ, σ^2.
3. For these two we have the log-likelihood functions

$$\begin{aligned}\ell(\lambda) &= \sum_{i=1}^{n} y_i \log(\lambda) - n\lambda - \sum_i \log(y_i!) \quad \text{or, ignoring constants} \\ &= \sum_{i=1}^{n} y_i \log(\lambda) - n\lambda\end{aligned}$$

$$\ell(\mu, \sigma^2) = -0.5n \log(2\pi) - 0.5n \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2$$

Optimizing the log-likelihood

1. `R` has several functions that optimize functions, e.g., `nlm` and `optim`
2. We use the `optim` function.
3. `optim` has many arguments, most of which can be ignored. The only arguments that must be supplied are two
   - `f` the function to be minimized
   - `p` a starting value for the variable. The dimension of p agrees with the dimension of the space in which `f` takes values.
   - It helps if we can specify a starting value reasonably close to the solution.
4. `optim (starting values, log-likelihood, data)`
5. Good starting values are hard to find, in general. In our particular problem, a good estimate of λ is the sample mean, which is the method of moments estimator.

Optimizing the Log-Likelihood

```
#the log-likelihood
poisson.lik<-function(mu,y){
  n<-nrow(y)
  logl<-sum(y)*log(mu)-n*mu
  return(-logl)
}

npoint <- 100
x=rpois(npoint,2)
dat=data.frame(x)

optim(1,poisson.lik,y=dat,method="BFGS")
$par
[1] 1.861386

$value
[1] 71.19157

$counts
function gradient
      26        6

$convergence
[1] 0

$message
NULL

optim(1,poisson.lik,y=dat,method="CG")
```

Optimizing the Log-Likelihood

```
### Normal-likelihood

normal.lik1<-function(theta,y){
  mu<-theta[1]
  sigma2<-theta[2]
  n<-nrow(y)
  logl<- -.5*n*log(2*pi) -.5*n*log(sigma2) - (1/(2*sigma2))*sum((y-mu)**2)
  return(-logl)
}

x=data.frame(rnorm(100))

optim(c(0,1),normal.lik1,y=x,method="BFGS")
$par
[1] 0.09550428 1.02343504

$value
[1] 143.052

$counts
function gradient
      17        5

$convergence
[1] 0

$message
NULL
```

Othe R resources

1. Check
   `https://cran.r-project.org/web/packages/EstimationTools/vignettes/maxlogL.pdf`
   for the function `maxlogL` in the `EstimationTools` package