Optimization and Data Science

Lecture 16: Optimization and Statistic

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

Contents

Hypothesis tests

- Let a sample $X_i \sim \mathcal{N}(\mu, \sigma^2), i = 1, \ldots, n$, be given ...
- ... or let us assume that a sample has this distribution.
- We compute the mean as estimator for the expectation μ:

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i,$$

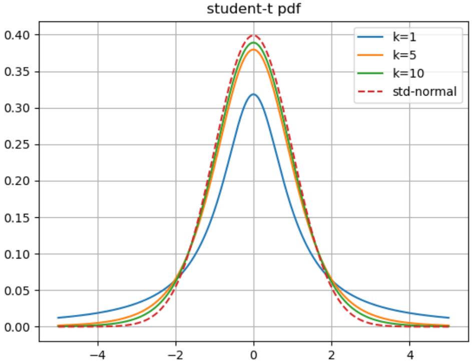
- ... and the estimator for the variance σ^2:

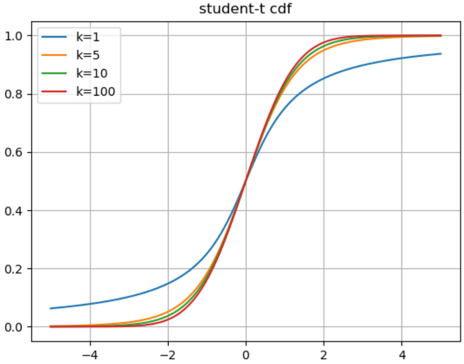
$$e_{\sigma^2} := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

$\rightsquigarrow$ The deviation (scaled with the factor $\sqrt{n/e_{\sigma^2}}$) of the mean $\bar{X}$ from the true expectation μ, is student-$t(n-1)$-distributed:

$$(\bar{X} - \mu)\sqrt{\frac{n}{e_{\sigma^2}}} \sim t(n-1).$$

Student-t-distribution





Confidence intervals for normal-distributed random variables

- We obtained:

$$P\left(-c \le (\bar{X}-\mu)\sqrt{\frac{n}{e_{\sigma^2}}} \le c\right) = 2\int_0^c f_{n-1}(x)dx = 2(F_{n-1}(c) - F_{n-1}(0)) = \gamma,$$

- ... where F_{n-1} is the student-t-cumulative distribution function.

$\rightsquigarrow$ For given γ, we find (using tables or library functions) $c > 0$ such that

$$F_{n-1}(c) = \frac{1}{2}(\gamma + F_{n-1}(0)). \tag{1}$$

$\rightsquigarrow$ Given γ, we find c and the bounds of the two-sided, symmetric confidence intervals

$$\begin{aligned} P\left(-c \le (\bar{X}-\mu)\sqrt{\frac{n}{e_{\sigma^2}}} \le c\right) &= P\left(c\sqrt{\frac{e_{\sigma^2}}{n}} \le \bar{X}-\mu \le c\sqrt{\frac{e_{\sigma^2}}{n}}\right) \\ &= P\left(\mu - c\sqrt{\frac{e_{\sigma^2}}{n}} \le \bar{X} \le \mu + c\sqrt{\frac{e_{\sigma^2}}{n}}\right) = \gamma. \end{aligned}$$

Testing a hypothesis

- Assumed: sample $X_i \sim \mathcal{N}(\mu, \sigma^2), i = 1, \ldots, n$, be given.
- Test the hypothesis that the expectation is μ using the γ-confidence interval:

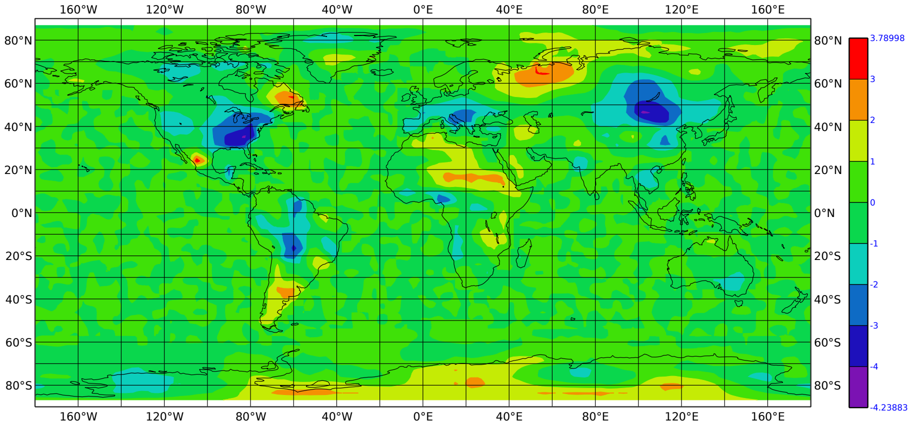
$$P\left(-c \leq (\bar{X} - \mu)\sqrt{\frac{n}{e_{\sigma^2}}} \leq c\right) = \gamma.$$

