RISK FUNCTION. The expectation of the quadratic estimation error
\[ \text{Err}_T = \mathbb{E}[(T - g(\theta))^2] \]
is called the risk function or the mean square error of the estimator \( T \). It measures the estimator performance.

REMARK. \( \text{Err}_T = \text{Var}_T[T] + \text{Bias}_T^2 \).
EXAMPLE. If \( T \) is unbiased, then \( \text{Err}_T = \text{Var}_T[T] \).

EXAMPLE. The arithmetic mean is the "best linear unbiased estimator".
Proof. With \( T = \sum \alpha_i X_i \), where \( \sum \alpha_i = 1 \) the risk function is \( \text{Err}_T[T] = \sum \alpha_i^2 \text{Var}_T[X_i] \) which is by Lagrange minimal for \( \alpha_i = 1/n \).

MAXIMUM LIKELIHOOD FUNCTION. The maximum likelihood function \( \ell(x_1, \ldots, x_n) \) is defined as the maximum of
\[ \ell(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n) \]
The maximum likelihood estimator is the random variable \( T(\omega) = (X_1(\omega), \ldots, X_n(\omega)) \). For discrete random variables, \( f(x_1, \ldots, x_n) \) would be replaced by \( f_j = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \). One also looks at the maximum a posteriori estimator, which is the maximum of
\[ \ell(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n) |\theta \]
where \( g(\theta) \) was the a priori distribution on \( \Theta \).

MINIMAX PRINCIPLE. Find min \( m \max R(\theta, T) \) to find the worst case.

BAYES PRINCIPLE. Find \( \min_T \int_{\Omega} R(\theta, T) d\mathbb{P}(\theta) \).

EXAMPLES.

1. \( f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \). The maximum likelihood function \( L_1(x_1, \ldots, x_n) = \frac{1}{n} e^{-\sum x_i^2/2} \) is maximal when \( \sum x_i^2 \) is minimal which means that \( \ell(x_1, \ldots, x_n) \) is the median of the data \( x_1, \ldots, x_n \).

2. \( f_2(x) = \theta^x e^{-\theta} x! \) Poisson distribution. The maximum likelihood function \( I_2(x_1, \ldots, x_n) = \sum \log(x_i - \theta) / x_i! \) is maximal for \( \theta = x_i/n \).

3. The maximum likelihood estimator for \( \theta = (m, \sigma^2) \) for Gaussian distributed random variables \( f_3(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} \) has the maximum likelihood function maximized for \( \ell(x_1, \ldots, x_n) = \frac{1}{n} \sum x_i^2 / n - \frac{1}{2n} \sum (x_i - \bar{x})^2 \).

FISHER INFORMATION. Define the Fisher information of a random variable \( X \) with density \( f_0 \)
\[ I(\theta) = \int \frac{\partial^2}{\partial \theta^2} \log(f_0(x)) f_0(x) dx. \]
Remark. In the multiparametric case, one defines the Fisher information matrix \( I_{ij}(\theta) = \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(f_0(x)) f_0(x) dx \).

LEMMA 1. \( I(\theta) = \mathbb{E}[(\partial \theta)^2(X)] = \mathbb{V}(\frac{\partial}{\partial \theta} f_0). \)
Proof. This follows from \( \mathbb{E}[\frac{\partial}{\partial \theta} f_0] = \int \frac{\partial}{\partial \theta} f_0(x) dx = 0 \).

LEMMA 2. \( I(\theta) = -\mathbb{E}[\log(f_0)^\prime]. \)
Proof. \( \mathbb{E}[(\partial \theta)^\prime] = \int \frac{\partial}{\partial \theta} \log(f_0(x)) f_0(x) dx = -\int \log(f_0(x)) f_0(x) dx = -\int (\log(f_0(x)) f_0(x) dx = \int f_0(x) \log(f_0(x)) dx. \)

THE SCORE FUNCTION is defined as the logarithmic derivative \( p_0 = f_0'/f_0 \). One has \( I(\theta) = \mathbb{E}[p_0^2] = \mathbb{V}[p_0]. \)

EXAMPLE. If \( X \) is Gaussian, the score function \( p_0 = f'(\theta)/f(\theta) = -x/(\sigma^2) \) is linear and has variance \( 1 \). The Fisher information \( I \) is \( 1/\sigma^2 \). We see that \( \mathbb{V}[X] = 1/I \). This is a special case \( n = 1, T = X; \theta = m \) is the mean of the following bound.
### Rao-Cramer Bound

