# Statistics

## Collection of Formulas

# Contents

# 1 Descriptive 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**

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

**Arithmetic Mean**

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

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

*Calculation Rules:*

- ⋆ $E(a + b \cdot X) = a + b \cdot E(X)$
- ⋆ $E(X \pm Y) = E(X) \pm E(Y)$

**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**

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

### 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:*

- ⋆ $Var(aX + b) = a^2 \cdot Var(X)$
- ⋆ $Var(X \pm Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

**(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.

### 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

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

These are also the values for the Lorenz curve.

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

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

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

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 $(\sigma^2)^{\frac{2}{3}}$

**(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**

$$\gamma = k - 3$$

### 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: $-1 \leq \gamma \leq 1$

**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

$$\rho = \frac{s_{rg_x rg_y}}{\sqrt{s_{rg_x rg_x} s_{rg_y rg_y}}}$$

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)**

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

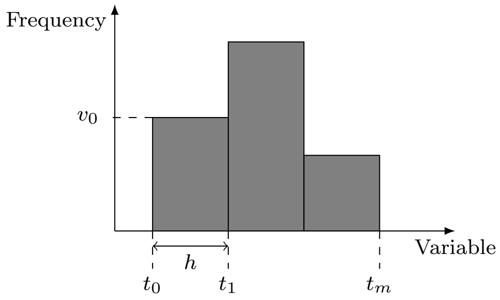
with
$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2$ or $s_{xy} = \frac{S_{xy}}{n}$
$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ or $s_{xx} = \frac{S_{xx}}{n}$
$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ or $s_{yy} = \frac{S_{yy}}{n}$

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

## 1.2 Tables

## 1.3 Diagrams

### 1.3.1 Histogram




sample: $X = \{x_1, x_2, ...; x_n\}$
$k$-th bin: $B_k = [t_k, t_{k+1})\,, k = \{0, 1, ..., m-1\}$
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)

# 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:
$n! = n \cdot (n-1) \cdot ... \cdot 1$
$\binom{n}{m} = \frac{n!}{m!(n-m)!}$

## 2.2 Probability Theory

**Laplace**

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

**Kolmogorov Axioms** mathematical definition of probability

(1) $0 \le P(A) \le 1 \qquad \forall A \in \mathcal{A} = \sigma\text{-algebra}(\Omega)$
(2) $P(\Omega) = 1$
(3) $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
$\forall A_i \in \mathcal{A}, i = 1, ..., \infty$ with $A_i \cap A_j = \emptyset$ for $i \neq j$

Implications:

- $P(\bar{A}) = 1 - P(A)$
- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $A \subseteq B \Rightarrow P(A) \le P(B)$

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

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

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

**Conditional Probability**

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

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

**Law of Total Probability**

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \qquad \text{for } \Omega = \bigcup_{i=1}^{\infty} A_i \text{ and } A_i \cap A_j = \emptyset$$

**Bayes' Theorem**

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

**Stochastic Independence**

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

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

## 2.3 Random Variables/Vectors

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

**Definition**

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

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 $f(\cdot)$** (positive, integrates out to 1):

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

  If $Y = g(X)$, then $f_Y(y) = \left|\frac{dg^{-1}(y)}{dy}\right| f_X(g^{-1}(y))$.

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

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

  with $\lim_{y\to-\infty} F_Y(y) = 0$ and $\lim_{y\to\infty} F_Y(y) = 1$

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

- **$k$th Moment**: $E(Y^k) = \int y^k f_Y(y)dy$,
  **$k$th centralized Moment**: $E(\{Y - E(Y)\}^k)$

**Moment Generating Function**

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

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

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* $\in \mathbb{R}^q$

**Definition**

$$(Y_1, Y_2, ..., Y_q)$$

with random variables $Y_i$

**Density and Cumulative Distribution Function**

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

$$P(a_1{\leq}Y_1{\leq}b_1, ..., a_q{\leq}Y_q{\leq}b_q) = \int_{a_1}^{b_1} ... \int_{a_q}^{b_q} f(y_1, .., y_q) dy_1 ... dy_q$$

**Marginal Density**

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

**Conditional Density** two-dimensional case

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

**Covariance and Correlation**

$$Cov(Y_j, Y_k) = E(Y_j Y_k) - E(Y_j)E(Y_k)$$

$$Cor(Y_j, Y_k) = \frac{Cov(Y_j, Y_k)}{\sqrt{Var(Y_j)Var(Y_k)}}$$

**Iterated Expectation**

$$\mathrm{E}(Y) = \mathrm{E}_X(\mathrm{E}(Y|X))$$

Proof:

$$\mathrm{E}(Y) = \int y f(y) dy = \int\int y f(y|x) dy f_X(x) dx = \mathrm{E}_X(\mathrm{E}(Y|X))$$

$$\mathrm{Var}(Y) = \mathrm{E}_X(\mathrm{Var}(Y|X)) + \mathrm{Var}_X(\mathrm{E}(Y|X))$$

Proof:

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

# 2.4 Probability Distributions

## 2.4.1 Discrete Distributions

**Discrete Uniform**

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

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

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

**Binomial** Successes in independent trials

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

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

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

**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

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

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

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

The model tends to overestimate the variance (Overdispersion).

*Approximation* of the Binomial for small p

**Geometric**

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

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

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

**Negative Binomial**

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

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

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

### 2.4.2 Continuous Distributions

**Continuous Uniform**

$$Y \sim \mathrm{U}(a, b) \text{ with } \alpha, \beta \in \mathbb{R}, a \leq b,\ y \in [a, b]$$
$$p(y|a, b) = \frac{1}{b-a}$$
$$\mathrm{E}(Y|a, b) = \frac{a+b}{2},\ \mathrm{Var}(Y|a, b) = \frac{(b-a)^2}{12}$$

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

$$Y \sim \mathrm{N}(\mu, \sigma^2) \text{ with } \mu \in \mathbb{R}, \sigma^2 > 0,\ y \in \mathbb{R}$$
$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$
$$\mathrm{E}(Y|\mu, \sigma^2) = \mu,\ \mathrm{Var}(Y|\mu, \sigma^2) = \sigma^2$$

**Log-Normal**

$$Y \sim \mathrm{LogN}(\mu, \sigma^2) \text{ eith } \mu \in \mathbb{R}, \sigma^2 > 0,\ y > 0$$
$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 y}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right)$$
$$\mathrm{E}(Y|\mu, \sigma^2) = \exp(\mu + \frac{\sigma^2}{2}),$$
$$\mathrm{Var}(Y|\mu, \sigma^2) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$$

