

---

# Statistics

## Collection of Formulas

---

## Contents

<b>1</b>	<b>Deskriptive Statistics</b>	<b>2</b>		<b>2</b>	3.4 Sufficiency und Consistency . . . . .	10
1.1	Summary Statistics . . . . .	2				
1.1.1	Location . . . . .	2	<b>4</b>	<b>Statistical Hypothesis Testing</b>		<b>10</b>
1.1.2	Dispersion . . . . .	2	4.1	Significance and Confidence Intervals . . . . .	10	
1.1.3	Concentration . . . . .	2	4.2	Tests for One Sample . . . . .	11	
1.1.4	Shape . . . . .	3	4.3	Tests for Two Samples . . . . .	11	
1.1.5	Dependence . . . . .	3	4.4	Tests for Goodness of Fit . . . . .	11	
1.2	Tables . . . . .	4	4.5	Multiple Tests . . . . .	12	
1.3	Diagrams . . . . .	4	<b>5</b>	<b>Regression</b>		<b>12</b>
1.3.1	Histogram . . . . .	4	5.1	Models . . . . .	12	
1.3.2	QQ-Plot . . . . .	4	5.1.1	Simple Linear Model . . . . .	12	
1.3.3	Scatterplot . . . . .	4	5.1.2	Multivariate Linear Model . . . . .	13	
<b>2</b>	<b>Probability</b>	<b>4</b>	5.1.3	Bayesian Linear Model . . . . .	13	
2.1	Combinatorics . . . . .	4	5.1.4	Quantile Regression . . . . .	14	
2.2	Probability Theory . . . . .	4	5.2	Analysis of Variances (ANOVA) . . . . .	14	
2.3	Random Variables/Vectors . . . . .	5	5.3	Goodness of Fit . . . . .	14	
2.4	Probability Distributions . . . . .	5	5.3.1	Coefficient of Determination . . . . .	14	
2.4.1	Discrete Distributions . . . . .	6	<b>6</b>	<b>Classification</b>		<b>14</b>
2.4.2	Continuous Distributions . . . . .	6	6.1	Diskriminant Analysis (Bayes) . . . . .	14	
2.4.3	Exponential Family . . . . .	7	<b>7</b>	<b>Cluster Analysis</b>		<b>14</b>
2.5	Limit Theorems . . . . .	7	<b>8</b>	<b>Bayesian Statistics</b>		<b>14</b>
<b>3</b>	<b>Inference</b>	<b>8</b>	8.1	Basics . . . . .	14	
3.1	Method of Moments . . . . .	8	8.2	Numerical Methods for the Posterior . . . . .	15	
3.2	Loss Functions . . . . .	8				
3.3	Maximum Likelihood (ML) . . . . .	9				

# 1 Deskriptive Statistics

## 1.1 Summary Statistics

### 1.1.1 Location

**Mode** Most frequent value of  $x_i$ . Two or more modes are possible (bimodal).

**Median**

$$\tilde{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{falls } n \text{ gerade} \end{cases}$$

**Quantile**

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{falls } n\alpha \notin \mathbb{N} \\ \frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)}) & \text{falls } n\alpha \text{ ganzzahlig} \end{cases}$$

with

$$k = \min x \in \mathbb{N}, \quad x > n\alpha$$

**Minimum/Maximum**

$$x_{\min} = \min_{i \in \{1, \dots, N\}} (x_i) \quad x_{\max} = \max_{i \in \{1, \dots, N\}} (x_i)$$

### 1.1.2 Dispersion

**Range**

$$R = x_{(n)} - x_{(1)}$$

**Interquartile Range**

$$d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

**(Empirical) Variance**

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

Estimates the second centralized moment.

*Calculation Rules:*

$$\star \operatorname{Var}(aX + b) = a^2 \cdot \operatorname{Var}(X)$$

### 1.1.3 Concentration

**Gini Coefficient**

$$G = \frac{2 \sum_{i=1}^n i x_{(i)} - (n+1) \sum_{i=1}^n x_{(i)}}{n \sum_{i=1}^n x_{(i)}} = 1 - \frac{1}{n} \sum_{i=1}^n (v_{i-1} + v_i)$$

with

$$u_i = \frac{i}{n}, \quad v_i = \frac{\sum_{j=1}^i x_{(j)}}{\sum_{j=1}^n x_{(j)}} \quad (u_0 = 0, \quad v_0 = 0)$$

**Arithmetic Mean**

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

Estimates the expectation  $\mu = E[X]$  (first moment).

*Calculation Rules:*

$$\star E(a + b \cdot X) = a + b \cdot E(X)$$

$$\star E(X \pm Y) = E(X) \pm E(Y)$$

**Geometric Mean**

$$\bar{x}_G = \sqrt[n]{\sum_{i=1}^n x_i}$$

For growth factors:  $\bar{x}_G = \sqrt[n]{\frac{B_n}{B_0}}$

**Harmonic Mean**

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

$$\star \operatorname{Var}(X \pm Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y)$$

**(Empirical) Standard Deviation**

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

**Coefficient of Variation**

$$\nu = \frac{s}{\bar{x}}$$

**Average Absolute Deviation**

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

Estimates the first absolute centralized moment.

These are also the values for the Lorenz curve.

$$\text{Range: } 0 \leq G \leq \frac{n-1}{n}$$

**Lorenz-Münzner Coefficient (normed  $G$ )**

$$G^+ = \frac{n}{n-1} G$$

$$\text{Range: } 0 \leq G^+ \leq 1$$

### 1.1.4 Shape

#### (Empirical) Skewness

$$\nu = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3$$

Estimates the third centralized moment, scaled with  $(\sigma^2)^{\frac{2}{3}}$

### 1.1.5 Dependence

#### for two nominal variables

##### $\chi^2$ -Statistic

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - \frac{n_{i+}n_{+j}}{n})^2}{\frac{n_{i+}n_{+j}}{n}} = n \left( \sum_{i=1}^k \sum_{j=1}^l \frac{n_{ij}^2}{n_{i+}n_{+j}} - 1 \right)$$

Range:  $0 \leq \chi^2 \leq n(\min(k, l) - 1)$

##### Phi-Coefficient

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

Range:  $0 \leq \Phi \leq \sqrt{\min(k, l) - 1}$

##### Cramér's $V$

$$V = \sqrt{\frac{\chi^2}{\min(k, l) - 1}}$$

Range:  $0 \leq V \leq 1$

##### Contingency Coefficient $C$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Range:  $0 \leq C \leq \sqrt{\frac{\min(k, l) - 1}{\min(k, l)}}$

##### Corrected Contingency Coefficient $C_{corr}$

$$C_{corr} = \sqrt{\frac{\min(k, l)}{\min(k, l) - 1}} \cdot \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Range  $0 \leq C_{corr} \leq 1$

##### Odds-Ratio

$$OR = \frac{ad}{bc} = \frac{n_{ii}n_{jj}}{n_{ij}n_{ji}}$$

Range:  $0 \leq OR < \infty$

#### for two ordinal variables

##### Gamma (Goodman and Kruskal)

$$\gamma = \frac{K - D}{K + D}$$

$K = \sum_{i < m} \sum_{j < n} n_{ij}n_{mn}$  Number of concordant pairs

$D = \sum_{i < m} \sum_{j > n} n_{ij}n_{mn}$  Number of reversed pairs

Range:  $-1 \leq \gamma \leq 1$

#### (Empirical) Kurtosis

$$k = \left[ n(n+1) \cdot \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^4 - 3(n-1) \right] \cdot \frac{n-1}{(n-2)(n-3)} + 3$$

Estimates the fourth centralized moment, scaled with  $(\sigma^2)^2$

##### Excess

$$\gamma = k - 3$$

#### Kendall's $\tau_b$

$$\tau_b = \frac{K - D}{\sqrt{(K + D + T_X)(K + D + T_Y)}}$$

with

$T_X = \sum_{i=m} \sum_{j < n} n_{ij}n_{mn}$  Number of ties w.r.t.  $X$

$T_Y = \sum_{i < m} \sum_{j=n} n_{ij}n_{mn}$  Number of ties w.r.t.  $Y$

Range:  $-1 \leq \tau_b \leq 1$

#### Kendall's/Stuart's $\tau_c$

$$\tau_c = \frac{2 \min(k, l)(K - D)}{n^2(\min(k, l) - 1)}$$

Range:  $-1 \leq \tau_c \leq 1$

#### Spearman's Rank Correlation Coefficient

$$\rho = \frac{n(n^2 - 1) - \frac{1}{2} \sum_{j=1}^J b_j(b_j^2 - 1) - \frac{1}{2} \sum_{k=1}^K c_k(c_k^2 - 1) - 6 \sum_{i=1}^n d_i^2}{\sqrt{n(n^2 - 1) - \sum_{j=1}^J b_j(b_j^2 - 1)} \sqrt{n(n^2 - 1) - \sum_{k=1}^K c_k(c_k^2 - 1)}}$$

or

$$\rho = \frac{s_{rg_x} r_{g_y}}{\sqrt{s_{rg_x} r_{g_x} s_{rg_y} r_{g_y}}}$$

Without ties:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

with

$d_i = R(x_i) - R(y_i)$  rank difference

Range:  $-1 \leq \rho \leq 1$

#### for two metric variables

##### Correlation Coefficient (Bravais-Pearson)

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$$

with

$$s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})^2 \quad \text{or} \quad s_{xy} = \frac{s_{xy}}{n}$$

$$s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{or} \quad s_{xx} = \frac{s_{xx}}{n}$$