- Given γ, bound $c = c(\gamma)$ computed form inverse student cdf (1).
- Scaled deviation of the sample mean from μ smaller than $c \rightsquigarrow$ hypothesis true.
- Test hypothesis that a sample $\{X_i, i = 1, \ldots, n\}$ has same mean as another one $\{Y_i, i = 1, \ldots, n\}$: Take mean $\bar{Y}$ instead of μ:

$$P\left(-c \leq (\bar{X} - \bar{Y})\sqrt{\frac{n}{e_{\sigma^2}}} \leq c\right) = \gamma.$$

- Scaled deviation of second sample mean from first one smaller than $c \rightsquigarrow$ hypothesis true.
- Often value $\gamma = 0.95$ is used $\rightsquigarrow$ corresponding value of c: “95 confidence level”.
- Values outside the γ-confidence interval are called **significant** w.r.t. this level.

Example: Test



Values of two-sided t-test for spatially distributed surface temperature of a modified atmosphere climate model (compared to the original version), absolute values below 2.05 are not significant at the 95 confidence level.

Contents

Maximum-likelihood estimator

Definition (Likelihood function and estimator)

Let $\{X_i, i = 1, \ldots, n\}$ be a sample whose distribution depends on a parameter $p \in \mathbb{R}$. The function $L : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, defined as

$$L(p; x) := \prod_{i=1}^{n} P(p; X_i = x_i), \text{ if the } X_i \text{ are discrete,}$$

$$L(p; x) := \prod_{i=1}^{n} f(p; x_i), \text{ if the } X_i \text{ are continuous with density } f,$$

is called **likelihood function**. The **maximum-likelihood estimator** is defined as

$$e(n, X_1, \ldots, X_n) := \operatorname*{argmax}_{p \in \mathbb{R}} L\left(p; (X_i)_{i=1}^{n}\right).$$

- The maximum-likelihood estimate is the value of the parameter p that is most likely for the given sample.

Example: maximum-likelihood estimator for a discrete random variable

- Repeated random experiment with two possible outcomes $\{0, 1\}$.
- Random variable $X = k :\Leftrightarrow k$ times result 1 in n tries.
- Unknown parameter p: probability $\in (0, 1)$ for result 1 in one single try.
- Distribution is binomial distribution:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

- Assume we have one realization. $\rightsquigarrow$ Likelihood function

$$L(p; k) = P(p; X = k).$$

- Maximum-likelihood estimate:

$$\hat{p} = \operatorname*{arg\,max}_{p \in (0,1)} L(p; k) = \operatorname*{arg\,max}_{p \in (0,1)} P(p; X = k) = \operatorname*{arg\,max}_{p \in (0,1)} \log P(p; X = k)$$

since logarithm function is monotone increasing.

Example: maximum-likelihood estimator for a discrete random variable

- We want to maximize the function

$$\begin{aligned}\phi(p) := \log P(p; X = k) &= \log\left(\binom{n}{k} p^k (1-p)^{n-k}\right)\\ &= \log\binom{n}{k} + k\log p + (n-k)\log(1-p).\end{aligned}$$

- Compute the first derivative of the function and apply the first order necessary optimality condition:

$$\phi'(p) = \frac{k}{p} - \frac{n-k}{1-p} = \frac{k(1-p)-(n-k)p}{p(1-p)} = \frac{k-np}{p(1-p)} = 0$$

- ... gives as candidate for a minimizer:

$$p^* = \frac{k}{n}.$$

Example: maximum-likelihood estimator for a discrete random variable

- We want to maximize the function

$$\phi(p) = \log \binom{n}{k} + k \log p + (n-k) \log(1-p).$$

- First derivative:

$$\phi'(p) = \frac{k - np}{p(1-p)} = 0 \quad \Leftrightarrow \quad p = \frac{k}{n} =: p^*.$$

- Compute the second derivative and apply the second order optimality condition:

$$\phi''(p) = \frac{-np(1-p) - (k-np)(1-2p)}{p^2(1-p)^2}, \quad \phi''(p^*) = \frac{-np(1-p)}{p^2(1-p)^2} < 0.$$

$\rightsquigarrow$ $p^* = \frac{k}{n}$ is the maximizer of ϕ and thus the maximum-likelihood estimate for the probability p of getting the value 1 in one try.