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

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

$$Y \sim \mathrm{t}_\nu(\mu, \sigma^2) \text{ with } \mu \in \mathbb{R}, \sigma^2, \nu > 0,\ y \in \mathbb{R}$$
$$p(y|\mu, \sigma^2, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\frac{\nu}{2})\Gamma(\sqrt{\nu\pi\sigma})} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$
$$\mathrm{E}(Y|\mu, \sigma^2, \nu) = \mu \text{ for } \nu > 1,$$
$$\mathrm{Var}(Y|\mu, \sigma^2, \nu) = \sigma^2 \frac{\nu}{\nu-2} \text{ for } \nu > 2$$

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

**Beta**

$$Y \sim \mathrm{Be}(a, b) \text{ with } a, b > 0,\ y \in [0, 1]$$
$$p(y|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}$$
$$\mathrm{E}(Y|a, b) = \frac{a}{a+b},$$
$$\mathrm{Var}(Y|a, b) = \frac{ab}{(a+b)^2(a+b+1)},$$
$$\mathrm{mod}(Y|a, b) = \frac{a-1}{a+b-2} \text{ for } a, b > 1$$

**Gamma**

$$Y \sim \mathrm{Ga}(a, b) \text{ with } a, b > 0,\ y > 0$$
$$p(y|a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by)$$
$$\mathrm{E}(Y|a, b) = \frac{a}{b},$$
$$\mathrm{Var}(Y|a, b) = \frac{a}{b^a},$$
$$\mathrm{mod}(Y|a, b) = \frac{a-1}{b} \text{ for } a \geq 1$$

**Inverse-Gamma**

$$Y \sim \mathrm{IG}(a, b) \text{ with } a, b > 0,\ y > 0$$
$$p(y|a, b) = \frac{b^a}{\Gamma(a)} y^{-a-1} \exp(-\frac{b}{y})$$
$$\mathrm{E}(Y|a, b) = \frac{b}{a-1} \text{ for } a > 1,$$
$$\mathrm{Var}(Y|a, b) = \frac{b^2}{(a-1)^2(a-2)} \text{ for } a \geq 2,$$
$$\mathrm{mod}(Y|a, b) = \frac{b}{a+1}$$

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

**Exponential** Time between Poisson events

$$Y \sim \mathrm{Exp}(\lambda) \text{ with } \lambda > 0,\ y \geq 0$$
$$p(y|\lambda) = \lambda \exp(-\lambda y)$$
$$\mathrm{E}(Y|\lambda) = \frac{1}{\lambda},\ \mathrm{Var}(Y|\lambda) = \frac{1}{\lambda^2}$$

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

$$Y \sim \chi^2(\nu) \text{ with } \nu > 0, ,\ y \in \mathbb{R}$$
$$p(y|\nu) = \frac{y^{\frac{\nu}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)}$$
$$\mathrm{E}(Y|\nu) = \nu,\ \mathrm{Var}(Y|\nu) = 2\nu$$

### 2.4.3 Exponential Family

**Definition**

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

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

with $h(y) \geq 0$, $t(y)$ vector of the canonical statistic, parameter vector $\theta$ and $\kappa(\theta)$ as 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 e. g. $\frac{\partial\kappa(\theta)}{\partial\theta_1} = \mathrm{E}(t_1(Y))$

**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**
- **Binomial**

## 2.5 Multivariate Distributions

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

$$Y \sim \mathrm{N}(\mu, \Sigma) \text{ with } \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d} s.p.d.,\ y \in \mathbb{R}^d$$
$$p(y|\mu,\Sigma) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right)$$
$$\mathrm{E}(Y|\mu,\Sigma) = \mu,\ \mathrm{Var}(Y|\mu,\Sigma) = \Sigma$$

**General Copulas**

$F(y_1, .., y_q) = C(F_1(y_1), ..., F_q(y_q))$ with $C : [0,1]^q \to [0,1]$ with $C$ monotonically increasing as a cdf on $[0,1]^q$

Modelled as follows:

1. marginal distributions $F_j(y_j) = C(F_j(y_j), 1..., 1)$
2. dependence structure $\hat{u}_i = (\hat{u}_{i1}, ..., \hat{u}_{iq}) \overset{iid}{\sim} C(.)$ with $\hat{u}_{ij} := \hat{F}_j(y_{ij})$.

   The copula density is $c(u_{1:q}) = \frac{\partial^q C(u_{1:q})}{\partial u_1 \ldots \partial u_q}$ and $f(y_{1:q}) = c(F_1(y_1), ..., F_q(y_q)) \prod_{j=1}^q f_j(y_j)$.

**Tail Dependence**

upper: $\lambda_u := \lim_{u\to 1} P(Y_1 \geq F_1^{-1}(u) | Y_2 \geq F_2^{-1}(u))$
$= \lim_{u\to 1} \frac{1-2u+C(u,u)}{1-u}$
lower: $\lambda_l := \lim_{u\to 0} P(Y_1 \leq F_1^{-1}(u) | Y_2 \leq F_2^{-1}(u))$
$= \lim_{u\to 0} \frac{C(u,u)}{u}$

**Gaussian Copula** coefficients for pairwise dependences

$$c(u_{1:q}) = \frac{1}{|R|^{1/2}} \exp\left(-\frac{1}{2}u^T R^{-1} u\right)$$

For $Y_{ij} \sim \mathrm{N}(\mu_j, \sigma_j)$: $f(y_{ij}; \mu_j, \sigma_j^2) = \frac{1}{\sigma_j}\phi(Z_{ij})$ with $Z_{ij}$ the standardized $Y_{ij}$. With $u_{ij} = \phi^{-1}(Z_{ij})$, $R$ can be estimated.

$\lambda_l = 0$, $\lambda_u = 0$

**Archimedean Copulas** few parameters even in high dimensions

$$\psi(;\theta) : [0,1] \to [0,\infty)$$

with the parametric generator function $\psi(u,\theta)$ continuous, strictly decreasing, convex, and $\psi(1,\theta) = 0\ \forall\theta$

$$C(u_{1:q};\theta) = \psi^{-1}(\psi(u_1;\theta) + ... + \psi(u_q;\theta);\theta)$$

- **Clayton** $\psi(t;\theta) = \frac{1}{\theta}(\theta^{-1} - 1)$: $\lambda_l = 2^{-1/\theta}$, $\lambda_u = 0$
- **Frank** $\psi(t;\theta) = -\log\frac{\exp(-\theta t)-1}{\exp(-\theta)-1}$: $\lambda_l = 0$, $\lambda_u = 0$
- **Gumbel** $\psi(t;\theta) = (-\log(t))^\theta$: $\lambda_l = 0$, $\lambda_u = 2 - 2^{1/\theta}$

