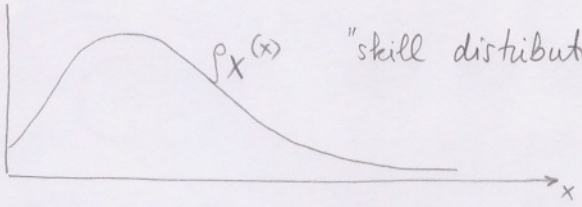
Frequentist Consistency of Frequentist & Bayesian Methods

(I) Introductory Example
(II) Setup & Notation
(III) What are we going to talk about?
(IV) Frequentist Consistency of Frequentist Methods (MLE)
(V) Asymptotic Normality of Frequentist Methods (MLE)
(VI) Frequentist Consistency of Bayesian Methods

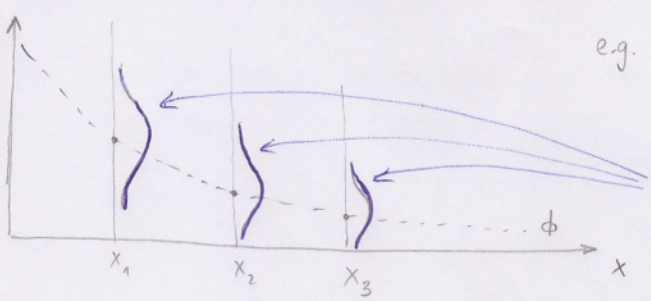
Sources: Tim Sullivan - Introduction to Uncertainty Quantification [S]
Richard Nickl - Statistical Theory [N]

(I) Introductory Example

- X = skiing skill of a person – "value of interest"
- Z = time measurement on some racing track

(a) 

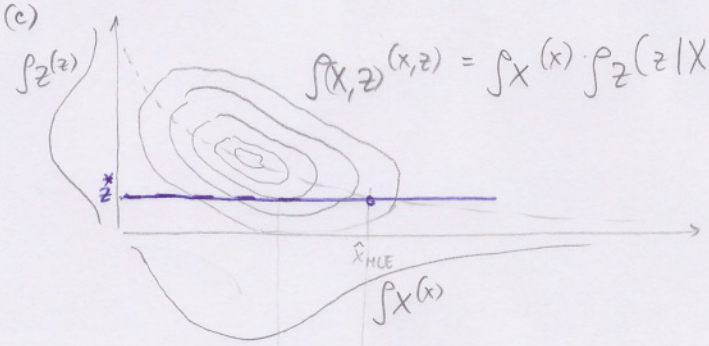
"skill distribution within the whole population"

(b) 

e.g. $Z = \phi(X) + E$ (model; uncertainty / measurement error)

"likelihood fct." $L(x,z) := f_Z(z \mid X = x)$

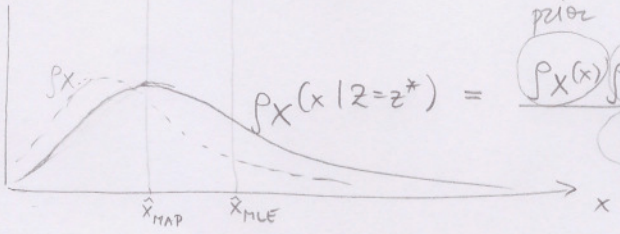
(a) and (b) give the joint probability density of X and Z:

(c) 

$$f_{(X,Z)}(x,z) = f_X(x) \cdot f_Z(z \mid X = x)$$

(f_X, f_Z as marginal densities)

Restricting (c) to $Z = z^*$ yields the posterior density $f_X(x \mid Z = z^*)$

(d) 

$$f_X(x \mid Z = z^*) = \frac{\overbrace{f_X(x)}^{\text{prior}} \overbrace{f_Z(z^* \mid X = x)}^{\text{likelihood}}}{\underbrace{f_Z(z^*)}_{\text{evidence}}} = \frac{f_X(x) \cdot L(x, z^*)}{\text{normalization}}$$

"Bayes' rule"

If we have many (independent) measurements for the same person, $Z_1 = z_1^*, \dots, Z_M = z_M^*$,

Bayes' rule becomes:

$$f_X(x \mid Z_1 = z_1, \dots, Z_M = z_M) = \frac{f_X(x) \left(\prod_{m=1}^{M} f_Z(z_m^* \mid X = x) \right)}{\text{norm.}} = f_{(Z_1,\dots,Z_M)}(z_1^*,\dots,z_M^* \mid X = x)$$

joint likelihood

(II) Setup & Notation (S)

- $X \in \mathbb{R}^d$ unknown parameter / value of interest
- $Z \in \mathbb{R}^n$ measurement with prob. density ρ_Z
- Input: - prob. density $\rho_X \in C^1(\mathbb{R}^d)$ of X
  - likelihood fct. $L(x,z) = \rho_Z(z|X=x)$
  - one or more realization(s) $z_1 = z_1^*, \dots, z_M = z_M^*$
- Output: - Estimation $\hat{x}_M$ of X
  - its uncertainty: – [the distribution of] the error $\hat{x}_M - X$
- We will always assume that there exists an (unknown) "true" parameter value $x_T \in \operatorname{supp}(\rho_X)$ which "induces" the measurements, i.e.

