Formula sheet for Statistical Theory and Modeling

Course code: ST2601

Stockholms universitet

Arithmetics

Powers

For all real numbers x, y and positive numbers a, b

- $a^x a^y = a^{x+y}$
- $\frac{a^x}{a^y} = a^{x-y}$
- $(ab)^x = a^x b^x$
- $(\frac{a}{b})^x = \frac{a^x}{b^x}$
- $\frac{1}{a^x} = a^{-x}$
- $(a^x)^y = a^{xy}$
- $a^0 = 1$
- $a^{\frac{1}{n}} = \sqrt[n]{a}$, where n is a positive integer

Natural logarithm

For positive numbers x, y

- $e^x = y \iff x = \ln(y)$
- $\ln(xy) = \ln(x) + \ln(y)$
- $\ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y)$
- $\ln(x^p) = p\ln(x)$

Some special functions

Factorial of positive integers n

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$$

and $0! = 1$.

Gamma function

Properties of the Gamma function

$\Gamma(n) = (n-1)!$ if n is a positive integer

$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for any $\alpha > 0$

Derivatives

Derivatives of elementary functions

k and a are constants.

- $\frac{d}{dx}k = 0$
- $\frac{d}{dx}x^n = nx^{n-1}$
- $\frac{d}{dx}e^{ax} = ae^{ax}$
- $\frac{d}{dx}\ln(x) = \frac{1}{x}$, $x > 0$
- $\frac{d}{dx}a^x = \frac{a^x}{\ln a}$

Derivatives of combined functions

$f(x)$ and $g(x)$ are differentiable functions, and k a constant.

- Constant rule

$$\frac{d}{dx}(k \cdot f(x)) = k \cdot f'(x)$$

- Sum rule

$$\frac{d}{dx}(f(x) + g(x)) = f'(x) + g'(x)$$

- Product rule

$$\frac{d}{dx}(f(x) \cdot g(x)) = f'(x)g(x) + f(x)g'(x)$$

- Quotient rule

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f'(x)g(x) - f(x)g'(x)}{\left(g(x)\right)^2}$$

- Reciprocal rule

$$\frac{d}{dx}\left(\frac{1}{g(x)}\right) = -\frac{g'(x)}{\left(g(x)\right)^2}$$

- Chain rule for a composite function

$$\frac{d}{dx}f(g(x)) = f'(g(x)) \cdot g'(x)$$

Integrals

Anti-derivatives

C and k are constants.

- $\int x^n \, dx = \frac{1}{n+1}x^{n+1} + C$, $n \neq -1$
- $\int e^{ax} \, dx = \frac{1}{a}e^{ax} + C$, $a \neq 0$
- $\int \frac{1}{x} \, dx = \ln|x|$, $x > 0$

Definite integral of $f(x)$ from a to b

$$\int_a^b f(x) \, dx = [F(x)]_a^b = F(b) - F(a)$$

Integrals of combined functions

$f(x)$ and $g(x)$ are integrable functions.

- Constant rule

$$\int k \cdot f(x) \, dx = k \cdot \int f(x) \, dx$$

- Sum rule

$$\int (f(x) + g(x)) \, dx = \int f(x) \, dx + \int g(x) \, dx$$

Combinatorics

Combinations and Permutations

How many ways can we choose k elements from n elements?		
	with replacement	without replacement
order	n^k	${}_nP_k = \frac{n!}{(n-k)!}$
no order	not included	${}_nC_k = \binom{n}{k} = \frac{n!}{(n-k)!k!}$

Descriptive Statistics

Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample Variance

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Sample standard deviation

$$s_x = \sqrt{s_x^2}$$

Sample covariance

$$s_{xy} = \mathrm{Cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Sample correlation

$$r_{xy} = \mathrm{Corr}(x, y) = \frac{s_{xy}}{s_x s_y}$$

Basic Probability

Addition Rule

$$P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \cap \mathbf{B})$$

Multiplication Rule

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{B}|\mathbf{A})P(\mathbf{A}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B})$$

Law of Total Probability - basic version

$$P(\mathbf{A}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) + P(\mathbf{A}|\mathbf{B}^c)P(\mathbf{B}^c)$$

where $\mathbf{B}^c$ is the complement of $\mathbf{B}$.