**Pair Copulas** flexible pairwise dependences

$$f_{123} = c_{12}c_{23}c_{23|1}\prod_{j=1}^3 f_j$$

**Generalized Extreme Value Distribution (GEV)**
for block maxima $M_n := \max(Y_{1:n})$:
$F_{M_n}(y) = P(M_n \leq y) = P(Y_{1:n} \leq y) = (F_Y(y))^n$

$$\lim_{n\to\infty} f_{M_n}(y) = \begin{cases} 1, & \text{if } F(y) = 1 \\ 0, & \text{otherwise} \end{cases}$$

For $\{a_n\}_{n=1}^\infty$, $\{b_n\}_{n=1}^\infty$ fixed sequences, the standardized maximum $\frac{M_n - a_n}{b_n}$ converges to a GEV as $n \to \infty$.

$$G(x) = \begin{cases} \exp(-(1+\gamma z)^{-1/\gamma}, & \text{for } \gamma \neq 0 \\ \exp(-\exp(-z)), & \text{for } \gamma = 0 \end{cases}$$

with location $\mu$, scale $\sigma$, and shape $\gamma$ and $z = \frac{x-\mu}{\sigma}$

- **Gumbel** $\gamma = 0$
- **Weibull** $\gamma > 0$
- **Frechet-Pareto** $\gamma < 0$

## 2.6 Limit Theorems

**Law of Large Numbers**

**Central Limit Theorem**

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

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

Proof:
For normal random variables $Z \sim \mathrm{N}(\mu, \sigma^2)$: $K_Z(t) = \mu t + \frac{1}{2}\sigma^2 t^2$. The first two derivatives $\frac{\partial^k K_Z(t)}{\partial t^k}\Big|_{t=0}$ are $\mu$ and $\sigma$. All other moments are zero.
For $Z_n = (Y_1 + Y_2 + ... + Y_n)/\sqrt{n}$:

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

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

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

Using the Taylor Expansion, we can write$K_{Z_n}(t) = 0 + \sqrt{n}\mu t + \frac{1}{2}\sigma^2 t^2 + ...$, where the terms in ... are tending towards 0 as $n \to \infty$.
Therefore: $K_{Z_n}(t) \overset{n\to\infty}{\longrightarrow} K_Z(t)$ with $Z \sim \mathrm{N}(\sqrt{n}\mu, \sigma^2)$.

# 3 Inference

## 3.1 Method of Moments

The theoretical moments are estimated by their empirical counterparts:

$$\mathrm{E}_{\hat{\theta}_{MM}}(Y^k) = m_k(y_1, ..., y_n)$$

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

## 3.2 Loss Functions

**Loss**

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

with parameter space $\Theta \subset \mathbb{R}$, $t \in \mathcal{T}$ with $t : \mathbb{R}^n \to \mathbb{R}$ 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 quantity. It is also the realisation of a random variable as it depends on a sample.

**Risk**

$$\begin{aligned} R(t(.), \theta) &= \mathrm{E}_\theta\left(\mathcal{L}(t(Y_1, ..., Y_n), \theta)\right) \\ &= \int_{-\infty}^{\infty} \mathcal{L}(t(Y_1, ..., Y_n), \theta) \prod_{i=1}^{n} f(y_i; \theta) dy_i \end{aligned}$$

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

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

**Mean Squared Error (MSE)**

$$\begin{aligned} MSE(t(.), \theta) &= \mathrm{E}_\theta\left(\{t(Y) - \theta\}^2\right) \\ &= \mathrm{Var}_\theta\left(t(Y_1, ..., Y_n)\right) + Bias^2((t(.); \theta) \end{aligned}$$

with $Bias(t(.); \theta) = \mathrm{E}_\theta\left(t(Y_1, ..., Y_n)\right) - \theta$

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

$$\begin{aligned} R(t(.), \theta) &= \mathrm{E}_\theta(\{t(Y) - \theta\}^2) \\ &= \mathrm{E}_\theta(\{t(Y) - \mathrm{E}_\theta(t(Y)) + \mathrm{E}_\theta(t(Y)) - \theta\}^2) \\ &= \mathrm{E}_\theta(\{t(Y) - \mathrm{E}_\theta(t(Y))\}^2) + \mathrm{E}_\theta(\{\mathrm{E}_\theta(t(Y)) - \theta\}^2) \\ &\quad + 2\mathrm{E}_\theta(\{t(Y) - \mathrm{E}_\theta\left(t(Y)\right)\}\{\mathrm{E}_\theta(t(Y)) - \theta\}) \\ &= \mathrm{Var}_\theta(t(Y_1, ..., Y_n)) + Bias^2((t(.); \theta) + 0 \end{aligned}$$

**Cramér-Rao Bound**

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

Proof:
For unbiased estimates: $\theta = \mathrm{E}_\theta(\hat{\theta}) = \int t(y) f(y; \theta) dy$

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

**Kullback-Leibler Divergence** Comparing distributions

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

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 $x = 1$.

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

Proof:

$$\begin{aligned} & R_{KL}(\theta, t(.)) = \\ &= \int_{-\infty}^{\infty} \mathcal{L}_{KL}(t(Y_1, ..., Y_n), \theta) \prod_{i=1}^{n} f(y_i; \theta) dy_i \\ &= \int \int \log \frac{f(\tilde{y}; \theta)}{f(\tilde{y}; t)} f(\tilde{y}; \theta) d\tilde{y} \prod_{i=1}^{n} f(y_i; \theta) dy_i \\ &= \int \int \left(\log f(\tilde{y}; \theta) - \log f(\tilde{y}; t)\right) f(\tilde{y}; \theta) d\tilde{y} \prod_{i=1}^{n} f(y_i; \theta) dy_i \\ &\approx -\int \underbrace{\left(\int \frac{\partial \log f(\tilde{y}; \theta)}{\partial \theta} f(\tilde{y}; \theta) d\tilde{y}\right)}_{0} (t - \theta) \prod_{i=1}^{n} f(y_i; \theta) dy_i \\ &+ \frac{1}{2} \int \underbrace{\left(-\int \frac{\partial^2 \log f(\tilde{y}; \theta)}{\partial \theta^2} f(\tilde{y}; \theta) d\tilde{y}\right)}_{I(\theta)} (t - \theta)^2 \prod_{i=1}^{n} f(y_i; \theta) dy_i \end{aligned}$$

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)

**Prerequisites**

- $Y_i \sim f(y;\theta)\ i.i.d.$
- $\theta \in \mathbb{R}^p$
- $f(.;\theta)$ Fisher-regular:
  - $\{y : f(y;\theta > 0)\}$ independent of $\theta$
  - Parameter space $\Theta$ is open
  - $f(y;\theta)$ twice differentiable
  - $\int \frac{\partial}{\partial\theta} f(y;\theta)dy = \frac{\partial}{\partial\theta}\int f(y;\theta)dy$