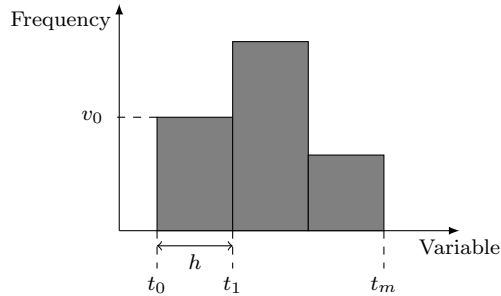
$$s_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{or} \quad s_{yy} = \frac{s_{yy}}{n}$$

Range:  $-1 \leq r \leq 1$

## 1.2 Tables

## 1.3 Diagrams

### 1.3.1 Histogram



sample:  $X = \{x_1, x_2, \dots, x_n\}$

$k$ -th bin:  $B_k = [t_k, t_{k+1})$ ,  $k = \{0, 1, \dots, m-1\}$

Number of observations in the  $k$ -th bin:  $v_k$

bin width:  $h = t_{k+1} - t_k, \forall k$

**Scott's Rule**

$$h^* \approx 3.5\sigma n^{-\frac{1}{3}}$$

For approximately normal distributed data (min. MSE)

### 1.3.2 QQ-Plot

### 1.3.3 Scatterplot

## 2 Probability

### 2.1 Combinatorics

	without replacement	with replacement
Permutations	$n!$	$\frac{n!}{n_1! \cdots n_s!}$
Combinations: without order	$\binom{n}{m}$	$\binom{n+m-1}{m}$
with order	$\binom{n}{m} m!$	$n^m$

with:

$$n! = n \cdot (n-1) \cdot \dots \cdot 1$$

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

### 2.2 Probability Theory

**Laplace**

$$P(A) = \frac{|A|}{|\Omega|}$$

**Kolmogorov Axioms** mathematical definition of probability

- (1)  $0 \leq P(A) \leq 1 \quad \forall A \in \mathcal{A}$
- (2)  $P(\Omega) = 1$
- (3)  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$   
 $\forall A_i \in \mathcal{A}, i = 1, \dots, \infty$  with  $A_i \cap A_j = \emptyset$  for  $i \neq j$

Implications:

- $P(\bar{A}) = 1 - P(A)$
- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(B) = \sum_{i=1}^n P(B \cap A_i)$ , for  $A_1, \dots, A_n$  complete decomposition of  $\Omega$  into pairwise disjoint events

**Probability (Mises)** frequentist definition of probability

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A(n)}{n}$$

with  $n$  repetitions of a random experiment and  $n_A(n)$  events  $A$

**Conditional Probability**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{für } P(B) > 0$$

$$\Rightarrow P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

**Law of Total Probability**

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$$

**Bayes' Theorem**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{for } P(A), P(B) > 0$$

**Stochastic Independence**

$$A, B \text{ independent} \Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$

$$X, Y \text{ independent} \Leftrightarrow f_{XY}(x, y) = f_X(x) \cdot f_Y(y) \quad \forall x, y$$

## 2.3 Random Variables/Vectors

### Random Variables $\in \mathbb{R}$

#### Definition

$$Y : \Omega \rightarrow \mathbb{R}$$

The Subset of possible values for  $\mathbb{R}$  is called support.

Notation: Realisations of  $Y$  are depicted with lower case letters.

$Y = y$  means, that  $y$  is the realisation of  $Y$ .

#### Discrete and Continuous Random Variables

If the support is uncountably infinite, the random variable is called *continuous*, otherwise it is called *discrete*.

- **Density  $f(\cdot)$ :**

For continuous variables:  $P(Y \in [a, b]) = \int_a^b f_Y(y) dy$

For discrete variables the density (and other functions) can be depicted like the corresponding function for continuous variables, if the notation is extended as follows:

$\int_{-\infty}^y f_Y(\tilde{y}) d\tilde{y} := \sum_{k: k \leq y} P(Y = k)$ . This notation is used.

- **Cumulative Distribution Function  $F(\cdot)$ :**

$$F_Y(y) = P(Y \leq y)$$

Relationship:

$$F_Y(y) = \int_{-\infty}^y f_Y(\tilde{y}) d\tilde{y}$$

#### Moments

- **Expectation (1. Moment):**  $\mu = E(Y) = \int y f_Y(y) dy$

- **Variance (2. centralized Moment):**

$$\sigma^2 = Var(Y) = E(\{Y - E(Y)\}^2) = \int (y - E(Y))^2 f(y) dy$$

Note:  $E(\{Y - \mu\}^2) = E(Y^2) - \mu^2$

Proof:

$$E(\{Y - \mu\}^2) = E(Y^2 - 2Y\mu + \mu^2) = E(Y^2) - 2\mu^2 + \mu^2 = E(Y^2) - \mu^2$$

- **kth Moment:**  $E(Y^k) = \int y^k f_Y(y) dy$ ,

**k. centralized Moment:**  $E(\{Y - E(Y)\}^k)$

#### Moment Generating Function

$$M_Y(t) = E(e^{tY})$$

$$\text{with } \left. \frac{\partial^k M_Y(t)}{\partial t^k} \right|_{t=0} = E(Y^k)$$

Cumulant Generating Function  $K_Y(t) = \log M_Y(t)$

A random variable is uniquely defined by its moment generating function and vice versa (as long as moments and cumulants are finite).

### Random Vectors $\in \mathbb{R}^q$

#### Density and Cumulative Distribution Function

$$F(y_1, \dots, y_q) = P(Y_1 \leq y_1, \dots, Y_q \leq y_q)$$

$$P(a_1 \leq Y_1 \leq b_1, \dots, a_q \leq Y_q \leq b_q)$$

$$= \int_{a_1}^{b_1} \dots \int_{a_q}^{b_q} f(y_1, \dots, y_q) dy_1 \dots dy_q$$

#### Marginal Density

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(y_1, \dots, y_k) dy_2 \dots dy_k$$

#### Conditional Density

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f(y_1, \dots, y_2)}{f(y_2)} \text{ for } f(y_2) > 0$$

#### Iterated Expectation

$$E(Y) = E_X(E(Y|X))$$

Proof:

$$E(Y) = \int y f(y) dy = \int \int y f(y|x) dy f_X(x) dx = E_X(E(Y|X))$$

$$Var(Y) = E_X(Var(Y|X)) + Var_X(E(Y|X))$$

Proof:

$$\begin{aligned} Var(Y) &= \int (y - \mu_Y)^2 f(y) dy \\ &= \int (y - \mu_Y)^2 f(y|x) f(x) dy dx \\ &= \int (y - \mu_{Y|x} + \mu_{Y|x} - \mu_Y)^2 f(y|x) f(x) dy dx \\ &= \int (y - \mu_{Y|x})^2 f(y|x) f(x) dy dx + \\ &\quad \int (\mu_{Y|x} - \mu_Y)^2 f(y|x) f(x) dy dx + \\ &\quad 2 \int (y - \mu_{Y|x})(\mu_{Y|x} - \mu_Y) f(y|x) f(x) dy dx \\ &= \int Var(Y|x) f(x) dx + \int (\mu_{Y|x} - \mu_Y)^2 f(x) dx \\ &= E_X(Var(Y|X)) + Var_X(E(Y|X)) \end{aligned}$$

## 2.4 Probability Distributions

## 2.4.1 Discrete Distributions

### Discrete Uniform

$$Y \sim U(\{y_1, \dots, y_k\}), y \in \{y_1, \dots, y_k\}$$

$$P(Y = y_i) = \frac{1}{k}, i = 1, \dots, k$$

$$E(Y) = \frac{k+1}{2}, \text{Var}(Y) = \frac{k^2-1}{12}$$

### Binomial Successes in independent trials

$$Y \sim \text{Bin}(n, \pi) \text{ with } n \in \mathbb{N}, \pi \in [0, 1], y \in \{0, \dots, n\}$$

$$P(Y = y|\lambda) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

$$E(Y|\pi, n) = n\pi, \text{Var}(Y|\pi, n) = n\pi(1-\pi)$$

### Poisson Counting model for rare events

only one event at a time, no autocorrelation, mean number of events over time is constant and proportional to length of the considered time interval

$$Y \sim \text{Po}(\lambda) \text{ with } \lambda \in [0, +\infty], y \in \mathbb{N}_0$$

## 2.4.2 Continuous Distributions

### Continuous Uniform

$$Y \sim U(a, b) \text{ with } a, b \in \mathbb{R}, a \leq b, y \in [a, b]$$

$$p(y|a, b) = \frac{1}{b-a}$$

$$E(Y|a, b) = \frac{a+b}{2}, \text{Var}(Y|a, b) = \frac{(b-a)^2}{12}$$

### Univariate Normal symmetric with $\mu$ and $\sigma^2$

$$Y \sim N(\mu, \sigma^2) \text{ with } \mu \in \mathbb{R}, \sigma^2 > 0, y \in \mathbb{R}$$

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

$$E(Y|\mu, \sigma^2) = \mu, \text{Var}(Y|\mu, \sigma^2) = \sigma^2$$

### Multivariate Normal symmetric with $\mu_i$ and $\Sigma$

$$Y \sim N(\mu, \Sigma) \text{ with } \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d} \text{ s.p.d.}, y \in \mathbb{R}^d$$

$$p(y|\mu, \Sigma) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)\right)$$

$$E(Y|\mu, \Sigma) = \mu, \text{Var}(Y|\mu, \Sigma) = \Sigma$$

### Log-Normal

$$Y \sim \text{LogN}(\mu, \sigma^2) \text{ with } \mu \in \mathbb{R}, \sigma^2 > 0, y > 0$$

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 y}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right)$$

$$E(Y|\mu, \sigma^2) = \exp\left(\mu + \frac{\sigma^2}{2}\right),$$

$$\text{Var}(Y|\mu, \sigma^2) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$$

Relationship:  $\log(Y) \sim N(\mu, \sigma^2) \Rightarrow Y \sim \text{LogN}(\mu, \sigma^2)$

**non-standardized Student's t** statistical Tests for  $\mu$  with unknown (estimated) variance and  $\nu$  degrees of freedom

$$P(Y = y|\lambda) = \frac{\lambda^y \exp^{-\lambda}}{y!}$$

$$E(Y|p) = \lambda, \text{Var}(Y|p) = \lambda$$

The model tends to overestimate the variance (Overdispersion).