Law of Total Probability - general partition

$$P(\mathbf{A}) = \sum_{k=1}^{K} P(\mathbf{A}|\mathbf{B}_k)P(\mathbf{B}_k)$$

where $\mathbf{B}_1, \ldots, \mathbf{B}_K$ is a partitioning of the sample space.

Bayes' Theorem - basic version

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{B})P(\mathbf{B})}{P(\mathbf{A})}$$

Bayes' Theorem - general partition

$$P(\mathbf{B}_k|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{B}_k)P(\mathbf{B}_k)}{P(\mathbf{A})}$$

Properties of One Random Variable

Expected value

If X is a discrete variable with probability function $p(x)$

$$\mu = \mathbb{E}(X) = \sum_{\text{all } x} x \cdot p(x)$$

If X is a continuous variable with density function $f(x)$

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x$$

Expected value of a function $g(X)$

If X is a discrete variable with probability function $p(x)$

$$\mathbb{E}(g(X)) = \sum_{\text{all } x} g(x) \cdot p(x)$$

If X is a continuous variable with density function $f(x)$

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f(x)\mathrm{d}x$$

Variance

If X is a discrete variable with probability function $p(x)$

$$\sigma^2 = \mathbb{V}(X) = \sum_{\text{all } x}(x-\mu)^2 \cdot p(x) = \mathbb{E}(X^2) - \mu^2$$

If X is a continuous variable with density function $f(x)$

$$\sigma^2 = \mathbb{V}(X) = \int_{-\infty}^{\infty}(x-\mu)^2 \cdot f(x)\mathrm{d}x = \mathbb{E}(X^2) - \mu^2$$

Standard deviation

$$\sigma = \mathbb{S}(X) = \sqrt{\mathbb{V}(X)}$$

Expected value linear combination (c and d are constants)

$$\mathbb{E}(c + d \cdot X) = c + d \cdot \mathbb{E}(X)$$

Variance linear combination

$$\mathbb{V}(c + d \cdot X) = d^2 \cdot \mathbb{V}(X)$$

Distribution of a transformation

Let X be a continuous random variable and $Y = g(X)$, where $g(x)$ is a monotone differentiable function with inverse function $x = g^{-1}(y)$. Then,

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \cdot \left|\frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y}\right|$$

Properties of Two Random Variables

Expected value of a linear combination

$$\mathbb{E}(cX + dY) = c\mathbb{E}(X) + d\mathbb{E}(Y)$$

Variance for a linear combination

$$\mathbb{V}(cX + dY) = c^2\mathbb{V}(X) + d^2\mathbb{V}(Y) + 2cd\mathrm{Cov}(X, Y)$$

Marginal distribution for X

If X and Y are discrete variables with joint probability function $p(x, y)$, then the marginal distribution of X is

$$p_X(x) = \sum_{\text{all } y} p(x, y)$$

If X and Y are continuous variables with joint density function $f(x, y)$, then the marginal density of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}y$$

Conditional distribution for Y given X

If X and Y are discrete variables with joint probability function $p(x, y)$, then the conditional distribution of Y is

$$p(y \mid x) = \frac{p(x, y)}{p_X(x)}, \quad p_X(x) > 0$$

where $p_X(x)$ is the marginal distribution for X.

If X and Y are continuous variables with joint density function $f(x, y)$, then the conditional density of Y is

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)}, \quad f_X(x) > 0$$

where $f_X(x)$ is the marginal density for X.

Law of iterated expectation

$$\mathbb{E}_Y(Y) = \mathbb{E}_X\left(\mathbb{E}_{Y|X}(Y|X)\right)$$

Covariance between two random variables X and Y

$$\mathrm{Cov}(X, Y) = \mathbb{E}\Big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\Big) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Covariance between two discrete random variables X and Y

$$\mathrm{Cov}(X, Y) = \sum_{\text{all pairs } (x,y)} p(x, y)(x - \mathbb{E}(X))(y - \mathbb{E}(Y))$$

where $p(x, y)$ is the joint distribution of X and Y.

