

5 Sandwich estimator and bootstrap

❖ estimating the covariance matrix for parameter estimates from IID or IIDP data

- ▣ data are random samples drawn from some parent distribution
- ▣ parameters are obtained by processing the data into a set of numbers
 - ▶ the parameters are random variates
 - ▶ the goal is to obtain an estimate of their variance
- ▣ two methods
 - ▶ **sandwich estimator**: analytical error propagation
 - ▶ **bootstrap**: numerical resampling of data

➤ reminder

In both cases the uncertainty of a parameter estimated is quantified by **variance**, i.e. a single number characterizing the width of the distribution from which the parameter estimate is sampled. An alternative and conceptually different uncertainty are error intervals obtained from profiling the likelihood function.

The sandwich estimator

❖ Taylor expansion of a score function $g(\theta)$ for estimating θ around the true θ_0

$$g(\theta) = g(\theta_0) + \frac{\partial g(\theta_0)}{\partial \theta}(\theta - \theta_0) + O(1/N)$$

■ defining condition for $\hat{\theta}$

$$g(\hat{\theta}) = 0 = g(\theta_0) + \frac{\partial g(\theta_0)}{\partial \theta}(\hat{\theta} - \theta_0) \quad \text{and thus} \quad (\hat{\theta} - \theta_0) = - \left(\frac{\partial g(\theta_0)}{\partial \theta} \right)^{-1} g(\theta_0)$$

■ covariance matrix of parameters

$$C_{\theta} = \langle (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T \rangle = \left\langle \left(\frac{\partial g(\theta_0)}{\partial \theta} \right)^{-1} g(\theta_0) g(\theta_0)^T \left(\frac{\partial g(\theta_0)}{\partial \theta} \right)^{-T} \right\rangle$$

■ result, asymptotically (LLN) replacing $\partial g(\theta_0)/\partial \theta$ by its expectation value:

$$C_{\theta} = \left\langle \frac{\partial g(\theta_0)}{\partial \theta} \right\rangle^{-1} \langle g(\theta_0) g(\theta_0)^T \rangle \left\langle \frac{\partial g(\theta_0)}{\partial \theta} \right\rangle^{-T}$$

application →

Covariance matrix for parameters from Maximum Likelihood fits

❖ defining condition for ML-parameters from IID measurements

$$g_k = \frac{\partial}{\partial \theta_k} \sum_{i=1}^N \ln f(x_i, \theta) = 0$$

■ expectation value $\langle g g^T \rangle$

$$\begin{aligned} \langle g_k g_l \rangle &= \left\langle \sum_{i=1}^N \frac{\partial \ln f(x_i, \theta)}{\partial \theta_k} \sum_{j=1}^N \frac{\partial \ln f(x_j, \theta)}{\partial \theta_l} \right\rangle \\ &= N(N-1) \left\langle \frac{\partial \ln f(x, \theta)}{\partial \theta_k} \right\rangle \left\langle \frac{\partial \ln f(x, \theta)}{\partial \theta_l} \right\rangle + N \left\langle \frac{\partial \ln f(x, \theta)}{\partial \theta_k} \frac{\partial \ln f(x, \theta)}{\partial \theta_l} \right\rangle \\ &= N \left\langle \frac{\partial \ln f(x, \theta)}{\partial \theta_k} \frac{\partial \ln f(x, \theta)}{\partial \theta_l} \right\rangle = N I_{kl} \quad \text{or} \quad \langle g g^T \rangle = N I \end{aligned}$$

Fisher information matrix

■ expectation values of the Jacobian

$$\begin{aligned}\left\langle \frac{\partial g_k}{\partial \theta_l} \right\rangle &= \sum_{i=1}^N \left\langle \frac{\partial^2 \ln f(x_i, \theta)}{\partial \theta_k \partial \theta_l} \right\rangle = N \left\langle \frac{\partial^2 \ln f(x, \theta)}{\partial \theta_k \partial \theta_l} \right\rangle = N \int dx f(x, \theta) \frac{\partial^2 \ln f(x, \theta)}{\partial \theta_k \partial \theta_l} \\&= N \int dx f(x, \theta) \frac{\partial}{\partial \theta_k} \left(\frac{1}{f(x, \theta)} \frac{\partial f(x, \theta)}{\partial \theta_l} \right) \\&= N \int dx f(x, \theta) \left(-\frac{1}{f^2(x, \theta)} \frac{\partial f(x, \theta)}{\partial \theta_k} \frac{\partial f(x, \theta)}{\partial \theta_l} + \frac{1}{f(x, \theta)} \frac{\partial^2 f(x, \theta)}{\partial \theta_k \partial \theta_l} \right) \\&= -N \int dx f(x, \theta) \frac{\partial \ln f(x, \theta)}{\partial \theta_l} \frac{\partial \ln f(x, \theta)}{\partial \theta_k} + N \frac{\partial^2}{\partial \theta_k \partial \theta_l} \int dx f(x, \theta) \\&= -N \left\langle \frac{\partial \ln f(x, \theta)}{\partial \theta_k} \frac{\partial \ln f(x, \theta)}{\partial \theta_l} \right\rangle = -N I_{kl} \quad \text{or} \quad \left\langle \frac{\partial g}{\partial \theta} \right\rangle = -N I\end{aligned}$$