*Approximation* of the Binomial for small p

### Geometric

$$Y \sim \text{Geom}(\pi) \text{ with } \pi \in [0, 1], y \in \mathbb{N}_0$$

$$P(Y = y|\pi) = \pi(1-\pi)^{y-1}$$

$$E(Y|\pi) = \frac{1}{\pi}, \text{Var}(Y|\pi) = \frac{1-\pi}{\pi^2}$$

### Negative Binomial

$$Y \sim \text{NegBin}(\alpha, \beta) \text{ with } \alpha, \beta \geq 0, y \in \mathbb{N}_0$$

$$P(Y = y|\alpha, \beta) = \binom{\alpha+y-1}{\alpha-1} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^y$$

$$E(Y|\alpha, \beta) = \frac{\alpha}{\beta}, \text{Var}(Y|\alpha, \beta) = \frac{\alpha}{\beta^2}(\beta+1)$$

$$Y \sim t_\nu(\mu, \sigma^2) \text{ with } \mu \in \mathbb{R}, \sigma^2, \nu > 0, y \in \mathbb{R}$$

$$p(y|\mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\Gamma(\sqrt{\nu\sigma^2})} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

$$E(Y|\mu, \sigma^2, \nu) = \mu \text{ for } \nu > 1,$$

$$\text{Var}(Y|\mu, \sigma^2, \nu) = \sigma^2 \frac{\nu}{\nu-2} \text{ for } \nu > 2$$

Relationship:  $Y|\theta \sim N(\mu, \frac{\sigma^2}{\theta}), \theta \sim \text{Ga}(\frac{\nu}{2}, \frac{\nu}{2}) \Rightarrow Y \sim t_\nu(\mu, \sigma)$   
 $t_\nu(\mu, \sigma^2)$  has heavier tails than the normal distribution.  
 $t_\infty(\mu, \sigma^2)$  approaches  $N(\mu, \sigma^2)$ .

### Beta

$$Y \sim \text{Be}(a, b) \text{ with } a, b > 0, y \in [0, 1]$$

$$p(y|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1} (1-y)^{b-1}$$

$$E(Y|a, b) = \frac{a}{a+b},$$

$$\text{Var}(Y|a, b) = \frac{ab}{(a+b)^2(a+b+1)},$$

$$\text{mod}(Y|a, b) = \frac{a-1}{a+b-2} \text{ for } a, b > 1$$

### Gamma

$$Y \sim \text{Ga}(a, b) \text{ with } a, b > 0, y > 0$$

$$p(y|a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by)$$

$$E(Y|a, b) = \frac{a}{b},$$

$$\text{Var}(Y|a, b) = \frac{a}{b^2},$$

$$\text{mod}(Y|a, b) = \frac{a-1}{b} \text{ for } a \geq 1$$

### Inverse-Gamma

$$Y \sim \text{IG}(a, b) \text{ with } a, b > 0, y > 0$$

$$p(y|a, b) = \frac{b^a}{\Gamma(a)} y^{-a-1} \exp(-\frac{b}{y})$$

$$E(Y|a, b) = \frac{b}{a-1} \text{ for } a > 1,$$

$$\text{Var}(Y|a, b) = \frac{b^2}{(a-1)^2(a-2)} \text{ for } a \geq 2,$$

$$\text{mod}(Y|a, b) = \frac{b}{a+1}$$

Relationship:  $Y^{-1} \sim \text{Ga}(a, b) \Leftrightarrow Y \sim \text{IG}(a, b)$

**Exponential** Time between Poisson events

$$Y \sim \text{Exp}(\lambda) \text{ with } \lambda > 0, y \geq 0$$

### 2.4.3 Exponential Family

#### Definition

The exponential family comprises all distributions, whose density can be written as follows:

$$f_Y(y, \theta) = \exp^{t^T(y)\theta - \kappa(\theta)} h(y)$$

with  $h(y) \geq 0$ ,  $t(y)$  vector of the canonical statistic,  $\theta$  as parameter and  $\kappa(\theta)$  the normalising constant.

#### Normalising Constant

$$1 = \int \exp^{t^T(y)\theta} h(y) dy \exp^{-\kappa(\theta)}$$

$$\Leftrightarrow \kappa(\theta) = \log \int \exp^{t^T(y)\theta} h(y) dy$$

$\kappa(\theta)$  is the cumulant generating function, therefore  $\frac{\partial \kappa(\theta)}{\partial \theta} = E(t(Y))$  and  $\frac{\partial^2 \kappa(\theta)}{\partial \theta^2} = \text{Var}(t(Y))$

## 2.5 Limit Theorems

#### Law of Large Numbers

#### Central Limit Theorem

$$Z_n \xrightarrow{d} N(0, \sigma^2)$$

with  $Z_n = \sum_{i=1}^n \frac{Y_i}{\sqrt{n}}$  and  $Y_i$  i.i.d. with expectation 0 and variance  $\sigma^2$

$$p(y|\lambda) = \lambda \exp(-\lambda y)$$

$$E(Y|\lambda) = \frac{1}{\lambda}, \text{Var}(Y|\lambda) = \frac{1}{\lambda^2}$$

**Chi-Squared** squared standard normal random variables with  $\nu$  degrees of freedom

$$Y \sim \chi^2(\nu) \text{ with } \nu > 0, y \in \mathbb{R}$$

$$p(y|\nu) = \frac{y^{\frac{\nu}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})}$$

$$E(Y|\nu) = \nu, \text{Var}(Y|\nu) = 2\nu$$

#### Members

- **Poisson**
- **Geometric**
- **Exponential**
- **Normal**  $t(y) = \left(-\frac{y^2}{2}, y\right)^T$ ,  $\theta = \left(\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}\right)^T$ ,  $h(y) = \frac{1}{\sqrt{2\pi}}$ ,  $\kappa(\theta) = \frac{1}{2} \left(-\log \frac{1}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right)$
- **Gamma**
- **Chi-Squared**
- **Beta**

Proof:

For normal random variables  $Z \sim N(\mu, \sigma^2)$ :  $K_Z(t) = \mu t + \frac{1}{2} \sigma^2 t^2$ . The first two derivatives  $\left. \frac{\partial^k K_Z(t)}{\partial t^k} \right|_{t=0}$  are  $\mu$  and  $\sigma$ . All other moments are zero.

For  $Z_n = (Y_1 + Y_2 + \dots + Y_n)/\sqrt{n}$ :

$$\begin{aligned} M_{Z_n}(t) &= E \left( e^{t(Y_1 + Y_2 + \dots + Y_n)/\sqrt{n}} \right) \\ &= E \left( e^{tY_1/\sqrt{n}} \cdot e^{tY_2/\sqrt{n}} \cdot \dots \cdot e^{tY_n/\sqrt{n}} \right) \\ &= E \left( e^{tY_1/\sqrt{n}} \right) E \left( e^{tY_2/\sqrt{n}} \right) \dots E \left( e^{tY_n/\sqrt{n}} \right) \\ &= M_Y^n(t/\sqrt{n}) \end{aligned}$$

Analogously:  $K_{Z_n}(t) = nK_Y(t/\sqrt{n})$ .

$$\begin{aligned} \left. \frac{\partial K_{Z_n}(t)}{\partial t} \right|_{t=0} &= \frac{n}{\sqrt{n}} \left. \frac{\partial K_Y(t)}{\partial t} \right|_{t=0} = \sqrt{n} \mu \\ \left. \frac{\partial^2 K_{Z_n}(t)}{\partial t^2} \right|_{t=0} &= \frac{n}{n} \left. \frac{\partial^2 K_Y(t)}{\partial t^2} \right|_{t=0} = \sigma^2 \end{aligned}$$

Using the Taylor Expansion, we can write  $K_{Z_n}(t) = 0 + \sqrt{n} \mu t + \frac{1}{2} \sigma^2 t^2 + \dots$ , where the terms in  $\dots$  are tending towards 0 as  $n \rightarrow \infty$ .

Therefore:  $K_{Z_n}(t) \xrightarrow{n \rightarrow \infty} K_Z(t)$  with  $Z \sim N(\sqrt{n} \mu, \sigma^2)$ .