**Central Functions**

- **Likelihood** $L(\theta; y_1, ..., y_n)$: $\prod_{i=1}^n f(y_i;\theta)$
- **log-Likelihood** $l(\theta; y_1, ..y_n)$: $\log L(\theta; y_1, ..., y_n) = \sum_{i=1}^n \log f(y_i;\theta)$
- **Score** $s(\theta; y_1, ..., y_n)$: $\frac{\partial l(\theta;y_1,..y_n)}{\partial\theta}$
- **Fisher-Information** $I(\theta)$: $-\mathrm{E}_\theta\left(\frac{\partial s(\theta;Y)}{\partial\theta}\right)$
- **observed Fisher-Information** $J(\theta)$: $-\mathrm{E}_\theta\left(\frac{\partial s(\theta;y)}{\partial\theta}\right)$

**Attributes of the Score-Function**

first Bartlett-Equation:

$$\mathrm{E}\left(s(\theta;Y)\right) = 0$$

Proof:

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

second Bartlett-Equation:

$$\mathrm{Var}_\theta\left(s(Y;\theta)\right) = \mathrm{E}_\theta\left(-\frac{\partial^2 logf(Y;\theta)}{\partial\theta^2}\right) = I(\theta)$$

Proof:

$$\begin{aligned}
0 = \frac{\partial 0}{\partial\theta} &= \frac{\partial}{\partial\theta}\int \frac{\partial \log f(y;\theta)}{\partial\theta} f(y;\theta)dy \qquad \text{see above} \\
&= \int \frac{\partial^2 \log f(y;\theta)}{\partial\theta^2} f(y;\theta)dy \\
&\quad + \int \frac{\partial \log f(y;\theta)}{\partial\theta}\frac{\partial f(y;\theta)}{\partial\theta}dy \\
&= \mathrm{E}_\theta\left(\frac{\partial^2 \log f(Y;\theta}{\partial\theta^2})\right) \\
&\quad + \int \frac{\partial \log f(y;\theta)}{\partial\theta}\frac{\partial \log f(y;\theta)}{\partial\theta} f(y;\theta)dy
\end{aligned}$$

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

Bartlett's second equation holds then as $\mathrm{E}\left(s(\theta;Y)\right) = 0$

**ML-Estimate**

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

for Fisher-regular distributions: $\hat\theta_{ML}$ has aysmptotically the smallest variance, given by the Cramér-Rao inequality,
$s\left(\hat\theta_{ML}; y_1, ..., y_n\right) = 0$
$\hat\theta \overset{a}{\sim} \mathrm{N}\left(\theta, I^{-1}(\theta)\right)$
If the true model is unknown, the distribution is
$\hat\theta \overset{a}{\sim} \mathrm{N}\left(\theta, I^{-1}(\theta)V(\theta)I^{-1}(\theta)\right)$ with $V(\theta)$ variance of the score function.
The ML-estimate is invariant: $\hat\gamma = g(\hat\theta)$ if $\gamma = g(\theta)$.

Proof:
$\gamma = g(\theta) \Leftrightarrow \theta = g^{-1}(\gamma)$
For the log-likelihood of $\gamma$ at the location $\hat\theta$ holds:

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

Then, the Fisher information is $\frac{\partial\theta}{\partial\gamma}I(\theta)\frac{\partial\theta}{\partial\gamma}$

Proof:

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

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

**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:

$$0 = s(\hat\theta_{ML}; y) \underset{Series}{\overset{Taylor}{\approx}} s(\theta; y) + \frac{\partial s(\theta;y)}{\partial\theta}(\hat\theta_{ML} - \theta) \Leftrightarrow$$

$$\hat\theta_{ML} \approx \theta - \left(\frac{\partial s(\theta;y)}{\partial\theta}\right)^{-1} s(\theta;y) \approx \theta - I^{-1}(\theta)s(\theta;y)$$

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 $\theta_{(t+1)} := \theta_{(t)} + \widehat{\mathrm{Var}}_{\theta_{(t)}}(t(Y))^{-1}\widehat{\mathrm{E}}_{\theta_{(t)}}(t(Y))$ 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 $2 \cdot lr(\theta,\hat{\theta}) \overset{a}{\sim} \chi^2_1$

Proof:

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

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

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

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

**Relation to Kullback-Leibler divergence**

$$\hat{\theta}_{ML} = \arg\ \min \mathrm{KL}(g,f)$$

with $f$ distributional model used and $g$ true model

Proof:

$$KL(g,f) = \int \log \frac{g(y)}{f(y)} g(y)dy$$
$$= \int \log(g(y))g(y)dy - \int \log(f(y))g(y)dy$$

To minimize that, the second component needs to be maximized. Its derivative is $\int s(\theta;y)g(y)dy = \mathrm{E}_g(s(\theta;Y)) = 0$

## 3.4 Consistency and Sufficiency

**Statistic**

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

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

**(Weak) Consistency**

$$MSE(\hat{\theta},\theta) \overset{n\to\infty}{\longrightarrow} 0 \Rightarrow \hat{\theta} \text{ consistent}$$

Proof:
$P(|\hat{\theta} - \mathrm{E}_\theta(\hat{\theta})| \geq \delta) \leq \frac{Var_\theta(\hat{\theta})}{\delta^2}$ using the inequality of Chebyshev and $MSE(t(.),\theta) = \mathrm{Var}_\theta(t(Y_1,...,Y_n)) + Bias^2((t(.);\theta)$

**Sufficiency**

A statistic $t(y_1,...,y_n)$ is sufficient for $\theta$, if the conditional distribution $f(y_1,...,y_n|t_0 = t(y_1,...,y_n);\theta)$ is independent of $\theta$.

**Neyman criterion:**

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

Proof:
"$\Rightarrow$":

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

"$\Leftarrow$":

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

Therefore:

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

**Minimal Sufficiency:**

$t(.)$ is sufficient and $\forall\, \tilde{t}(.)\, \exists\, h(.)$ s.t. $t(y) = h(\tilde{t}(y))$

# 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.:

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

| | "$H_0$" | "$H_1$" |
|---|---|---|
| $H_0$ | $1-p$ (correct) | $p$ (type I error) |
| $H_1$ | $\beta$ (type II error) | $1-\beta$ (correct) |

**Power** concerns the type II error

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

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

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

The $p$-value is uniformly distributed on $[0,1]$ under $H_0$.