Example: maximum-likelihood estimator for a continuous random variable

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2), i = 1, \ldots, n$ be a sample with unknown expectation μ.
- Likelihood funktion (using the rules for the exponential function):

$$L(\mu; x) = \prod_{i=1}^{n} f(\mu; x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

- Because of the strict monotonic growth of the exponential function, we have:

$$\operatorname*{argmax}_{\mu} L(\mu; x) = \operatorname*{argmax}_{\mu} \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right) = \operatorname*{argmin}_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2 = \operatorname*{argmin}_{\mu} \phi(\mu)$$

- We get

$$\phi'(\mu) = -2 \sum_{i=1}^{n} (x_i - \mu) = 0 \Leftrightarrow \mu = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ and } \phi''(\mu) = 2n > 0.$$

$\leadsto$ the maximum likelihood estimator for the expectation μ is the mean.

Contents

Recall simple example: Data-fitting

- Given: data points

$$(t_k, z_k)_{k=1,\ldots,m},\ t_k, z_k \in \mathbb{R}.$$

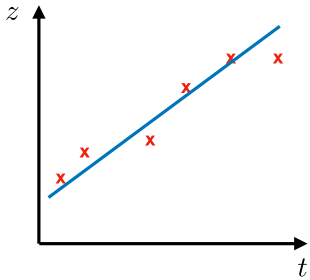
- Task: Find affine-linear function that satisfies

$$y(t_k) = at_k + b \approx z_k, \quad k = 1, \ldots, m.$$

- Minimize distance between points and function:

$$\min_{x=(a,b)} \sum_{k=1}^{m} (y(x; t_k) - z_k)^2,$$

where y depends on x.



- Minimizing the sum of non-negative values means: minimize every term in the sum:

$$\min_{x=(a,b)} (y(x; t_k) - z_k)^2, \quad k = 1, \ldots, m.$$

Recall simple example: Data-fitting

- Minimizing one term in the sum:

$$
\begin{aligned}
\min_x \, (y(x;t_k)-z_k)^2 \;&\Leftrightarrow\; \min_x \frac{(y(x;t_k)-z_k)^2}{2\sigma^2} \qquad \text{for arbitrary } \sigma^2>0,\\
&\Leftrightarrow\; \max_x \left(-\frac{(y(x;t_k)-z_k)^2}{2\sigma^2}\right)\\
&\Leftrightarrow\; \max_x \exp\left(-\frac{(y(x;t_k)-z_k)^2}{2\sigma^2}\right)\\
&\Leftrightarrow\; \max_x \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y(x;t_k)-z_k)^2}{2\sigma^2}\right)
\end{aligned}
$$

- This is the density of the normal distribution with variance σ^2.

$\leadsto$ minimizer x^* of one term in the sum is the maximum-likelihood estimate for the model parameter x, if the difference of model $y(t_k)$ and data z_k is considered as random variable.

Recall simple example: Data-fitting

- Minimizing one term in the sum:

$$\min_x (y(x;t_k) - z_k)^2 \Leftrightarrow \max_x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y(x;t_k)-z_k)^2}{2\sigma^2}\right)$$

  where σ^2 is the variance of $y(x;t_k) - z_k$, for example the measurement error.
- Minimizing the sum in the data-fitting cost function:

$$\min_x \sum_{k=1}^m (y(x;t_k) - z_k)^2 \Leftrightarrow \max_x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\sum_{k=1}^m \frac{(y(x;t_k)-z_k)^2}{2\sigma^2}\right)$$

- This is the density of the multivariate normal distribution with common variances σ^2 for all k.
- $\rightsquigarrow$ Interpretation: data considered as random variable Z_k with $\mathbb{E}(Z_k) = z_k$, then

$$\min_x (y(t_k) - z_k)^2$$

  means: Find parameter x such that model output $y(t_k)$ fits expectation of the data.

Generalization

- Standard least-squares cost function:

$$\min_x \sum_{k=1}^{m} (y_k - z_k)^2 = \min_x (y-z)^\top (y-z), \text{ with } y_k := y(x; t_k).$$

- Weighted least-squares function:

$$\min_x \sum_{k=1}^{m} \frac{1}{2\sigma_k^2}(y_k - z_k)^2 = \min_x \frac{1}{2}(y-z)^\top \Sigma^{-1}(y-z) \quad \text{with } \Sigma := \operatorname{diag}(\sigma_k^2) \in \mathbb{R}^{m\times m}.$$

- Including covariance matrix

$$\Sigma = Cov(Z), Z = (Z_k)_{k=1}^m \,,$$

$\rightsquigarrow$ generalized least-squares function:

$$\min_x \frac{1}{2}(y-z)^\top \Sigma^{-1}(y-z).$$

Contents

Normal-distributed samples

- Uniform distributed samples can be generated by standard (pseudo-) random number generators, see lecture 14.
- Box-Muller algorithm: Generation of normal-distributed random numbers:

- Let two uniform-distributed random vectors $X, Y \in \mathbb{R}^n$ be given. Then the function

$$G(X, Y) = \sqrt{-2 \log X}(\cos(2\pi Y), \sin(2\pi Y)).$$

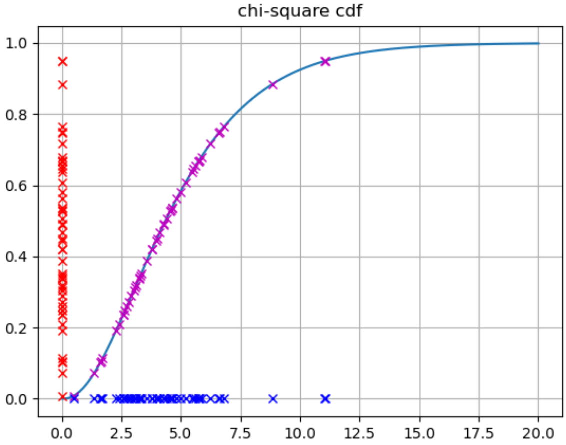
generates two standard-normal-distributed random vectors.

Random (Monte-Carlo) sampling

- Random sampling for arbitrary cdf F_X.
- Let $\{U_i, i = 1 \ldots, n\} \subset [0, 1]$ be a uniform-distributed sample.
- Determine $x = x(u)$ with

$$u = F_X(x) = P(X \leq x),$$
$$\text{i.e., } X := \inf\{y \in \mathbb{R} : U \leq F_X(y)\}.$$

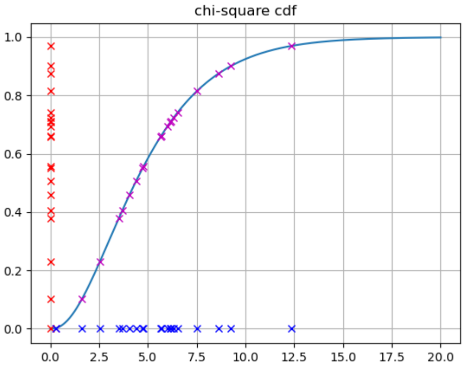
- X is now distributed with cdf F_X.
- Inverse cdfs can be found in tables or libraries.
- Uniform sample on whole interval $[0, 1]$ leads to clustering.

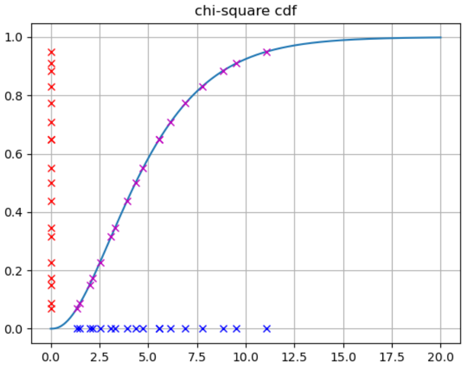


50 sample points

Stratified sampling

- Perform random sampling on a number of equidistant subintervals.
- Avoids clustering.



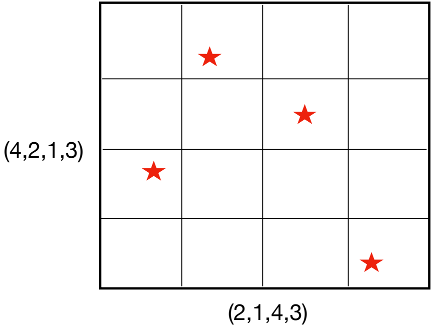


standard random (20 points) vs. stratified sampling (10×2 points)

Latin hypercube samples

- Generalization of stratified sampling in higher dimensions.
- Dimension n, total number of samples: m.
- The interval in each dimension is split into m equidistant subintervals.
- For every dimension $j = 1, \dots, n$, define a permutation of the subintervals:

$$\Pi_j(1, \dots, m) := (\pi_{1j}, \dots, \pi_{mj}) \,.$$



- Uniformly distributed **Latin-Hypercube sample** points $x_i = (x_{ij})_{j=1}^n \in [0,1]^n$ defined as

$$x_{ij} = \frac{\pi_{ij} - 1 + s_{ij}}{m}, \quad i = 1, \dots m, j = 1, \dots, n.$$

where s_{ij} are uniformly distributed random numbers in $[0,1]$.
- Latin Hypercube samples have better convergence properties than simple random samples.

What is important

- Confidence intervals can be used to test statistical hypotheses.
- This is is based on the assumption that the data are normal-distributed.
- A hypothesis test can be seen as different way to measure differences between data sets, taken into account the variance of the data.
- The inverse cdf is needed, whose values can be taken from tables or software libraries.
- The least-squares cost functions (we had in the regession problems) can be interpreted as a special kind of estimator, the maximum-likelihood estimator.
- Stratified and Latin Hypercube sampling are important ways to generate random samples.
- To generate samples for a given non-uniform probability distribution, we need again the inverse cdf.