## 3 Inference

### 3.1 Method of Moments

The theoretical moments are estimated by their empirical counterparts:

$$E_{\hat{\theta}_{MM}}(Y^k) = m_k(y_1, \dots, y_n)$$

For the exponential family:  $\hat{\theta}_{MM} = \hat{\theta}_{ML}$

### 3.2 Loss Functions

**Loss**

$$\mathcal{L} : \mathcal{T} \times \Theta \rightarrow \mathbb{R}^+$$

with parameter space  $\Theta \subset \mathbb{R}$ ,  $t \in \mathcal{T}$  with  $t : \mathbb{R}^n \rightarrow \mathbb{R}$  a statistic, that estimates the parameter  $\theta$ ,  $\mathcal{L}(\theta, \theta) = 0$  holds

- **absolute loss (L1):**  $\mathcal{L}(t, \theta) = |t - \theta|$
- **quadratic loss (L2):**  $\mathcal{L}(t, \theta) = (t - \theta)^2$

As  $\theta$  is unknown, the loss is a theoretical measure. Additionally, it is the realisation of a random variable as it is dependent on a concrete sample.

**Risiko**

$$\begin{aligned} R(t(\cdot), \theta) &= E_{\theta}(\mathcal{L}(t(Y_1, \dots, Y_n), \theta)) \\ &= \int_{-\infty}^{\infty} \mathcal{L}(t(Y_1, \dots, Y_n), \theta) \prod_{i=1}^n f(y_i; \theta) dy_i \end{aligned}$$

**Minimax Approach**

The risk still depends on the true parameter  $\theta$ . Tentative estimation: Choose  $\theta$ , so that the risk is maximal and then  $t(\cdot)$ , so that the risk is minimized (minimizing the worst case):

$$\hat{\theta}_{minimax} = \arg \min_{t(\cdot)} \left( \max_{\theta \in \Theta} R(t(\cdot); \theta) \right)$$

**Mean Squared Error (MSE)**

$$\begin{aligned} MSE(t(\cdot), \theta) &= E_{\theta}(\{t(Y) - \theta\}^2) \\ &= \text{Var}_{\theta}(t(Y_1, \dots, Y_n)) + \text{Bias}^2(t(\cdot); \theta) \end{aligned}$$

with  $\text{Bias}(t(\cdot); \theta) = E_{\theta}(t(Y_1, \dots, Y_n)) - \theta$

Proof:

Let  $\mathcal{L}(t, \theta) = (t - \theta)^2$

$$\begin{aligned} R(t(\cdot), \theta) &= E_{\theta}(\{t(Y) - \theta\}^2) \\ &= E_{\theta}(\{t(Y) - E_{\theta}(t(Y)) + E_{\theta}(t(Y)) - \theta\}^2) \\ &= E_{\theta}(\{t(Y) - E_{\theta}(t(Y))\}^2) + E_{\theta}(\{E_{\theta}(t(Y)) - \theta\}^2) \\ &\quad + 2E_{\theta}(\{t(Y) - E_{\theta}(t(Y))\}\{E_{\theta}(t(Y)) - \theta\}) \\ &= \text{Var}_{\theta}(t(Y_1, \dots, Y_n)) + \text{Bias}^2(t(\cdot); \theta) + 0 \end{aligned}$$

**Cramér-Rao Inequality**

$$MSE(\hat{\theta}, \theta) \geq \text{Bias}^2(\hat{\theta}, \theta) + \frac{\left(1 + \frac{\partial \text{Bias}(\hat{\theta}, \theta)}{\partial \theta}\right)^2}{I(\theta)}$$

Proof:

For unbiased estimates:  $\theta = E_{\theta}(\hat{\theta}) = \int t(y)f(y; \theta)dy$

$$\begin{aligned} 1 &= \int t(y) \frac{\partial f(y; \theta)}{\partial \theta} dy \\ &= \int t(y) \frac{\partial \log f(y; \theta)}{\partial \theta} f(y; \theta) dy \\ &= \int t(y) s(y; \theta) f(y; \theta) dy \\ &= \int (t(y) - \theta) (s(y; \theta) - 0) f(y; \theta) dy \quad \begin{array}{l} \text{1. Bartlett equation} \\ E_{\theta}(s(\theta; y)) = 0 \end{array} \\ &= \text{Cov}_{\theta}(t(Y); s(\theta; Y)) \\ &\geq \sqrt{\text{Var}_{\theta}(t(Y))} \sqrt{\text{Var}_{\theta}(s(\theta; Y))} \quad \text{Cauchy-Schwarz} \\ &= \sqrt{MSE(t(Y); \theta)} \sqrt{I(\theta)} \end{aligned}$$

**Kullback-Leibler Divergence** Comparing distributions

$$KL(t, \theta) = \int_{-\infty}^{\infty} \log \frac{f(\tilde{y}; \theta)}{f(\tilde{y}; t)} f(\tilde{y}; \theta) d\tilde{y}$$

The KL divergence is not a distance as it is not symmetric. It is 0 for  $t = \theta$  and  $\geq 0$  otherwise.

Proof:

Follows from  $\log(x) \leq x - 1 \forall x \geq 0$ , with equality for  $x = 1$ .

$R_{KL}(t(\cdot), \theta)$  is approximated by the MSE.

Proof:

$$\begin{aligned} R_{KL}(t(\cdot), \theta) &= \\ &= \int_{-\infty}^{\infty} \mathcal{L}_{KL}(t(Y_1, \dots, Y_n), \theta) \prod_{i=1}^n f(y_i; \theta) dy_i \\ &= \int \int \log \frac{f(\tilde{y}; \theta)}{f(\tilde{y}; t)} f(\tilde{y}; \theta) d\tilde{y} \prod_{i=1}^n f(y_i; \theta) dy_i \\ &= \int \int (\log f(\tilde{y}; \theta) - \log f(\tilde{y}; t)) f(\tilde{y}; \theta) d\tilde{y} - \prod_{i=1}^n f(y_i; \theta) dy_i \\ &\approx - \int \underbrace{\left( \int \frac{\partial \log f(\tilde{y}; \theta)}{\partial \theta} f(\tilde{y}; \theta) d\tilde{y} \right)}_0 (t - \theta) \prod_{i=1}^n f(y_i; \theta) dy_i \\ &\quad + \frac{1}{2} \int \underbrace{\left( - \int \frac{\partial^2 \log f(\tilde{y}; \theta)}{\partial \theta^2} f(\tilde{y}; \theta) d\tilde{y} \right)}_{I(\theta)} (t - \theta)^2 \prod_{i=1}^n f(y_i; \theta) dy_i \end{aligned}$$

The last step is approximated by the Taylor Expansion:  
 $\log f(\tilde{y}, t) \approx \log f(\tilde{y}, \theta) + \frac{\partial \log f(\tilde{y}, \theta)}{\partial \theta} (t - \theta) + \frac{1}{2} \frac{\partial^2 \log f(\tilde{y}, \theta)}{\partial \theta^2} (t - \theta)^2$



### 3.3 Maximum Likelihood (ML)

#### Voraussetzungen

- $Y_i \sim f(y; \theta)$  i.i.d.
- $\theta \in \mathbb{R}^p$
- $f(\cdot; \theta)$  Fisher-regulär:
  - $\{y : f(y; \theta) > 0\}$  unabhängig von  $\theta$
  - Möglicher Parameterraum  $\Theta$  ist offen
  - $f(y; \theta)$  zweimal differenzierbar
  - $\int \frac{\partial}{\partial \theta} f(y; \theta) dy = \frac{\partial}{\partial \theta} \int f(y; \theta) dy$

#### Zentrale Funktionen

- **Likelihood**  $L(\theta; y_1, \dots, y_n): \prod_{i=1}^n f(y_i; \theta)$
- **log-Likelihood**  $l(\theta; y_1, \dots, y_n):$   
 $\log L(\theta; y_1, \dots, y_n) = \sum_{i=1}^n \log f(y_i; \theta)$
- **Score**  $s(\theta; y_1, \dots, y_n): \frac{\partial l(\theta; y_1, \dots, y_n)}{\partial \theta}$
- **Fisher-Information**  $I(\theta): -E_{\theta} \left( \frac{\partial s(\theta; Y)}{\partial \theta} \right)$
- **beobachtete Fisher-Information**  $I_{obs}(\theta):$   
 $-E_{\theta} \left( \frac{\partial s(\theta; y)}{\partial \theta} \right)$

#### Eigenschaften der Score-Funktion

erste Bartlett-Gleichung:

$$E(s(\theta; Y)) = 0$$

Proof:

$$\begin{aligned} 1 &= \int f(y; \theta) dy \\ 0 &= \frac{\partial 1}{\partial \theta} = \int \frac{\partial f(y; \theta)}{\partial \theta} dy = \int \frac{\partial f(y; \theta) / \partial \theta}{f(y; \theta)} f(y; \theta) dy \\ &= \int \frac{\partial}{\partial \theta} \log f(y; \theta) f(y; \theta) dy = \int s(\theta; y) f(y; \theta) dy \end{aligned}$$

zweite Bartlett-Gleichung:

$$\text{Var}_{\theta}(s(Y; \theta)) = E_{\theta} \left( -\frac{\partial^2 \log f(Y; \theta)}{\partial \theta^2} \right) = I(\theta)$$

Proof:

$$\begin{aligned} 0 &= \frac{\partial 0}{\partial \theta} = \frac{\partial}{\partial \theta} \int \frac{\partial}{\partial \theta} \log f(y; \theta) f(y; \theta) dy \quad \text{siehe oben} \\ &= \int \left( \frac{\partial^2}{\partial \theta^2} \log f(y; \theta) \right) f(y; \theta) dy \\ &\quad + \int \frac{\partial \log f(y; \theta)}{\partial \theta} \frac{\partial f(y; \theta)}{\partial \theta} dy \\ &= E_{\theta} \left( \frac{\partial^2}{\partial \theta^2} \log f(Y; \theta) \right) \\ &\quad + \int \frac{\partial \log f(y; \theta)}{\partial \theta} \frac{\partial \log f(y; \theta)}{\partial \theta} f(y; \theta) dy \end{aligned}$$

$$\Leftrightarrow E_{\theta}(s(\theta; Y)s(\theta; Y)) = E_{\theta} \left( -\frac{\partial^2}{\partial \theta^2} \log f(Y; \theta) \right)$$

Bartletts zweite Gleichung gilt dann, weil  $E(s(\theta; Y)) = 0$

#### ML-Schätzer

$$\hat{\theta}_{ML} = \arg \max l(\theta; y_1, \dots, y_n)$$

für Fisher-reguläre Verteilungen:  $\hat{\theta}_{ML}$  hat asymptotisch die kleinstmögliche Varianz, gegeben durch die Cramér-Rao-Ungleichung,  $s(\hat{\theta}_{ML}; y_1, \dots, y_n) = 0$

$$\hat{\theta} \stackrel{a}{\sim} N(\theta, I^{-1}(\theta))$$

Der ML-Schätzer ist invariant:  $\hat{\gamma} = g(\hat{\theta})$  wenn  $\gamma = g(\theta)$ .