Correlation between two random variables X and Y

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\mathbb{S}(X) \cdot \mathbb{S}(Y)}$$

Properties of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be **independent identically distributed** random variables with expected value $\mu = E(X_i)$ and variance $\sigma^2 = \mathrm{Var}(X_i)$. For the sample mean $\overline{X}_n = \sum_{i=1}^{n} X/n$ we have:

Expected value of the sample mean

$$\mathrm{E}(\overline{X}_n) = \mu$$

Variance of the sample mean

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

Law of Large Numbers

$$\overline{X}_n \xrightarrow{p} \mu$$

where $\xrightarrow{p}$ is convergence in probability: for all $\epsilon > 0$

$$P(|\overline{X}_n - \mu| > \epsilon) \to 0 \text{ when } n \to \infty$$

Central Limit Theorem (informally)

$$\overline{X}_n \overset{\text{approx}}{\sim} \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ for large } n$$

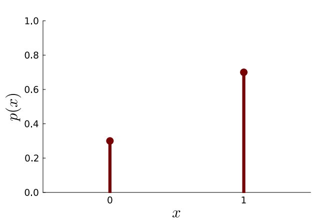
Rule of Thumb: the approximation is accurate if $n \geq 30$.

Discrete Distributions

Bernoulli distribution $X \sim \mathrm{Bernoulli}(p)$

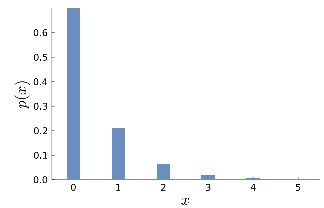
$$p(x) = \begin{cases} q & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases} \quad \text{where } q = 1 - p.$$

$$\mathbb{E}(X) = p$$
$$\mathbb{V}(X) = pq$$



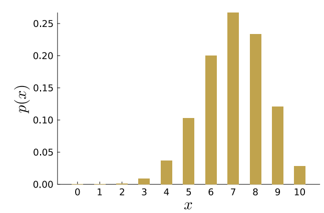
Geometric distribution $X \sim \mathrm{Geom}(p)$

$$p(x) = q^x p \text{ for } x = 0, 1, 2, \ldots$$
$$\mathbb{E}(X) = \frac{1-p}{p}$$
$$\mathbb{V}(X) = \frac{1-p}{p^2}$$



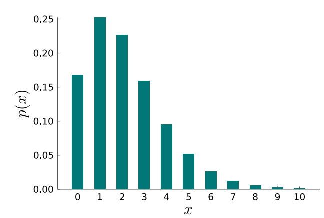
Binomial Distribution: $X \sim \mathrm{Binomial}(n, p)$

$$p(x) = \binom{n}{x} p^x q^{n-x} \text{ for } x = 0, 1, 2, \ldots, n$$
$$\mathbb{E}(X) = np$$
$$\mathbb{V}(X) = npq$$



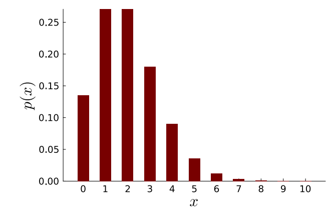
Negative Binomial distribution: $X \sim \mathrm{NegBin}(r, p)$

$$p(x) = \binom{x+r-1}{x} p^r q^{x-r} \text{ for } x = 0, 1, 2, \ldots$$
$$\mathbb{E}(X) = \frac{r(1-p)}{p}$$
$$\mathbb{V}(X) = \frac{r(1-p)}{p^2}$$



Poisson Distribution: $X \sim \mathrm{Pois}(\lambda)$

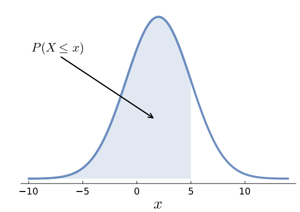
$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ for } x = 0, 1, 2, \ldots$$
$$\mathbb{E}(X) = \lambda$$
$$\mathbb{V}(X) = \lambda$$



Continuous Distributions

Normal Distribution: $X \sim \mathrm{N}(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2} \text{ for } -\infty < x < \infty$$
$$\mathbb{E}(X) = \mu$$
$$\mathbb{V}(X) = \sigma^2$$