**Confidence Interval**

$$[t_l(Y_{1:n}), t_r(Y_{1:n})] \text{ Confidence Interval}$$
$$\Leftrightarrow$$
$$P_\theta\left(t_l(Y_{1:n}) \leq \theta \leq t_r(Y_{1:n})\right) \geq 1-\alpha$$

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

**Corresponding Test**

$$\theta_0 \notin [t_l(y_{1:n}), t_r(y_{1:n})] \Leftrightarrow \text{“}H_1\text{”}$$

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

**Test for $\mu$, known $\sigma^2$ (Simple Gauss-Test)**
$H_0: \ \mu = \mu_0 \quad vs. \quad H_1: \ \mu \neq \mu_0$

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

**Test for $\mu$, unknown $\sigma^2$ (Simple t-Test)**
$H_0: \ \mu = \mu_0 \quad vs. \quad H_1: \ \mu \neq \mu_0$

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

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

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

**Wald Test**
$H_0: \ \theta = \theta_0 \quad vs. \quad H_1: \ \theta \neq \theta_0$

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

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

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

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

**Likelihood Ratio Test**
$H_0: \ \theta = \theta_0 \quad vs. \quad H_1: \ \theta \neq \theta_0$

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

**Neyman-Pearson Test**
$H_0: \ \theta = \theta_0 \quad vs. \quad H_1: \ \theta = \theta_1$

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

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: $\varphi^* = \begin{cases} 1 & \text{if } \frac{f(y;\theta_0)}{f(y;\theta_1)} \leq \mathrm{e}^c \\ 0 & \text{otherwise} \end{cases}$

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

$$\begin{aligned}
&P(\varphi^*{=}1|\theta_1) - P(\varphi{=}1|\theta_1) = \\
&= \int \{\varphi^*(y) - \varphi(y)\} f(y;\theta_1) dy \\
&\geq \frac{1}{\mathrm{e}^c} \int_{\varphi^*=1} \{\varphi^*(y) - \varphi(y)\} f(y;\theta_0) dy \qquad f(y;\theta_1) \geq \frac{f(y;\theta_0)}{\mathrm{e}^c} \\
&+ \frac{1}{\mathrm{e}^c} \int_{\varphi^*=0} \{\varphi^*(y) - \varphi(y)\} f(y;\theta_0) dy \qquad f(y;\theta_1) \leq \frac{f(y;\theta_0)}{\mathrm{e}^c} \\
&= \frac{1}{\mathrm{e}^c} \int \{\varphi^*(y) - \varphi(y)\} f(y;\theta_0) dy = 0
\end{aligned}$$

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

## 4.3 Tests for Goodness of Fit

**Discrete (Chi-Squared)**

$H_0\colon\ X_i \sim F_0 \quad vs. \quad H_1\colon\ X_i \sim F \neq F_0$

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

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 $\chi^2_{K-1-p}$-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)**

$H_0\colon\ X_i \sim F_0 \quad vs. \quad H_1\colon\ X_i \sim F \neq F_0$

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

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:

$$P(\sup_x |F_n(x) - F(x;\theta)| \leq t) =$$
$$= P(\sup_y |F^{-1}(y;\theta) - x| \leq t) \qquad {\scriptstyle x\in[0,1],\, x=F^{-1}(y;\theta) \atop F(F^{-1}(y;\theta);\theta)=y}$$
$$\overset{*}{=} P(\sup_y |\frac{1}{n}\sum_{i=1} \mathbb{1}_{\{U_i \leq y\}} - y| \leq t) \qquad \text{with } U_i \sim U(0,1)$$
$$^*F_n(F^{-1}(y;\theta)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq F^{-1}(y;\theta)\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{F(y;\theta)\leq y\}}$$

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

**Pivotal Statistic**

$$g(Y;\theta) \text{ pivotal}$$
$$\Leftrightarrow$$
$$\text{distribution of } g(Y;\theta) \text{ independent of } \theta$$

**Approximative Pivotal Statistic**

$H_0\colon\ X_i \sim F$ pivotal $\quad vs. \quad H_1\colon\ X_i \sim F$ not pivotal

$$g(\hat\theta;\theta) = \frac{\hat\theta - \theta}{\sqrt{\mathrm{Var}(\hat\theta)}} \overset{\alpha}{\sim} \mathrm{N}(0,1)$$

with $\hat\theta = t(Y) \overset{\alpha}{\sim} \mathrm{N}(\theta, \mathrm{Var}(\hat\theta))$

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

Proof:

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

## 4.4 Multiple Tests

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

For m tests:

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

$$FWER := P(\exists k : \text{“}H_1 k\text{”} | \forall k :\ H_0 k)$$

**Bonferoni Adjustment**

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

**Šidák Adjustment** only for independent tests

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

Proof:

$$\alpha \overset{!}{=} P\left(\cup_{k=1}^{m}(p_k \leq \alpha) | H_{0k}, k = 1, ..., m\right)$$
$$= 1 - (1-\alpha)^{1/m}$$

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

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

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 $n << m$ problems

$$FDR = \mathrm{E}\left(\frac{\#\text{“}H1\text{”}|H_0}{\#\text{“}H1\text{”}}\right)$$

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

j is largest index s.t. $p(j) \leq \alpha j/m$, reject all $H_0 i$ for $i \leq j$

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

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

For the normal linear model:

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

**Attributes of the Regression Line**

$$\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x}) \\
\hat{e}_i &= y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= y_i - (\bar{y} + \hat{\beta}_1(x_i - \bar{x})) \\
\sum_{i=1}^n \hat{e}_i &= \sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y} - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) \\
&= n\bar{y} - n\bar{y} - \hat{\beta}_1(n\bar{x} - n\bar{x}) = 0 \\
\bar{\hat{y}} &= \frac{1}{n}\sum_{i=1}^n \hat{y}_i = \frac{1}{n}(n\bar{y} + \hat{\beta}_1(n\bar{x} - n\bar{x})) = \bar{y}
\end{aligned}$$

**Estimates (OLS)**

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

Proof:
$Cov(x,y) = Cov(x, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}) = \hat{\beta}_1 Var(x)$
$\iff \hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)}$

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

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

The estimates are the same as for the ML procedure.

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

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

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

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

### 5.1.2 Multivariate Linear Model

**Theoretical Model**

$$Y = X\beta + u$$

**Empirical Model**

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

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

$$y = (y_1, ..., y_n)^T,\ e = (e_1, ..., e_n)^T,\ X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$$

**Assumptions**

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

For the normal linear model:

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

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

$$\hat{\beta} = \left((X^TX)^{-1}\right)X^Ty$$

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

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

The estimates are the same as for the OLS procedure.
$\beta$ is the **B**est **L**inear **U**nbiased **E**stimator

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

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

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

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

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

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

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

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

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 captured by confidence intervals) plus residual variance.