Proof:

$$\gamma = g(\theta) \Leftrightarrow \theta = g^{-1}(\gamma)$$

Für die Loglikelihood von  $\gamma$  an der Stelle  $\hat{\theta}$  gilt:

$$\frac{\partial l(g^{-1}(\hat{\gamma}))}{\partial \gamma} = \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \underbrace{\frac{\partial l(\hat{\theta})}{\partial \theta}}_{=0} = 0$$

Die Fisher-Information ist dann  $\frac{\partial \theta}{\partial \gamma} I(\theta) \frac{\partial \theta}{\partial \gamma}$

Proof:

$$\begin{aligned} I_{\gamma}(\gamma) &= -E \left( \frac{\partial^2 l(g^{-1}(\hat{\gamma}))}{\partial \gamma^2} \right) = -E \left( \frac{\partial}{\partial \gamma} \left( \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \frac{\partial l(\theta)}{\partial \theta} \right) \right) \\ &= -E \left( \underbrace{\frac{\partial^2 g^{-1}(\gamma)}{\partial \gamma^2} \frac{\partial l(\theta)}{\partial \theta}}_{\text{Erwartungswert 0}} + \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \frac{\partial^2 l(\theta)}{\partial \theta^2} \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \right) \\ &= \frac{\partial g^{-1}(\gamma)}{\partial \gamma} I(\theta) \frac{\partial g^{-1}(\gamma)}{\partial \gamma} = \frac{\partial \theta}{\partial \gamma} I(\theta) \frac{\partial \theta}{\partial \gamma} \end{aligned}$$

Delta-Regel:  $\gamma \stackrel{a}{\sim} N(\hat{\gamma}, \frac{\partial \theta}{\partial \gamma} I^{-1}(\theta) \frac{\partial \theta}{\partial \gamma})$

**Numerical computation of the ML estimate** Fisher-Scoring as statistical version of the Newton-Raphson procedure

1. Initialize  $\theta_{(0)}$
2. Repeat:  $\theta_{(t+1)} := \theta_{(t)} + I^{-1}(\theta_{(t)})s(\theta_{(t)}; y)$
3. Stop if  $\|\theta_{(t+1)} - \theta_{(t)}\| < \tau$ ; return  $\hat{\theta}_{ML} = \theta_{(t+1)}$

Proof:

$$\begin{aligned} 0 &= s(\hat{\theta}_{ML}; y) \stackrel{\text{Taylor}}{\approx} s(\theta; y) + \frac{\partial s(\theta; y)}{\partial \theta} (\hat{\theta}_{ML} - \theta) \Leftrightarrow \\ \hat{\theta}_{ML} &\approx \theta - \left( \frac{\partial s(\theta; y)}{\partial \theta} \right)^{-1} s(\theta; y) \approx \theta - I^{-1}(\theta) s(\theta; y) \end{aligned}$$

As  $\frac{\partial s(\theta; y)}{\partial \theta}$  is often complicated, its expectation  $I(\theta)$  is used.

The second part in 2 can be weighted with a step size  $\delta$  or  $\delta(t) \in (0, 1)$ , e. g. to ensure convergence.

If  $I(\theta)$  can't be analytically derived, simulation from  $f(y; \theta_{(t)})$  can be used. For the exponential family, step 2 then changes to  $\theta_{(t+1)} := \theta_{(t)} + \hat{\text{Var}}_{\theta_{(t)}}(t(Y))^{-1} E_{\theta_{(t)}}(t(Y))$  as the ML estimate is the expectation.

#### Log Likelihood Ratio

$$lr(\theta, \hat{\theta}) := l(\hat{\theta}) - l(\theta) = \log \frac{L(\hat{\theta})}{L(\theta)}$$

with  $2 \cdot lr(\theta, \hat{\theta}) \stackrel{a}{\sim} \chi^2_1$

Proof:

$$l(\theta) \stackrel{\text{Taylor Series}}{\approx} l(\hat{\theta}) + \underbrace{\frac{\partial l(\hat{\theta})}{\partial \theta}}_{=0} (\theta - \hat{\theta}) + \frac{1}{2} \underbrace{\frac{\partial^2 l(\hat{\theta})}{\partial \theta^2}}_{\approx I^{-1}(\theta) s(\theta; Y)} (\underbrace{\theta - \hat{\theta}}_{\approx -I(\theta)})^2$$

$$\approx l(\hat{\theta}) - \frac{1}{2} \frac{s^2(\theta, Y)}{I(\theta)}$$

$s(\theta, Y)$  is asymptotically normal.

If  $\theta \in \mathbb{R}^p$  the corresponding distribution is  $\chi_p^2$ .

## 3.4 Sufficiency und Consistency

**Statistic**

$$t: \mathbb{R}^n \rightarrow \mathbb{R}$$

$t(Y_1, \dots, Y_n)$  depends on sample size  $n$  and is a random variable

**Suffizienz**

Eine Statistik  $t(Y_1, \dots, Y_n)$  ist suffizient für  $\theta$ , wenn die bedingte Verteilung  $f(y_1, \dots, y_n | t_0 = t(Y_1, \dots, Y_n); \theta)$  unabhängig von  $\theta$  ist.

**Neyman-Kriterium:**

$$t(Y_1, \dots, Y_n) \text{ suffizient} \Leftrightarrow f(y; \theta) = h(y)g(t(y); \theta)$$

Proof:

“ $\Rightarrow$ ”:

$$f(y; \theta) = \underbrace{f(y | t = t(y); \theta)}_{h(y)} \underbrace{f_t(t(y); \theta)}_{g(t(y); \theta)}$$

“ $\Leftarrow$ ”:

$$f_t(t; \theta) = \int_{t=t(y)} f(y; \theta) dy = \int_{t=t(y)} h(y) g(t; \theta) dy$$

Damit:

$$f(y | t = t(y); \theta) = \frac{f(y, t = t(y); \theta)}{f_t(t, \theta)} = \begin{cases} \frac{h(y)g(t; \theta)}{g(t; \theta)} & t = t(y) \\ 0 & \text{sonst} \end{cases}$$

**Minimalsuffizienz:**

$t(\cdot)$  ist suffizient und  $\forall \tilde{t}(\cdot) \exists h(\cdot)$  s.t.  $t(y) = h(\tilde{t}(y))$

**(schwache) Konsistenz**

$$MSE(\hat{\theta}, \theta) \xrightarrow{n \rightarrow \infty} 0 \Rightarrow \hat{\theta} \text{ konsistent}$$

Proof:

$P(|\hat{\theta} - E_{\hat{\theta}}| \geq \delta) \leq \frac{\text{Var}_{\theta}(\hat{\theta})}{\delta^2}$  using the inequality of Chebyshev and  $MSE(t(\cdot), \theta) = \text{Var}_{\theta}(t(Y_1, \dots, Y_n)) + \text{Bias}^2(t(\cdot); \theta)$

## 4 Statistical Hypothesis Testing

### 4.1 Significance and Confidence Intervals

**Significance Test**

Assuming two states  $H_0$  and  $H_1$  and two corresponding decisions “ $H_0$ ” and “ $H_1$ ”, a decision rule (a threshold  $c \in \mathbb{R}$  for the test statistic  $T(X)$ ) is constructed s.t.:

$$P(\text{“}H_1\text{”} | H_0) \leq \alpha$$

	“ $H_0$ ”	“ $H_1$ ”
$H_0$	$1 - \alpha$ (correct)	$\alpha$ (type I error)
$H_1$	$\beta$ (type II error)	$1 - \beta$ (correct)