  $z_1, \dots, z_M \sim Z_T$, Z_T given by the density $\rho_Z(z|X=x_T)$

  Def: A parametric model $\big(\underbrace{\{\rho_Z(\cdot|X=x) \mid x \in \mathbb{R}^d\}}_{\text{likelihood model}}, \underbrace{\mathbb{P}}_{\substack{\text{data-generating} \\ \text{distribution}}}\big)$ is said to be "correctly or well-specified" if there exist an $x_T \in \mathbb{R}^d$ s.t. $\mathbb{P}$ has the prob. density $\rho_Z(z|X=x_T)$.

- Def: The maximum likelihood estimator (MLE) $\hat{x}_M$ based on $z_1, \dots, z_M$ is given by

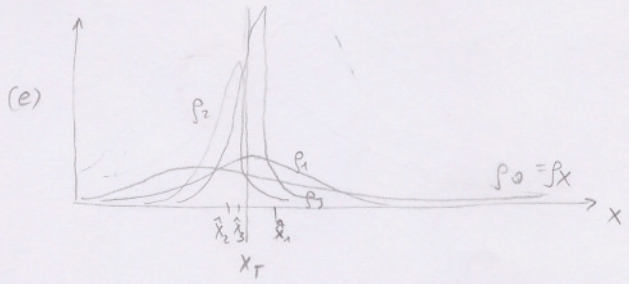
$$\hat{x}_M = \operatorname*{argmax}_x \prod_{m=1}^{M} L(x, z_m) = \operatorname*{argmax}_x \sum_{m=1}^{M} \log L(x, z_m) = \operatorname*{argmin}_x \sum_{m=1}^{M} q(x, z_m)$$

where $\boxed{q(x,z) := -\log(x,z)}$ is the negative log-likelihood

(III) What are we going to talk about?

Example 2(a): Same as introductory example, but with many measurements

Let $\rho_M(x) := \rho_X(x \mid Z_1 = z_1^*, \dots, Z_M = z_M^*)$ and $\hat{x}_M$ be the MLE based on $Z_1, \dots, Z_M$



Natural Questions: What happens to $\hat{x}_M$ and ρ_M for $M \to \infty$?

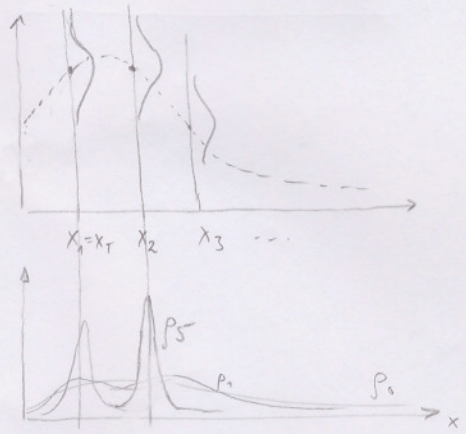
(i) $\hat{x}_M \longrightarrow x_T$? In what sense?

(ii) ρ_M concentrates around x_T? How exactly?

(iii) What are sufficient conditions for (i) and (ii)?

Let's start with (iii)

Example 2(b): Same as 2(a), but with different "model" ϕ (more general, different likelihood model $\{L(x,\cdot) \mid x \in \mathbb{R}^d\}$)



(i) and (ii) go wrong!

So, a necessary condition is the so-called identifiability assumption:

Def: A likelihood model $\{L(x,\cdot) \mid x \in \mathbb{R}^d\}$ fulfills the identifiability assumption (IA), if

$$\forall x_1, x_2;\ x_1 \neq x_2 : \quad \rho_Z(\cdot \mid X = x_1) \neq \rho_Z(\cdot \mid X = x_2) \quad (\text{in } L^1) \qquad \text{(IA)}$$

Further criteria we are going to use:

(C1) $L(x,z) > 0\ \forall x,z$, L is continuous in x (and $\int L(x,z)\,dz = 1\ \forall x$)

(C2) $\mathbb{E}\left[\sup_x |L(x, Z_T)|\right] < \infty$

(C3) $L(\cdot, z) \in C^2(U)\ \forall z$ on some nbh. $x_T \in U \subset \mathbb{R}^d$, s.t.