**Quantile**

$$Q(\tau) = \inf\{y : F(y) \geq \tau\}$$

If $F$ is invertable: $Q(\tau) = F^{-1}(\tau)$, $\tau \in (0,1)$

**Model**

$$Q(\tau|x_{1:p}) = X\beta$$

For median regression: $\hat\beta = \arg\min \sum_{i=1}^n |y_i - x_i^T\beta|$

In general:

$$\hat Q(\tau) = \arg\min_\beta \left(\sum_{i=1}^n \delta_\tau(y_i - x_i^T\beta)\right)$$

with check function $\delta_\tau(y) = y\left(\tau - \mathbb{1}_{\{y<0\}}\right)$

Proof:

$$\begin{aligned} Q(\tau) &= \arg\min_q E(\delta_\tau(Y-q)) \\ &= \arg\min_q \left\{(\tau{-}1)\int_{-\infty}^{q}(y{-}q)f(y)dy + \tau\int_{q}^{\infty}(y{-}q)f(y)dy\right\} \end{aligned}$$

Differentiating w.r.t. $q$ gives $(\tau{-}1)\int_{-\infty}^q f(y)dy - \tau\int_q^\infty f(y)dy$
$= (1{-}\tau)F(q) - \tau(1{-}F(q) = F(q) - \tau$

**Estimates**
The estimates for $\beta$ can be computed with linear programming and are normally distributed with mean $\beta$.

### 5.1.5 Flexible Regression

**Assumptions**

- **Independent Observations** $y_1, ...y_n$ are independent
- **Constant Variation** $Var(Y|x) = \sigma^2$
- **Normality** $e_i|x_{1:p} \sim N(0,\sigma^2)$ ; $Y|x \sim N(\hat y, \sigma^2)$

**Knot Placement**

- equidistant
- based on quantiles (more structure where data is dense)
- all data points plus penalization

**Penalized Regression Splines**

$$||y-X\beta||^2 + \lambda \int_{x_1}^{x_n} \left[f''(x)\right]^2 dx = ||y-X\beta||^2 + \lambda \beta^T D\beta$$

$l_p(\beta, \sigma^2, \lambda) = l(\beta, \sigma^2) - \frac{\lambda}{2\sigma_\epsilon^2}\beta^T D\beta$

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

**Difference Penalty**

- first order: $\beta^T D\beta = \sum_{j=1}^p (\beta_{j+1} - \beta_j)^2$
- second order: $\beta^T D\beta = \sum_{j=1}^p (\beta_{j+1} - 2\beta_j + \beta_{j-1})^2$

**Choosing $\boldsymbol{\lambda}$** Model complexity

$$\dim(\lambda) = tr\left\{(X^T X + \lambda D)^{-1}(X^T X)\right\}$$

$AIC(\lambda) = fit(\lambda) + 2\dim(\lambda)$

Numerically complex. Alternative: **Bayes**

$\beta \sim \mathrm{N}(0, \sigma_\beta^2 D^-)$ with $(D^-)^- = D$ (generalized inverse)

$$\log f(\beta, \sigma^2; \sigma_\beta^2 | y) \propto l(\beta, \sigma^2) - \frac{rk(D^-)}{2}\log(\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2}\beta^T D^- \beta$$

As $\lambda = \frac{1}{\sigma_\beta^2}$, marginal posterior for $\sigma_\beta^2$ can be derived. E. g. set $\lambda$ to the posterior mode estimate.

### 5.1.6 Generalized Regression

**Assumptions**

- **Independent Observations** $y_1, ...y_n$ are independent
- **Linearity of the Mean** $\mathrm{E}(Y|x_{1:p}) = X\beta$ or $\mathrm{E}(e|x_{1:p}) = 0$
- **Exponential Family** $Y|x \sim \exp\{t(y)\theta(x) - \kappa(\theta(x))\}\, h(y)$

**Link Function**

Linear predictor $\eta = X\beta$; $\mu = \frac{\partial\kappa(\theta)}{\partial\theta} = \mathrm{E}(t(Y); \theta)$

$$\mu = g^{-1}(\eta)$$

If $\lambda = 0$, *canonical link*:

$$g(\theta) = \eta$$

- score function: $s(\beta) = X^T\,(t(y) - \mathrm{E}(t(Y); \eta))$
- estimate $\hat{\beta} = X^T \mathrm{E}(t(Y); \hat{\eta}) = X^T t(y)$
- Fisher matrix $I(\beta) = X^T W X$
  with $W$ diagonal and $W_{ii} = \frac{\partial^2 \kappa(\eta_i)}{\partial \eta^2} = Var(t(Y_i), \eta_i)$

Examples:

- **Logistic**: $\mathrm{logit} P(Y_i{=}1|x_i) = \log \frac{P(Y_i=1|x_i)}{1-P(Y_i=1|x_i)} = \eta$
  $Var(Y_i|x_i) = P(Y_i{=}1|x_i) \cdot (1{-}P(Y_i{=}1|x_i))$
- **Poisson**: $log\mathrm{E}(Y_i|x_i) = \eta$
  $Var(Y_i|x_i) = \mathrm{E}(Y_i|x_i) = e^\eta$

### 5.1.7 Weighted Regression

**Different Precision** variance heterogeneity: $e_i \sim \mathrm{N}(0, \sigma_i^2)$

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

with $W = diag(\frac{1}{a_1}, ..., \frac{1}{a_n})$ and $a_i = \frac{\sigma_i^2}{\sigma^2}$

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

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

**Different Group Representation**

$$\mathrm{Y}_i | x_{i,1:p}, z_i \sim \mathrm{N}(x_{i,1:p}\beta_{z_i}, \sigma^2)$$

with $z_i$ indicating group affiliation

## 5.2 Goodness of Fit

### 5.2.1 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 Bayesian Statistics

## 6.1 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:

$$f_{post}(\theta|X) = \frac{f(X|\theta) \cdot f_\theta(\theta)}{\int f(X|\tilde{\theta}) f_\theta(\tilde{\theta}) d\tilde{\theta}}$$
$$= C \cdot f(X|\theta) \cdot f_\theta(\theta) \quad \text{choose C s.t.} \int f_{post}(\theta|X) = 1$$
$$\propto f(X|\theta) \cdot f_\theta(\theta)$$

**Point Estimates**

$$\hat{\theta}_{postmean} = E_0(\vartheta|x) = \int_{\vartheta \in \Theta} \vartheta f_\theta(\vartheta|x) d\vartheta$$
$$\hat{\theta}_{postmode} = \underset{\vartheta}{\operatorname{argmax}} f_\theta(\vartheta, x)$$
$$\hat{\theta}_{Bayesrisk} = \underset{t(.)}{\operatorname{argmin}} R_{Bayes}(t(.))$$