**Power** concerns the type II error

$$\text{power} = P(\text{“}H_1\text{”} | H_1) = 1 - \beta$$

**p-Value** measures the amount of evidence against  $H_0$

$$p\text{-value} \leq \alpha \Leftrightarrow \text{“}H_0\text{”}$$

**Confidence Interval**

$$[t_l(Y), t_r(Y)] \text{ Confidence Interval}$$

$$\Leftrightarrow$$

$$P_{\theta}((t_l(Y) \leq \theta \leq t_r(Y))) \geq 1 - \alpha$$

with  $1 - \alpha$  confidence level und  $\alpha$  significance level

**Corresponding Test**

$$\theta \notin [t_l(y), t_r(y)] \Leftrightarrow \text{“}H_1\text{”}$$

**Specificity** or True Negative Rate (1—empirical type I error)

$$TNR = \frac{\#TN}{\#N} = \frac{\#TN}{\#TN + \#FP}$$

**Sensitivity** or True Positive Rate, Recall (empirical power)

$$TPR = \frac{\#TP}{\#P} = \frac{\#TP}{\#TP + \#FN}$$

## 4.2 Tests for One Sample

**Normal Distribution**  $X_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$

**Test for  $\mu$ , known  $\sigma^2$  (Simple Gauss-Test)**

$H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$

$$T(X) = \frac{\bar{X} - \mu_0}{\sigma} \stackrel{H_0}{\sim} N(0, 1)$$

**Test for  $\mu$ , unknown  $\sigma^2$  (Simple t-Test)**

$H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$

$$T(X) = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}} \stackrel{H_0}{\sim} t_{n-1}$$

with  $\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}$

**ML Estimate**  $\hat{\theta} \stackrel{a}{\sim} N(\theta, I^{-1}(\theta))$

**Wald Test**

$H_0: \theta = \theta_0$  vs.  $H_1: \theta \neq \theta_0$

$$T(X) = |\hat{\theta} - \theta_0| \stackrel{H_0}{\sim} N(0, I^{-1}(\theta_0))$$

As  $\hat{\theta}$  converges to  $\theta_0$  under  $H_0$ , it can also be used to calculate the variance:  $I^{-1}(\hat{\theta})$ .

**Score Test**

$H_0: \theta = \theta_0$  vs.  $H_1: \theta \neq \theta_0$

$$T(X) = |s(\theta_0; y)| \stackrel{H_0}{\sim} N(0, I(\theta_0))$$

Advantage compared to the Wald Test:  $\hat{\theta}$  does not have to be calculated.

**Likelihood Ratio Test**

$H_0: \theta = \theta_0$  vs.  $H_1: \theta \neq \theta_0$

$$T(X) = 2(l(\hat{\theta}) - l(\theta_0)) \stackrel{H_0}{\sim} \chi_1^2$$

**Neyman-Pearson Test**

$H_0: \theta = \theta_0$  vs.  $H_1: \theta = \theta_1$

$$T(X) = l(\theta_0) - l(\theta_1)$$

For a given significance level  $\alpha$ , the Neyman Pearson Test is the most powerful test for comparing two estimates for  $\theta$ .

Proof:

Decision rule of the NP-Test:  $\varphi^* = \begin{cases} 1 & \text{if } \frac{f(y; \theta_0)}{f(y; \theta_1)} \leq e^c \\ 0 & \text{otherwise} \end{cases}$

Need to show:  $P(\varphi(Y)=1|\theta_1) \leq P(\varphi^*(Y)=1|\theta_1) \forall \varphi$

$$P(\varphi^*=1|\theta_1) - P(\varphi=1|\theta_1) =$$

$$= \int \{\varphi^*(y) - \varphi(y)\} f(y; \theta_1) dy$$

$$\geq \frac{1}{e^c} \int_{\varphi^*=1} \{\varphi^*(y) - \varphi(y)\} f(y; \theta_0) dy \quad f(y; \theta_1) \geq \frac{f(y; \theta_0)}{e^c}$$

$$+ \frac{1}{e^c} \int_{\varphi^*=0} \{\varphi^*(y) - \varphi(y)\} f(y; \theta_0) dy \quad f(y; \theta_1) \leq \frac{f(y; \theta_0)}{e^c}$$

$$= \frac{1}{e^c} \int \{\varphi^*(y) - \varphi(y)\} f(y; \theta_0) dy = 0$$

$$\text{As } \alpha = \int \varphi^*(y) f(y; \theta_0) dy = \int \varphi(y) f(y; \theta_0) dy$$

## 4.3 Tests for Two Samples

## 4.4 Tests for Goodness of Fit

**Discrete (Chi-Squared)**

$H_0: X_i \sim F_0$  vs.  $H_1: X_i \sim F \neq F_0$

$$T(X) = \sum_{k=1}^K \frac{(n_k - l_k)^2}{l_k} \stackrel{H_0}{\sim} \chi_{K-1-p}^2$$

with the following contingency table:

	1	2	...	K
observed	$n_1$	$n_2$	...	$n_K$
expected under $H_0$	$l_1$	$l_2$	...	$l_K$

$l_k > 5$  and  $l_k > n - 5$  for the  $\chi_{K-1-p}^2$ -distribution to hold,  $F_0$  needs to be known, but its  $p$  parameters can be estimated. The test can be applied to discretized continuous variables.

**Continuous (Kolmogorov-Smirnov Test)**

$H_0: X_i \sim F_0$  vs.  $H_1: X_i \sim F \neq F_0$

$$T(X) = \sup_x |F_n(x) - F(x; \theta)| \stackrel{H_0}{\sim} KS$$

with the distribution function  $F(x; \theta)$  and the empirical counterpart  $F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}$

Proof:

$$P(\sup_x |F_n(x) - F(x; \theta)| \leq t) =$$

$$= P(\sup_y |F^{-1}(y; \theta) - x| \leq t) \quad \begin{matrix} x \in [0, 1], x = F^{-1}(y; \theta) \\ F(F^{-1}(y; \theta); \theta) = y \end{matrix}$$

$$= P(\sup_y |\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i \leq y\}} - y| \leq t) \quad \text{with } U_i \sim U(0, 1)$$

$$* F_n(F^{-1}(y; \theta)) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq F^{-1}(y; \theta)\}} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{F(Y_i; \theta) \leq y\}}$$

For an estimated parameter the distribution of  $T(X)$  is not independent of  $F_0$ :  $T(X) \stackrel{H_0}{\sim} KS$  only holds asymptotically.

**Pivotal Statistic**

$g(Y; \theta)$  pivotal

$\Leftrightarrow$

Distribution of  $g(Y; \theta)$  independent of  $\theta$

**Approximative Pivotal Statistic**

$H_0: X_i \sim F$  pivotal vs.  $H_1: X_i \sim F$  not pivotal

$$g(\hat{\theta}; \theta) = \frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \stackrel{\circ}{\sim} N(0, 1)$$

with  $\hat{\theta} = t(Y) \stackrel{\circ}{\sim} N(\theta, \text{Var}(\hat{\theta}))$

## 4.5 Multiple Tests

**Family-Wise Error Rate (FWER)** as  $p\text{-value} \sim U(0, 1)$

For  $m$  tests:

$$\alpha \leq P(\cup_{k=1}^m (p_k \leq \alpha) | H_{0k}, k = 1, \dots, m) \leq m\alpha$$

$$FWER := P(\exists k : "H_1 k" | \forall k : H_{0k})$$

**Bonferroni Adjustment**

$$\alpha_B = \frac{\alpha}{m}$$

**Šidák Adjustment** only for independent tests

$$\alpha_S = 1 - (1 - \alpha)^{1/m}$$

## 5 Regression

### 5.1 Models

#### 5.1.1 Simple Linear Model

**Theoretical Model**

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

**Empirical Model**

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

**Assumptions**

- **Independent Observations**  $y_1, \dots, y_n$  are independent
- **Linearity of the Mean**  $E(Y|x) = \beta_0 + \beta_1 x$  or  $E(e|x) = 0$
- **Constant Variation**  $\text{Var}(Y|x) = \sigma^2$

For the normal linear model:

- **Normality**  $e|x \sim N(0, \sigma^2)$  ;  $Y|x \sim N(\hat{y}, \sigma^2)$

**Attributes of the Regression Line**

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\ \hat{e}_i &= y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= y_i - (\bar{y} + \hat{\beta}_1 (x_i - \bar{x})) \end{aligned}$$

$$KI = \left[ \hat{\theta} - z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{\theta})}, \hat{\theta} + z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{\theta})} \right]$$

Proof:

$$1 - \alpha \approx P\left(z_{\frac{\alpha}{2}} \leq \frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \leq z_{1-\frac{\alpha}{2}}\right)$$

Proof:

$$\begin{aligned} \alpha &\stackrel{!}{=} P(\cup_{k=1}^m (p_k \leq \alpha) | H_{0k}, k = 1, \dots, m) \\ &= 1 - (1 - \alpha)^{1/m} \end{aligned}$$

**Holm's Procedure** also takes power into account

Order the p-values:  $p_{(1)} \leq \dots \leq p_{(m)}$

Step  $x \in \mathbb{N}^+$ : if  $p(x) > \frac{\alpha}{m+1-x}$  reject  $H_{01}$  to  $H_{0x}$  and stop, else move on to step  $x + 1$ .

**False Discovery Rate (FDR)** balances type I and II errors, especially for  $n \ll m$  problems

$$FDR = E\left(\frac{\# "H_1" | H_0}{\# "H_1" }\right)$$

Order the p-values:  $p_{(1)} \leq \dots \leq p_{(m)}$ , choose  $\alpha \in (0, 1)$

$j$  is largest index s. t.  $p(j) \leq \alpha j / m$ , reject all  $H_{0i}$  for  $i \leq j$

It can be shown that  $FDR \leq m_0 \alpha / m$ , with  $m_0 = \#H_0$

$$\begin{aligned} \sum_{i=1}^n \hat{e}_i &= \sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y} - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) \\ &= n\bar{y} - n\bar{y} - \hat{\beta}_1 (n\bar{x} - n\bar{x}) = 0 \\ \bar{\hat{y}} &= \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} (n\bar{y} + \hat{\beta}_1 (n\bar{x} - n\bar{x})) = \bar{y} \end{aligned}$$

**Estimates (OLS)**

$$\hat{\beta}_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \cdot \sqrt{\frac{S_{yy}}{S_{xx}}} = r \sqrt{\frac{S_{yy}}{S_{xx}}}$$

Proof:

$$\begin{aligned} \text{Cov}(x, y) &= \text{Cov}(x, \hat{\beta}_0 + \hat{\beta}_1 x) = \hat{\beta}_1 \text{Var}(x) \\ &\iff \hat{\beta}_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \end{aligned}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Proof:

$$E[y] = E[\hat{\beta}_0 + \hat{\beta}_1 x + \hat{e}] \iff \hat{\beta}_0 = E[y] - \hat{\beta}_1 E[x]$$

The estimates are the same as for the ML procedure.