- $\mathbb{E}\left[D_x^2 q(x_T, Z_T)\right] \in GL(d, \mathbb{R})$ and $\mathbb{E}\left[\|\partial_x q(x_T, Z_T)\|^2\right] < \infty$
- $\exists K = B_{x_T}(r)$ compact ball around x_T with radius $r > 0$ s.t.

$$\mathbb{E}\left[\sup_{x \in K} \|D_x^2 q(x, Z_T)\|\right] < \infty, \quad \int \sup_{x \in K} \|\partial_x L(x,z)\|\,dz < \infty, \quad \int \sup_{x \in K} \|D^2 L(x,z)\|\,dz < \infty$$

(IV) Frequentist Consistency of the MLE:

Thm: Assuming Setup (S), (IA) and (C1), (C2), the MLE $\hat{x}_M$ based on $z_1, \dots, z_M$ is consistent,
[N, Thm 2] [S, Thm 6.13] i.e. $\hat{x}_M \xrightarrow{\mathbb{P}} x_T$

Proof (sketch): (1) $\sup_{x \in \mathbb{R}^d} \left| \underbrace{\mathbb{E}\left[q(x, Z_T)\right]}_{=: Q(x) \text{ (fct.)}} - \underbrace{\frac{1}{M} \sum_{m=1}^{M} q(x, Z_m)}_{=: Q_M(x) \text{ (r.v.)}} \right| \xrightarrow[M \to \infty]{\mathbb{P}\text{-a.s.}} 0$ "uniform law of large numbers"

⌜proof see [N., Prop 2]⌟

(2) x_T minimizes Q (uniquely!): $x_T = \operatorname{argmin} Q(x)$

⌜$\forall x \neq x_T$: $Q(x_T) - Q(x) = \int q(x_T, z) L(x_T, z)\, dz - \int q(x, z) L(x_T, z)\, dz = \underbrace{\int \log\left[\frac{L(x,z)}{L(x_T,z)}\right] L(x_T, z)\, dz}_{-D_{KL}\left(L(x_T,\cdot) \,\|\, L(x,\cdot)\right) > 0 \text{ (IA)}}$

$\overset{\text{strict Jensen ineq.}}{<}$ (IA) $\log \int \frac{L(x,z)}{L(x_T,z)} L(x_T, z)\, dz = \log 1 = 0$ ⌟

(3) Part (a) implies $\operatorname{argmin}_x Q_M(x) \xrightarrow{\mathbb{P}} \operatorname{argmin} Q(x)$
$\|$ Def. $\hat{x}_M$ $\quad$ $\|$ (2) x_T

⌜Technical proof see [N, Thm 1]⌟

Picture

(V) Asymptotic Normality of the MLE

Def: The Fisher information $I: \mathbb{R}^d \to \mathbb{R}^{d\times d}$ of a C^1-likelihood model $\{\rho_Z(\cdot|X=x) \mid x\in\mathbb{R}^d\}$ is defined as $I(x) = \mathbb{C}ov\left[(\partial_x q)(x, Z_x)\right]$ where $q(x,z) = -\log \rho_Z(z|X=x)$ and Z_x has the prob. density $\rho_Z(\cdot|X=x)$

In particular, $I(x_T) = \mathbb{C}ov\left[(\partial_x q)(x_T, Z_T)\right] \overset{\text{technical detail}}{=} \mathbb{E}\left[(D_x^2 q)(x_T, Z_T)\right]$. ⊛
(under regularity assumptions)

Thm: Assuming Setup (S), (IA) and (C1)–(C3), the MLE $\hat{x}_M$ based on $Z_1, \dots, Z_M$ fulfills
[N, Thm 3] [S, Thm 6.16]
$$\sqrt{M}\,(\hat{x}_M - x_T) \xrightarrow[M\to\infty]{\text{in distr.}} \mathcal{N}(0, I(x_T)^{-1}).$$

Proof (sketch): $0 \overset{\text{Def. } \hat{x}_M}{=} \partial_x Q_M(\hat{x}_M) \overset{\text{Taylor}}{=} \partial_x Q_M(x_T) + D_x^2 Q_M(\tilde{x}_M^{\in[x_T,\hat{x}_M]})\cdot(\hat{x}_M - x_T)$ ⊛⊛

So, $\underbrace{\sqrt{M}(\hat{x}_M - x_T)}_{\text{guy of interest}} = -\underbrace{\left(D_x^2 Q_M(\tilde{x}_M)\right)^{-1}}_{\text{(A)}}\underbrace{\left(\sqrt{M}\,\partial_x Q_M(x_T)\right)}_{\text{(B)}}$