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

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

with posterior Bayes risk:
$R_{postBayes}(t(.)|y) = \int \mathcal{L}(t(y), \vartheta) f_\theta(\vartheta|y) = \mathrm{E}_{\theta|y}(\mathcal{L}(t(y), \theta)|y)$

For squared loss: $\hat{\theta}_{postBayesrisk} = \hat{\theta}_{postmean}$

**Credibility Interval**

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

- 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**

$f_\theta(\theta) = const.$ 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** transformation-invariant
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 $\mathrm{E}(KL(f_\theta(.), f_{post}(., x))$

**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 |
|---|---|---|
| $\pi \sim \mathrm{Be}(\alpha, \beta)$ | $\mathrm{Bin}(n, \pi)$ | $\mathrm{Be}(\alpha{+}k, \beta{+}n{-}k)$ |
| $\mu \sim \mathrm{N}(\gamma, \tau^2)$ | $\mathrm{N}(\mu, \sigma^2)$ | $\mathrm{N}(.,.) \overset{n\to\infty}{\longrightarrow} \mathrm{N}(\bar{y}, \frac{\sigma^2}{n})$ |
| $\sigma^2 \sim \mathrm{IG}(\alpha, \beta)$ | $\mathrm{N}(\mu, \sigma^2)$ | $\mathrm{IG}(\alpha{+}\frac{n}{2}, \beta{+}\frac{1}{2}\sum_{i=1}^n (y_i{-}\mu)^2)$ |
| $\lambda \sim \mathrm{Ga}(\alpha, \beta)$ | $\mathrm{Po}(\lambda)$ | $\mathrm{Ga}(\alpha{+}n\bar{y}, \beta{+}n)$ |

## 6.2 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**

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

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

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

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 $\theta_1, ..., \theta_N$, 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. $\frac{f(x)}{g(x)} \leq M \; \forall x$ with $f(x) > 0$ when $g(x) > 0$
  1. Draw candidate $y \sim g(x)$
  2. Acceptance probability $\alpha$ for $y$: $\alpha = \frac{f(x)}{Mg(x)}$
  3. Draw $u \sim U(0,1)$ and accept if $u \leq \alpha$, else: step 1

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

- **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 $\theta_1^*, ..., \theta_n^*$:

1. Draw candidate $\theta^*$ from proposal distribution $q\left(\theta|\theta_{(t)}^*\right)$

2. Accept $\theta_{(t+1)}^* = \theta^*$ with probability
   $$\alpha(\theta_{(t)}|\theta^*) = \min\left\{1, \frac{f_{post}(\theta^*|y) \, q\left(\theta_{(t)}^*|\theta^*\right)}{f_{post}\left(\theta_{(t)}^*|y\right) q\left(\theta^*|\theta_{(t)}^*\right)}\right\}$$
   else choose $\theta_{(t+1)}^* = \theta_{(t)}^*$

This sequence has a stationary distribution for $n \to \infty$.
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 seperately:
$$\theta_{t+1,i}^* \sim f_{\theta_i|y,\theta\backslash\theta_i}\left(\theta_i^*|y, \theta_{t^*,i}\right)$$
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 $q_\theta = \min_{q_\theta \in Q} KL(f_{post}(.|X), q_\theta(.))$

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

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

# 7 Sampling

**Bootstrap**

1. Draw $y_i^*$: $n$ samples with replacement from $y$
2. Calculate the statistic of interest $t(y_i^*)$
3. Repeat this B times
4. *Plug-in Principle*: Whenever the distribution function is involved in estimating a statistic, use the empirical bootstrapped version instead.

In a **Parametric Bootstrap** the parameter is first estimated from the data and then Bootstrap samples are drawn from the resulting distribution.

**Bootstrap Probability**

$$P(Y_i \in Y^*) = 1 - (1 - \frac{1}{n})^n \overset{n\to\infty}{\to} 1 - e^{-1} \approx 0.632$$

**Subsampling**

- **replacement** $m$-out-of-$n$ bootstrap
- **non-replacement** subsampling directly from true $F$

**Permutation Test** for two variables

1. Calculate $t(x, y)$, e. g. differences in mean, correlation...
2. Draw samples $x^*$, $y^*$ of size $n$ from $x$ and $y$ without replacement ("shuffel")
3. Calculate $t(x^*, y^*)$
4. $p$-value $= \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\{t(x_b^*, y_b^*) \geq t(x,y)\}}$

For a **Boostrap Test** do step 2 with replacement.

**Bootstrap in Regression**

- **Residual based**: 1. Get Bootstrap sample $e_i^*$ from fitted residuals $\hat{e} = y - X\hat{\beta}$, 2. Calculate new response $y_i^* = x_i\hat{\beta} + e_i^*$, 3. Calculate $\hat{\beta}^*$
- **Model based** 1. Draw a sample from $e_i \sim \mathrm{N}(0, \hat{\sigma}^2)$, 2. Calculate new response $y_i^* = x_i\hat{\beta} + e_i^*$, 3. Calculate $\hat{\beta}^*$
- **Pairwise** 1. Draw $(y_i^*, x_i^*)$ from the original sample for $i = 1, .., n$, 2. Calculate $\hat{\beta}^*$
- **Wild** Set $\hat{e}_i^* = V_i^* \hat{e}_i$, with $V_i^*$ from the 2-point distribution $P(V_i^* = \frac{\sqrt{5}+1}{2}) = \frac{\sqrt{5}-1}{2\sqrt{5}}$ and $P(V_i^* = -\frac{\sqrt{5}-1}{2}) = \frac{\sqrt{5}+1}{2\sqrt{5}}$, chosen as $\mathrm{E}(V_i^*) = 0$, $Var(V_i^*) = 1$, $\mathrm{E}(V_i^{*3}) = 1$

**Consistency of a Bootstrap Estimator**

$$\lim_{n\to\infty} P_n \left\{ \sup_t |G_n(t, F_n) - G_\infty(t, F)| > \epsilon \right\} = 0 \ \ \forall\epsilon$$

with $F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq y\}}$ empirical distribution function, $G_n(t, F) = P(T_n \leq t)$ exact finite sample distribution, and $P_n$ joint probability of the sample

The bootstrap estimate is inconsistent for the maximum of a sample or if the $\theta$ lies on the boundary of $\Theta$.