**Estimates (ML)**  $Y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n x_i \hat{\beta}_1$$

$$\hat{\beta}_1 = \sum_{i=1}^n x_i (y_i - \hat{\beta}_0) / \sum_{i=1}^n x_i^2$$

## 5.1.2 Multivariate Linear Model

**Theoretical Model**

$$Y = X\beta + u$$

**Empirical Model**

$$Y = X\hat{\beta} + e$$

$$\hat{Y} = X\hat{\beta}$$

$$y = (y_1, \dots, y_n)^T, e = (e_1, \dots, e_n)^T, X = \begin{pmatrix} 1 & x_1 & \dots & x_p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_p \end{pmatrix}$$

**Assumptions**

- **Independent Observations**  $y_1, \dots, y_n$  are independent
- **Linearity of the Mean**  $E(Y|x_{1:p}) = X\beta$  or  $E(e|x_{1:p}) = 0$
- **Constant Variation**  $Var(Y|x) = \sigma^2$

For the normal linear model:

- **Normality**  $e_i|x_{1:p} \sim N(0, \sigma^2)$ ;  $Y|x \sim N(\hat{y}, \sigma^2)$

**Estimates (ML)**  $Y|x_{1:p} \sim N(X\beta, \sigma^2)$

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

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = I^{-1}(\beta)$$

Proof:

$$l(\beta, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - x_i \hat{\beta}_1)^2$$

The  $\beta$ -estimates are the same as for the OLS procedure.

Proof:

$$l(\beta_0, \beta_1, \sigma^2) = \sum_{i=1}^n \left\{ -\frac{1}{2} \log \sigma^2 - \frac{1}{2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2} \right\}$$

The estimates are the same as for the OLS procedure.

$\beta$  is the **Best Linear Unbiased Estimator**

Proof:

$$\text{Unbiased because of the Gauß-Markov Theorem: } E(\hat{\beta}) = (X^T X)^{-1} X^T E(Y|X) = (X^T X)^{-1} X^T X \beta = \beta$$

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta}); \quad \hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

The ML-estimate for  $\sigma^2$  is biased.

Proof:

$$H := X(X^T X)^{-1} X^T \text{ hat matrix; } HH = H = H^T \text{ (idempotent)}$$

$$\begin{aligned} E((Y - X\hat{\beta})^T (Y - X\hat{\beta})) &= E((Y^T (I_n - H))^T ((I_n - H)Y)) \\ &= E(\text{tr}(Y^T (I_n - H)Y)) \\ &= E(\text{tr}((I_n - H)YY^T)) \\ &= \text{tr}((I_n - H)E(YY^T)) \\ &= \text{tr}((I_n - H)E(X\beta\beta^T X^T + \sigma I_n)) \\ &= \sigma^2 \text{tr}((I_n - H)) \\ &= \sigma^2 (n - p) \end{aligned}$$

$$s^2 = \frac{1}{n - p} (y - X\hat{\beta})^T (y - X\hat{\beta}); \quad \hat{\beta} \sim t_{n-p}(\beta, s^2 (X^T X)^{-1})$$

with  $s$  an unbiased estimator

## 5.1.3 Bayesian Linear Model

**Prior** flat prior

$$f_{\beta, \sigma^2}(\beta, \sigma^2) = \frac{1}{\sigma^2}$$

**Posterior**

Resulting posterior:

$$f_{post}(\beta, \sigma^2|y) \propto (\sigma^2)^{-\frac{n}{2}+1} e^{-\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta)}$$

Note:  $f_{post}(\beta, \sigma^2|y) = f(\beta|\sigma^2, y)f(\sigma^2|y)$

$$\begin{aligned} \beta|\sigma^2, y &\sim N(\hat{\beta}, \sigma^2 (X^T X)^{-1}) \\ \sigma^2|y &\sim \text{IG}\left(\frac{n-p}{2}, \frac{s^2(n-p)}{2}\right) \\ \beta|y &\sim t_{n-p}(\hat{\beta}, s^2 (X^T X)^{-1}) \end{aligned}$$

The two distributions for  $\beta$  mirror the results for  $\hat{\beta}$  in the linear model.

### 5.1.4 Quantile Regression

**Prediction Interval** range of  $1 - \alpha$  fraction of the data

$$Var(\hat{Y}|x_{1:p}) = Var(X\hat{\beta}) + \sigma^2$$

Determined by estimation variance (usually depicted in confidence intervals) plus residual variance.

## 5.2 Analysis of Variances (ANOVA)

$$SS_{Total} = SS_{Explained} + SS_{Residual}$$

with

$$SS_{Total} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SS_{Explained} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$SS_{Residual} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = S_{yy} - \hat{\beta}^2 S_{xx}$$

## 5.3 Goodness of Fit

### 5.3.1 Coefficient of Determination

$$R^2 = \frac{SS_{Explained}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}} = r^2$$

Range:  $0 \leq R^2 \leq 1$

## 6 Classification

### 6.1 Diskriminant Analysis (Bayes)

## 7 Cluster Analysis

## 8 Bayesian Statistics

### 8.1 Basics

**Bayes Theorem**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{für } P(A), P(B) > 0$$

or more general:

$$\begin{aligned} f_{post}(\theta|X) &= \frac{f(X|\theta) \cdot f_{\theta}(\theta)}{\int f(X|\tilde{\theta}) f_{\theta}(\tilde{\theta}) d\tilde{\theta}} \\ &= C \cdot f(X|\theta) \cdot f_{\theta}(\theta) \quad \text{choose } C \text{ so that } \int f(\theta|X) = 1 \\ &\propto f(X|\theta) \cdot f_{\theta}(\theta) \end{aligned}$$

**Point Estimates**

$$\hat{\theta}_{postmean} = E_0(\vartheta|x) = \int_{\vartheta \in \Theta} \vartheta f_{\theta}(\vartheta|x) d\vartheta$$

$$\hat{\theta}_{postmode} = \underset{\vartheta}{\operatorname{argmax}} f_{\theta}(\vartheta, x)$$

$$\hat{\theta}_{Bayesrisk} = \underset{t(\cdot)}{\operatorname{argmin}} R_{Bayes}(t(\cdot))$$

with Bayes risk:  $R_{Bayes}(t(\cdot)) = \int_{\Theta} R(t(\cdot), \vartheta) f_{\theta}(\vartheta) d\vartheta$

$$\hat{\theta}_{postBayesrisk} = \underset{t(\cdot)}{\operatorname{argmin}} R_{postBayes}(t(\cdot)|y)$$

with posterior Bayes risk:

$$R_{postBayes}(t(\cdot)|y) = \int L(t(y), \vartheta) f_{\theta}(\vartheta|y) = E_{\theta|y}(L(t(y), \theta)|y)$$

**Credibility Interval**

$$P_{\theta}(\theta \in [t_l(y), t_r(y)] | y) = \int_{t_l(y)}^{t_r(y)} f_{\theta}(\vartheta|y) d\vartheta = 1 - \alpha$$

- symmetric:  $\int_{-\infty}^{t_l(y)} f_{\theta}(\vartheta|y) d\vartheta = \int_{t_r(y)}^{\infty} f_{\theta}(\vartheta|y) d\vartheta = \frac{\alpha}{2}$
- highest density:  $HDI = \theta | f_{\theta}(\theta|y) \geq c$ , choose  $c$  s. t.  $\int_{\vartheta \in HDI(y)} f_{\theta}(\vartheta|y) d\vartheta = 1 - \alpha$

**Bayes Factor** evidence contained in data for  $M_1$  vs.  $M_2$

$$\frac{P(M_1|y)}{P(M_0|y)} = \underbrace{\frac{f(y|M_1)}{f(y|M_0)}}_{\text{Bayes Factor}} \frac{P(M_1)}{P(M_0)}$$

with marginal likelihood  $f(y|M_i) = \int f(y|\vartheta) f_{\theta}(\vartheta|M_i) d\vartheta$

### Priors

**Flat (uninformative) Prior**

$f_{\theta}(\theta) = \text{const.}$  for  $\theta > 0$ , therefore:  $f(\theta|X) = C \cdot f(X|\theta)$

As  $\int f_{\theta}(\theta) = 1$  not possible like this, this is not a real density.

Changes for transformations of the parameter.

Proof: For  $\gamma = g(\theta)$ :  $f_{\gamma}(\gamma) = f_{\theta}(g^{-1}(\gamma)) \left| \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \right|$

No prior is truly uninformative.

### Jeffrey's Prior

Remains unchanged for transformations of the parameter.