► result

$$C_{\theta} = (-N I)^{-1} (N I) (-N I)^{-T} = (N I)^{-1}$$

- the covariance matrix of the parameters is the inverse of the total Fisher information
- ML fits have asymptotically the best possible accuracy
- estimate of covariance matrix from a given data sample: Law of Large Numbers

$$C^{-1} = N I_{kl} = -N \left\langle \frac{\partial^2 \ln f(x_i, \theta)}{\partial \theta_k \partial \theta_l} \right\rangle \stackrel{N \rightarrow \infty}{=} - \sum_{i=1}^N \frac{\partial^2 \ln f(x_i, \theta)}{\partial \theta_k \partial \theta_l} = - \frac{\partial^2 \ln L(\theta)}{\partial \theta_k \partial \theta_l}$$

- second derivative of the negative-log-likelihood function at the minimum
- covariance matrix estimates for parameters from ML and LSQ fits by MINUIT

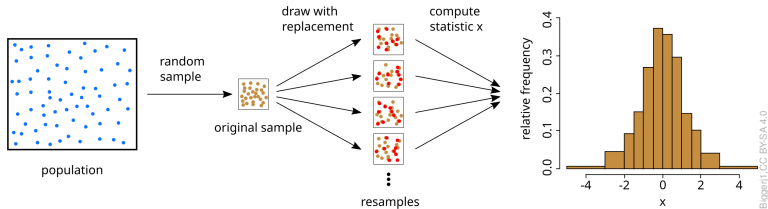
$$C_{\theta} = \left(\frac{\partial^2 \ln L}{\partial \theta^2} \right)^{-1} \quad \text{and} \quad C_{\theta} = \left(\frac{1}{2} \frac{\partial^2 \chi^2}{\partial \theta^2} \right)^{-1}$$

Concluding remarks on the sandwich estimator

- universal framework to estimate asymptotically correct covariance matrices
 - ▶ defined through derivatives at the true parameter values
 - ▶ in practice evaluated at the sample estimates
- works for any kind of parameter estimate based on $g(\theta) = 0$ with
 - ▶ one defining condition $g_k = 0$ for each parameter θ_k
 - ▶ covers also the case that some of the parameters are nuisance parameters
- inverse curvature of a cost function at the minimum works only in a special cases
 - ▶ OK for unweighted Maximum Likelihood fits
 - ▶ OK for Least Squares fits with proper variances for the measurements
 - ▶ note: in particular it fails for weighted Maximum Likelihood fits
 - see paper [arXiv:1911.01303](https://arxiv.org/abs/1911.01303) by Christoph Langenbruch

❖ “Münchhausen trick” for estimating uncertainties

- use the data as a proxy for the parent distribution from which the data are drawn



source: <https://commons.wikimedia.org/w/index.php?curid=135426288>

► estimating the uncertainty of a statistic $S(x_1, \dots, x_n)$

- randomly draw with replacement n values from $\{x_1, \dots, x_n\}$ and calculate S
- do this m -times to obtain values $S_1, \dots, S_m$
- the variance of the distribution of the S_k is an estimate for the variance of S

❖ illustration why this works

- consider the arithmetic average of N IID values x_i with variance $V(x_i) = \sigma^2$

$$S = \frac{1}{N} \sum_{i=1}^N x_i \quad \text{with} \quad V(S) = \frac{\sigma^2}{N}$$

- average of a bootstrap sample

$$S_b = \frac{1}{N} \sum_{i=1}^N w_i x_i \quad \text{with} \quad w_i \in \{0, 1, 2, \dots, N\}$$

- ▶ each value i is drawn with probability $p_i = 1/N$
- ▶ w_i is a vector of multinomial distributed random integers with $\sum w_i = N$,
- ▶ the covariance matrix elements $C_{ij}(w)$ of the weights are

$$C_{ij}(w) = N(p_i \delta_{ij} - p_i p_j) = \delta_{ij} - \frac{1}{N}$$

error propagation →

■ variance of S_b

$$\begin{aligned} V(S_b) &= \sum_{i,j=1}^N \frac{\partial S_b}{\partial w_i} \frac{\partial S_b}{\partial w_j} C_{ij}(w) \\ &= \sum_{i,j=1}^N \frac{x_i x_j}{N^2} \left(\delta_{ij} - \frac{1}{N} \right) = \frac{1}{N} \left[\frac{1}{N} \sum_{i=1}^N x_i^2 - \left(\frac{1}{N} \sum_{i=1}^N x_i \right)^2 \right] = \frac{\hat{\sigma}^2}{N} \end{aligned}$$

- ▶ $\hat{\sigma}^2$ is the empirical variance of the x_i
- ▶ the bootstrap estimates the variance of S as $\hat{\sigma}^2 / N$
- ▶ draw with replacement is crucial to get the correct result

■ generalization if data are given by a histogram with N entries

- ▶ normalize the histogram to get probabilities for all bins
- ▶ generate bootstrap samples from the multinomial distribution with N trials

❖ bootstrap in case the number of measurements is Poisson distributed

- common case in particle physics, rarely considered in classical statistics
- actually simpler case, since it avoids the multinomial constraint
- algorithm:
 - ▶ accept each x_i multiple times with $n = \{0, 1, 2, \dots\}$
 - ▶ determine n as a Poisson distributed random number with average $\mu = 1$
- generalization if data are given by a histogram with N entries
 - ▶ take histogram contents as Poisson averages
 - ▶ generate bootstrap samples as Poisson-distributed random numbers in each bin

problems and exercises

Consider an experiment to determine the average lifetime of an unstable particle, which recorded N IIDP events drawn from an exponential PDF with mean value $\langle t \rangle = 1$. Calculate the mean value and do an extended bootstrap to estimate variance and standard deviation of the mean value. Take N to be a Poisson distributed number with mean value $\langle N \rangle = 400$. Compare the bootstrap estimate to the expected variance of $1/\langle N \rangle$.