(A): It can be shown that (since $\hat{x}_M \xrightarrow{\mathbb{P}} x_T$ and thereby $\tilde{x}_M \xrightarrow{\mathbb{P}} x_T$, law of large numbers)
$$\underbrace{D_x^2 Q_M(\tilde{x}_M)}_{\frac{1}{M}\sum_{m=1}^M q(\tilde{x}_M, Z_m)} \xrightarrow[M\to\infty]{\mathbb{P}} \mathbb{E}\left[D_x^2 q(x_T, Z_T)\right] \underset{⊛}{=} I(x_T)$$

(B) $\cdot\ \mathbb{E}\left[\partial_x q(x_T, Z_T)\right] = \partial_x \underbrace{\mathbb{E}[q(x_T, Z_T)]}_{Q(x_T)} = 0$ since x_T minimizes Q.

$\cdot\ \sqrt{M}\cdot\underbrace{\partial_x Q_M(x_T)}_{=\frac{1}{M}\sum_{m=1}^M \partial_x q(x_T, Z_m)} \xrightarrow[M\to\infty]{\text{in distr.}}_{\text{(CLT)}} \mathcal{N}(0, \mathbb{C}ov\left[\partial_x q(x_T, Z_T)\right]) = \mathcal{N}(0, I(x_T))$

So, $\sqrt{M}\,(\hat{x}_M - x_T) = -\underbrace{\left(D_x^2 Q_M(\tilde{x}_M)\right)^{-1}}_{\xrightarrow{\mathbb{P}} I(x_T)\ \text{(A)}} \cdot \underbrace{\sqrt{M}\,\partial_x Q_M(x_T)}_{\xrightarrow{\text{in distr.}} \mathcal{N}(0, I(x_T))\ \text{(B)}} \xrightarrow[M\to\infty]{\text{in distr.}}_{\text{(Gauß)}} \mathcal{N}\left(0, I(x_T)^{-1} I(x_T)\, I(x_T)^{-1}\right) = \mathcal{N}(0, I(x_T)^{-1})$ □

Reminder:
(CLT) $Y_1, \dots, Y_M \overset{iid}{\sim} Y \Rightarrow \sqrt{M}\left(\frac{1}{M}\sum_{m=1}^M Y_m - \mathbb{E}[Y]\right) \xrightarrow[M\to\infty]{\text{in distr.}} \mathcal{N}(0, \mathbb{C}ov(Y))$

(Gauß) $Y \sim \mathcal{N}(\mu, \Sigma)$, $A \in GL(d,\mathbb{R}) \Rightarrow A\cdot Y \sim \mathcal{N}(A\cdot\mu,\ A\Sigma A^T)$

Remark: In dimension $d>1$, the Taylor-expansion ⊛⊛ has to be performed in each dimension separately: $\forall j = 1,\dots,d$: $0 = \left[\partial_x Q_M(\hat{x}_M)\right]_j + \left[D_x^2 Q_M(\underbrace{\tilde{x}_M^j}_{\in[x_T,\hat{x}_M]})\right]_{j,\cdot}\cdot(\hat{x}_M - x_T)$

Then replace $D_x^2 Q_M(\tilde{x}_M)$ by H^M defined by $H^M_{j,\cdot} = \left[D_x^2 Q_M(\tilde{x}_M^j)\right]_{j,\cdot}$ $\forall j$:
$$0 = \partial_x Q_M(\hat{x}_M) - \partial_x Q_M(x_T) + H^M\cdot(\hat{x}_M - x_T)$$

Since $D_x^2 Q_M(\tilde{x}_M^j) \xrightarrow[M\to\infty]{\mathbb{P}} I(x_T)$, also $H^M \xrightarrow{\mathbb{P}} I(x_T)$, so this detail does not affect the rest of the proof.

(VI) Frequentist Consistency of Bayesian Methods

Thm ([N, Thm. 5], [S, Thm. 6.17], Bernstein - von Mises Theorem):

Assuming Setup (S), (IA) and (C1)-(C3) and denoting

- by $\hat{x}_M$ the MLE based on $Z_1, \dots, Z_M$,
- by $\mathbb{P}_M$ the posterior distribution with density $\rho_M(x) = \rho_X(x \mid Z_1 = z_1^*, \dots, Z_M = z_M^*)$,
- by $\|P - Q\|_{TV} := \sup_{B \in \mathcal{B}or(\mathbb{R}^d)} |P(B) - Q(B)|$ the total variation distance between P, Q,

we have:

$$\left\| \mathbb{P}_M - \mathcal{N}\left(\hat{x}_M, \tfrac{1}{M} I(x_T)^{-1}\right) \right\|_{TV} \xrightarrow[M \to \infty]{\mathbb{P}} 0$$