<table>
<thead>
<tr>
<th>( \text{Var}[T] \geq \frac{1}{n\theta^2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased case:  ( \text{Ent}[T] \geq \frac{1}{2\theta^2} )</td>
</tr>
</tbody>
</table>

**Proof:**

1. \( \theta + B(\theta) = E_\theta[T] = \int \ell(x_1, \ldots, x_n) L_\theta(x_1, \ldots, x_n) \, dx_1 \ldots dx_n \).
2. \( 1 + B'(\theta) = \int \ell(x_1, \ldots, x_n) L_\theta(x_1, \ldots, x_n) \, dx_1 \ldots dx_n = \int \ell(x_1, \ldots, x_n) \frac{L_\theta(x_1, \ldots, x_n)}{L_\theta(x_1, \ldots, x_n)} \, dx_1 \ldots dx_n = E_\theta[T] \).
3. \( 1 = \int L_\theta(x_1, \ldots, x_n) \, dx_1 \ldots dx_n \) implies \( 0 = \int \frac{L_\theta(x_1, \ldots, x_n)}{L_\theta(x_1, \ldots, x_n)} \, dx_1 \ldots dx_n = E[L_\theta/L_\theta] \).
4. Using 3) and 2) \( \text{Cov}[T, L_\theta/L_\theta] = E[T L_\theta/L_\theta] - 0 = 1 + B'(\theta) \).
5. \( 1 + B'(\theta)^2 = \text{Cov}^2[T, L_\theta/L_\theta] \leq \text{Var}_\theta[T] \text{Var}_\theta[T/L_\theta] = \text{Var}_\theta[T] \sum_{i=1}^n E_\theta[\rho_i(x_i)/f_i(x_i)] = \text{Var}_\theta[T] n I(\theta) \).

### Shannon Entropy

Closely related to the Fisher information is the **Shannon entropy** of a random variable \( X \):

\[ S(\theta) = -\int f(x) \log(f(x)) \, dx \]

and the **power entropy**

\[ N(\theta) = \frac{1}{2\pi e^2} e^{2S(\theta)}. \]

### Information Inequalities

- **Fisher information inequality:**
  \[ I_{X,Y} \geq 2^{-1} I_X + I_Y \]
- **Power entropy inequality:**
  \[ N_{X+Y} \geq N_X + N_Y \]
- **Uncertainty property:**
  \[ I_X N_X \geq 1 \]

In all cases, equality holds if and only if the random variables are Gaussian.

**Proof.**

- a) \( I_{X+Y} = \text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] \geq 2^{-1} (\text{Var}[X] + \text{Var}[Y]) \).
- b) \( N_{X+Y} \geq N_X + N_Y \) is proven using the **Jensen inequality**. Take then \( \epsilon = I_Y/(I_X + I_Y) \).

### Rao-Cramer Bound

A random variable \( X \) with mean \( m \) and variance \( \sigma^2 \) satisfies: \( I_X \geq 1/\sigma^2 \). Equality holds if and only if \( X \) is the Normal distribution.

**Proof.**

This is a special case of Rao-Cramer inequality, where \( \theta \) is fixed, \( n = 1 \). The bias is automatically zero. A direct computation giving also uniqueness: \( E[(aX + b)\rho(X)] = \int (aX + b)f(x) \, dx = -a \int f(x) \, dx = -a \) implies

\[ 0 \leq E[(\rho(X) + (X - m)/\sigma)^2] = E[(\rho(X)^2] + 2E[(X - m)\rho(X)]/\sigma^2 + E[(X - m)^2/\sigma^4] \leq I_X - 2/\sigma^2 + 1\sigma_2 \]

Equality holds if and only if \( \rho_X \) is linear, i.e. when \( X \) is Normal.

### Corollary: Fisher information extremization

The normal distribution has the smallest Fisher information among all distributions with the same variance \( \sigma^2 \).

### Entropy characterizations of distributions

Maximizers of the Shannon entropy are

1. The uniform distribution on \([a, b]\).
2. The exponential distribution on \([0, \infty]\).
3. The Gaussian distribution on the real line.

These are results from the calculus of variations with constraints. For 1), one has to extremize \( F(f) = \int_a^b \log(f) \, f \, dx \) under the constraint \( G(f) = \int_a^b f(x) \, dx = 1 \). The Lagrange equations are \( 1 - \log(f) = \lambda \), so that \( f = 1/(b-a) \) is constant.

### Literature

This document partly based on lecture notes of a solid probability and statistics course at ETH given by H. Föllmer, which all mathematics students had to take.