For Fisher-regular distributions:  $f(\theta) \propto \sqrt{I_\theta(\theta)}$

Proof:

For  $\gamma = g(\theta)$  and  $f_\theta(\theta) = \sqrt{I_\theta(\theta)}$ :

$$f_\gamma(\gamma) \propto f_\theta(g^{-1}(\gamma)) \left| \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \right| \propto \sqrt{\frac{\partial g^{-1}(\gamma)}{\partial \gamma} I_\theta(g^{-1}(\gamma)) \frac{\partial g^{-1}(\gamma)}{\partial \gamma}} = \sqrt{I_\gamma(\gamma)}$$

Maximizes the information gained from the data (under appropriate regulatory conditions), i.e. maximizes

$E(KL(f_\theta(\cdot), f_{post}(\cdot, x)))$

### Empirical Bayes

Let the prior depend on a hyper-parameter:  $f_\theta(\theta, \gamma)$

Choose  $\gamma$  s.t.  $L(\gamma) = f(x; \gamma) = \int f(x; \vartheta) f_\theta(\vartheta, \gamma) d\vartheta$  is maximal.

Using the data to find the prior contradicts the Bayes approach of incorporating prior knowledge.

### Hierarchical Prior

$$x|\theta \sim f(x; \theta); \quad \theta|\gamma \sim f_\theta(\theta, \gamma); \quad \gamma \sim f_\gamma(\gamma)$$

### Conjugate Priors

If Prior and Posterior belong to the same family of distributions for a given likelihood function, they are called conjugate.

Examples:

Prior	Likelihood	Posterior
$\pi \sim \text{Be}(\alpha, \beta)$	$\text{Bin}(n, \pi)$	$\text{Be}(\alpha+k, \beta+n-k)$
$\mu \sim \text{N}(\gamma, \tau^2)$	$\text{N}(\mu, \sigma^2)$	$\text{N}(\cdot, \cdot) \xrightarrow{n \rightarrow \infty} \text{N}(\bar{y}, \frac{\sigma^2}{n})$
$\sigma^2 \sim \text{IG}(\alpha, \beta)$	$\text{N}(\mu, \sigma^2)$	$\text{IG}(\alpha+\frac{n}{2}, \beta+\frac{1}{2} \sum_{i=1}^n (y_i - \mu)^2)$
$\lambda \sim \text{Ga}(\alpha, \beta)$	$\text{Po}(\lambda)$	$\text{Ga}(\alpha+n\bar{y}, \beta+n)$

## 8.2 Numerical Methods for the Posterior

**Numerical Integration** here: trapezoid approximation

$$\int_{\Theta} f(y|\vartheta) f_\theta(\vartheta) d\vartheta \approx \sum_{k=1}^K \frac{f(y; \theta_k) f_\theta(\theta_k) + f(y; \theta_{k-1}) f_\theta(\theta_{k-1})}{2} (\theta_k - \theta_{k-1})$$

only normalisation constant unknown, works well for one-dimensional integrals

### Laplace Approximation

$$\int_{\Theta} f(y|\vartheta) f_\theta(\vartheta) d\vartheta \approx f(y; \hat{\theta}_P) f_\theta(\hat{\theta}_P) (2\pi)^{p/2} |J_P(\hat{\theta}_P)|^{\frac{1}{2}}$$

with the one-dimensional  $J_P := -\frac{\partial^2 l_{(n)}(\theta, y)}{\partial \theta^2} - \frac{\partial^2 \log f_\theta(\theta)}{\partial \theta^2}$  Fisher information considering the prior,  $\hat{\theta}_P$  posterior mode estimate s.t.  $s_{P, \theta}(\hat{\theta}_P) = 0$

Proof:

For  $n$  independent samples:

$$f_{post}(\theta|y) = \frac{\prod_{i=1}^n f(y_i|\theta) f_\theta(\theta)}{\int \prod_{i=1}^n f(y_i|\theta) f_\theta(\theta) d\theta}$$

Denominator:  $\int e^{\{\sum_{i=1}^n \log f(y_i|\theta) + \log f_\theta(\theta)\}} d\theta =$

$$\int e^{\{l(\theta; y) + \log f_\theta(\theta)\}} d\theta \approx \int e^{\{l_P(\hat{\theta}_P) - \frac{1}{2} J_P(\hat{\theta}_P)(\vartheta - \hat{\theta}_P)^2\}} d\vartheta$$

Resembles the normal distribution, therefore the inverse of the normalisation constant can be calculated, which gives the inverse of the Laplace approximation in the univariate case.

Works well for large  $n$  and is numerically simple also for big  $p$ .

### Monte Carlo Approximations

The denominator can be written as  $E_\theta(f(y; \theta)) =$

$\int_{\Theta} f(y|\vartheta) f_\theta(\vartheta) d\vartheta$ , which can be estimated by the arithmetic mean for a sample of  $\theta_1, \dots, \theta_N$ , which needs to be drawn from the prior. The following methods to draw from non-standard distributions can be used for that.

#### • Inverse CDF

$F(X)$  known. Since  $F(x) = u$ ,  $F^{-1}(u) = x$ ,  $u \sim U(0, 1)$

1. Draw  $u \sim U(0, 1)$
2. Compute  $F^{-1}(u)$  to get a value  $x$

Proof:

$$P(x \leq y) = P(F^{-1}(u) \leq y) = P(u \leq F(y)) = F(y)$$

#### • Rejection Sampling

An umbrella distribution  $g(x)$  can be found s.t.

$$\frac{f(x)}{g(x)} \leq M \quad \forall x \text{ with } f(x) > 0 \text{ when } g(x) > 0$$

1. Draw candidate  $y \sim g(x)$
2. Acceptance probability  $\alpha$  for  $y$ :  $\alpha = \frac{f(y)}{Mg(y)}$
3. Draw  $u \sim U(0, 1)$  and accept if  $u \leq \alpha$ , else: step 1

Proof:

$$\begin{aligned} P\left(Y \leq x | U \leq \frac{f(Y)}{Mg(Y)}\right) &= \frac{P\left(Y \leq x, U \leq \frac{f(Y)}{Mg(Y)}\right)}{P\left(U \leq \frac{f(Y)}{Mg(Y)}\right)} \\ &= \frac{\int_{-\infty}^x \int_0^{\frac{f(y)}{g(y)}} du g(y) dy}{\int_{-\infty}^{\infty} \int_0^{\frac{f(y)}{g(y)}} du g(y) dy} = \frac{\int_{-\infty}^x \frac{f(y)}{g(y)} g(y) dy}{\int_{-\infty}^{\infty} \frac{f(y)}{g(y)} g(y) dy} \\ &= \frac{\int_{-\infty}^x f(y) dy}{\int_{-\infty}^{\infty} f(y) dy} = P(X \leq x) \end{aligned}$$

#### • Importance Sampling

Directly estimate  $E_\theta(f(y; \theta))$ .

For sampling distribution  $g(x)$

$$\frac{1}{N} \sum_{i=1}^N \frac{f(x)}{g(x)}$$

is a consistent estimator.

Proof:

$$E_g\left(\frac{1}{N} \sum_{i=1}^N \frac{f(x)}{g(x)}\right) = \int \frac{f(x)}{g(x)} g(x) dx = \int f(x) dx = f(x)$$

**Markov Chain Monte Carlo** sample from  $f_{post}(\theta|X)$

$f(y)$  unknown, however:

$$\frac{f_{post}(\theta|x)}{f_{post}(\tilde{\theta}|x)} = \frac{f(x|\theta)f_{\theta}(\theta)}{f(y)} \frac{f(y)}{f(x|\tilde{\theta})f_{\theta}(\tilde{\theta})} = \frac{f(x|\theta)f_{\theta}(\theta)}{f(x|\tilde{\theta})f_{\theta}(\tilde{\theta})}$$

**Metropolis-Hastings:** Draw Markov Chain  $\theta_1^*, \dots, \theta_n^*$ :

1. Draw candidate  $\theta^*$  from proposal distribution  $q(\theta|\theta_{(t)}^*)$
2. Accept  $\theta_{(t+1)}^* = \theta^*$  with probability

$$\alpha(\theta_{(t)}|\theta^*) = \min \left\{ 1, \frac{f_{post}(\theta^*|y) q(\theta_{(t)}^*|\theta^*)}{f_{post}(\theta_{(t)}^*|y) q(\theta^*|\theta_{(t)}^*)} \right\}$$

else choose  $\theta_{(t+1)}^* = \theta_{(t)}^*$

This sequence has a stationary distribution for  $n \rightarrow \infty$ .

Choice of  $q$ : trade-off between exploring  $\Theta$  and reaching a high  $\alpha$ .

Burn-in and thinning out give *i.i.d.* samples from  $f_{post}(\theta|X)$ .

**Gibbs Sampling:** For high dimensions  $\alpha$  is close to zero.

Sample from the marginal distributions separately:

$$\theta_i^* \sim f_{\theta_i|y, \theta \setminus \theta_i}(\theta_i^*|y, \theta_{t^*,i}^*)$$

with  $\theta_{t^*,i}$  most recent estimates without  $\theta_i$

A Gibbs sampled sequence converges to  $f_{post}(\theta|X)$  as stationary.

Can also be used on its own, if marginal densities are known.

### Variational Bayes Principles

Approximate  $f_{post}(\theta|X)$  by  $q_{\theta} = \min_{q_{\theta} \in Q} KL(f_{post}(\cdot|X), q_{\theta}(\cdot))$

Restrict  $q_{\theta}$  to independence:  $q_{\theta}(\theta) = \prod_{k=1}^p q_k(\theta_k)$

Update each component iteratively. Works well for big  $p$ .