If $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Y = c + d \cdot X$ then

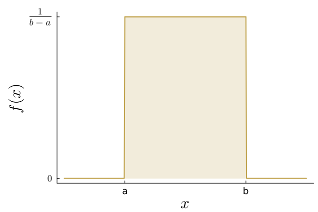
$$Y \sim N\left(c + d \cdot \mu,\, d^2 \cdot \sigma^2\right)$$

If $X \sim \mathrm{N}(\mu, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

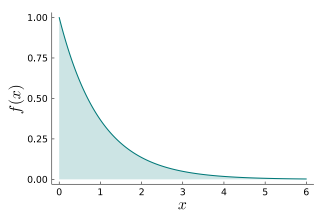
Uniform distribution: $X \sim \mathrm{U}(a,b)$

$$f(x) = \frac{1}{b-a}, \text{ for } -a < x < b$$
$$\mathbb{E}(X) = (a+b)/2$$
$$\mathbb{V}(X) = (b-a)^2/12$$



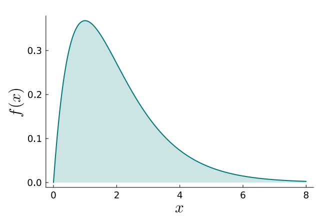
Exponential distribution: $X \sim \mathrm{Exp}(\beta)$

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, \text{ for } x \geq 0$$
$$\mathbb{E}(X) = \beta$$
$$\mathbb{V}(X) = \beta^2$$



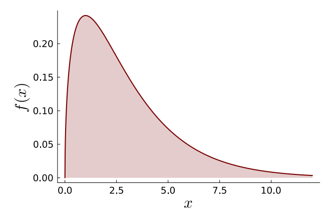
Gamma distribution $X \sim \mathrm{Gamma}(\alpha, \beta)$

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \text{ for } x \geq 0$$
$$\mathbb{E}(X) = \alpha\beta$$
$$\mathbb{V}(X) = \alpha\beta^2$$



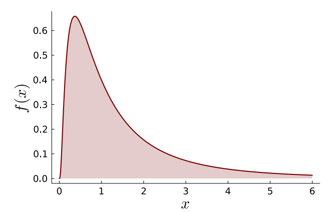
χ^2-distribution $X \sim \mathrm{Chi2}(\nu)$

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} \text{ for } x \geq 0$$
$$\mathbb{E}(X) = \nu$$
$$\mathbb{V}(X) = 2\nu$$



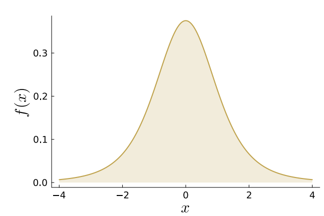
Log-normal distribution $X \sim \mathrm{LogNormal}(\mu, \sigma^2)$

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(\log(x)-\mu)^2} \text{ for } 0 < x < \infty$$
$$\mathbb{E}(X) = \exp(\mu + \sigma^2/2)$$
$$\mathbb{V}(X) = \left(\exp(\sigma^2) - 1\right)\exp(2\mu + \sigma^2)$$



Student t-distribution $X \sim t(\nu)$

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} \text{ for } x \geq 0$$
$$\mathbb{E}(X) = 0 \text{ if } \nu > 1$$
$$\mathbb{V}(X) = \frac{\nu}{\nu-2} \text{ if } \nu > 2$$

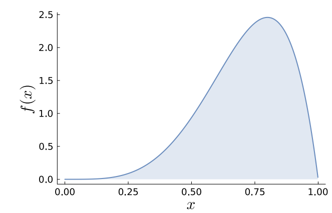


Beta distribution $X \sim \mathrm{Beta}(\alpha, \beta)$

$$f(x) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \text{ for } 0 \leq x \leq 1$$
$$\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$$
$$\mathbb{V}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

where

$$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$



Multivariate normal distribution

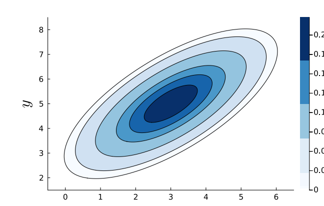
$$(Y_1, Y_2, \ldots, Y_p)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where μ is the p-element mean vector and Σ is the $p \times p$ covariance matrix.
In the bivariate case with $p = 2$:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

and ρ_{12} is the correlation between Y_1 and Y_2.



Maximum likelihood estimation

Log-likelihood $\ell(\theta)$ for discrete variables

If $Y_1, Y_2, \ldots, Y_n$ are *iid* with probability function $p(y_i \mid \theta)$

$$\ell(\theta) = \sum_{i=1}^{n} \log p(y_i \mid \theta)$$

Log-likelihood $\ell(\theta)$ for continuous variables

If $Y_1, Y_2, \ldots, Y_n$ are *iid* with density function $f(y_i \mid \theta)$

$$\ell(\theta) = \sum_{i=1}^{n} \log f(y_i \mid \theta)$$

Maximum likelihood estimator $\hat{\theta}$

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta)$$

Observed information

$$\mathcal{J}_n(\hat{\theta}) = -\ell''(\hat{\theta})$$

Fisher information

$$\mathcal{I}_n(\theta) = \mathbb{E}_{(Y_1,\ldots,Y_n)}\left(-\ell''(\theta)\right)$$

Approximate sampling distribution of the MLE

Informally, for large n

$$\hat{\theta} \overset{\text{approx}}{\sim} N\left(\theta, \mathcal{I}^{-1}(\theta)\right)$$

Linear Gaussian regression model

Regression model

For the ith observation

$$y_i = \mathbf{x}_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$$

where $\mathbf{x}_i$ is a p-element vector with covariate/features.

For all n observations

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma_\varepsilon^2 \cdot \mathbf{I}_n)$$

Least squares/Maximum likelihood estimate

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Estimate of error variance σ_ε^2

$$s_e^2 = \frac{\mathbf{e}^\top \mathbf{e}}{n-p}$$

where $\mathbf{e}$ is the n-element vector with residuals

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\beta$$

Estimated sampling distribution

$$\hat{\beta} \sim N(\beta, s_e^2 (\mathbf{X}^\top \mathbf{X})^{-1})$$

Prediction for $\mathbf{x} = \mathbf{x}_i$

$$\hat{y}_i = \mathbf{x}_i^\top \hat{\beta}$$

Non-Gaussian regression models

Logistic regression

For the ith observation

$$y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(p_i)$$

where

$$p_i = \Pr(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{x}_i^\top \beta}}$$

and $\mathbf{x}_i$ is a p-element vector with covariate/features.

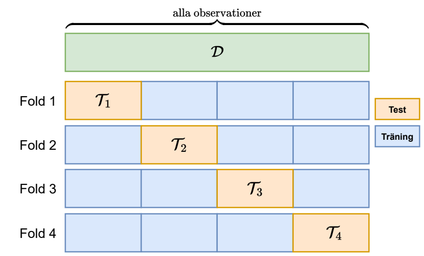
Poisson regression

For the ith observation

$$y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \text{Poisson}\left(\exp(\mathbf{x}_i^\top \beta)\right).$$

Cross validation

The observations of the data $\mathcal{D} = \{1, 2, \ldots, n\}$ are split into K parts, where each observation belongs to exactly one part.



Estimation of the predictive power on new data:

$$\text{SSE}_{\text{cv}} = \sum_{i \in \mathcal{T}_1} \left(y_i - \hat{y}_i^{(1)}\right)^2 + \ldots + \sum_{i \in \mathcal{T}_K} \left(y_i - \hat{y}_i^{(K)}\right)^2,$$

$$\text{RMSE}_{\text{cv}} = \sqrt{\frac{\text{SSE}_{\text{cv}}}{n}},$$

- $\mathcal{T}_k \subset \mathcal{D}$ are all observations that are *testdata* in fold k
- $\sum_{i \in \mathcal{T}_k}$ is the sum over all testdata in fold k
- $\hat{y}_i^{(k)}$ is the prediction of y_i in fold k from a model estimated on all data *except* testdata in $\mathcal{T}_k$.

Created by Mattias Villani, Stockholm University.