**Mallow's Metric**

$$\rho_p(F, G) = \inf_{\mathcal{T}_{XY}} \{E(|X - Y|)^p\}^{\frac{1}{p}}$$

for $F, G$ in the set of distributions where $\int_{-\infty}^{\infty} |t|^p dF(t) < \infty$; $(X, Y) \sim T \in \mathcal{T}_{XY}$ with $X \sim F$ and $Y \sim G$

**Theorem of Beran and Ducharme**

$G_n(., F_n)$ is consistent if $\forall\epsilon > 0, F$ the following holds:

1. $\lim_{n\to\infty} P_n(\rho(F_n, F) > \epsilon) = 0$
2. $G_\infty(t, F)$ is a continuous function of t
3. $\forall t$ and sequences $\{H_n\}$ s. t. $\lim_{n\to\infty} \rho(H_n, F) = 0$ holds: $G_n(t, H_n) \to G_\infty(t, F)$

# 8 Model Selection

**AIC (Akaike Information Criterion)**

$$AIC = -2\sum_{i=1}^{n} \log f(y_i;\hat{\theta}) + 2p$$

The AIC estimates $2\mathrm{E}_Y\left\{\mathrm{KL}(g,f)\right\} - 2\int \log(g(y))g(y)dy$. The latter component is unknown, so the absolute value of the AIC is not informative. The AIC favours complex models.

For regressions: $AIC = 2n\log(\hat{\sigma}^2) + 2(p+2)$

**The AIC as theoretical cross validation**
The AIC minimizes $\mathrm{E}_{Y_{1:n}}\left\{\mathrm{E}_Y\left[Y-\hat{\mu}\right]^2\right\}$ if we use the MSE instead of the Kullback-Leibler divergence. This can be estimated via cross validation.

**Bias Corrected AIC**

$$AIC_{corr} = -2\sum_{i=1}^{n} \log f(y_i;\hat{\theta}) + 2p\left(\frac{n}{n-p-1}\right)$$

should be preferred if $\frac{n}{p} < 40$

**BIC (Bayesian Information Criterion)**

$$BIC = -2\sum_{i=1}^{n} \log f(y_i;\hat{\theta}) + \log(n)p$$

approximately maximizes the posterior probability of a model and selects less complex models as the AIC

**DIC (Deviance Information Criterion)** Bayesian AIC

$$DIC = D(y,\hat{\theta}_{postmean}) + 2p_D = \int D(y,\vartheta)f_{post}(\vartheta|y)d\vartheta + p_D$$

with deviance $D(y;\theta) := -2l(\theta)$ the difference in likelihood compared to the full model and $\Delta D(y;\theta,\hat{\theta}) = 2\left\{l(\hat{\theta}) - l(\theta)\right\} \overset{a}{\sim} \chi^2_p$ the difference in deviance

$p_D := \mathrm{E}(\Delta D(y;\theta,\hat{\theta}_{\substack{post\\mean}}|y) = \int D(y,\vartheta)f_{post}(\vartheta|y)d\vartheta - D(y,\hat{\theta}_{\substack{post\\mean}})$
The integral can be approximated using MCMC.

**Model Averaging** Using probabilities as weights

$$P(M_k|y) := \frac{\exp(-\frac{1}{2}\Delta IC_k)}{\sum_{k'=1}^{K}\exp(-\frac{1}{2}\Delta IC'_k)}$$

with $\Delta IC_k = IC_k - min(IC)$
For regressions: $P(\text{covariate } x|y) = \sum_{k=1}^{K} \mathbb{1}_{\{x_i \text{ in } M_k\}} P(M_k|y)$

**Inference After Model Selection** neglect is a quiet scandal

$$\begin{aligned} Var(\hat{\theta}) &= \mathrm{E}_{model}(Var(\hat{\theta}|model)) + Var_{model}(E(\hat{\theta}|model)) \\ &= \sum_{k=1}^{K}\pi_k Var_k(\hat{\theta}) + \sum_{k=1}^{K}\pi_k(\theta_k - \bar{\theta})^2 \end{aligned}$$

The last component depends on the true parameter and will be biased if the estimates are used.

Solutions:

- $\widehat{Var}(\hat{\theta}) = \left[\sum_{k=1}^{K}\pi_k\sqrt{\widehat{Var}_k(\hat{\theta}_k) + (\hat{\theta}_k - \hat{\bar{\theta}})^2}\right]^2$
- Use the Variance of the full (saturated) model
- Use bootstrap for confidence intervals

**Lasso** least absolute shrinkage and selection operator

$$l_p(\theta,\lambda) = l(\theta) - \lambda\sum_{j=1}^{p}|\theta_j|$$

This penalized log likelihood can be solved with iterative quadratic programming using a Taylor expansion.
Using Bayesian view the penalty corresponds to a prior: $f_{\theta_j}(\theta_j) \propto \exp(-|\theta_j|)$ (Laplace prior)

# 9 Dimensionality Reduction

**Covariance Matrix $\Sigma$**

- symmetric, $\in \mathbb{R}^{n \times n}$ therefore $\frac{q(q+1)}{2}$ parameters
- positive definite, i. e. $\forall a \in \mathbb{R}^q : a^T \Sigma a \geq 0$

**Marginal Independence**

$$\Sigma_{jk} = 0 \Leftrightarrow Y_{ij} \text{ and } Y_{ik} \text{ are independent}$$

**Conditional Independence**

$\Omega = \Sigma^{-1}_{jk} = 0 \Leftrightarrow Y_{ij}$ and $Y_{ik}$ are independent given all other $Y$ with concentration matrix $\Omega$

Proof:
$$f(y_{.j}, y_{.k} | y_{.\overline{j,k}}) = \frac{f(y)}{f_{.\overline{j,k}}} \propto f(y) \overset{\mathrm{N}(\mu,\Sigma)}{\propto} \exp\left\{-\frac{1}{2} y^T \Sigma^{-1} y\right\}$$

**Graphical Models**

visualize conditional dependences in a graph

**Principal Component Analysis (PCA)**

1. Use singular value decomposition $\Sigma = U \Lambda U^T$ with $U$ matrix of orthonomal eigenvectors and $\Lambda = diag(\lambda_1, ..., \lambda_q)$ matrix of sorted eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_q$ of $\Sigma$
2. Prune smallest $k = q - r$ eigenvalues in $\tilde{\Lambda}$
3. Simplify model with spectral decomposition $\tilde{Y} = \tilde{V} \tilde{\Lambda}^{1/2} \tilde{U}^T$ with $\tilde{V}$, $\tilde{U}$ first $r$ eigenvectors of $YY^T$ and $Y^T Y$ respectively
4. explained variance $\sum_{i=1}^{r} \lambda_i / \sum_{i=1}^{q} \lambda_i$

Proof:
Karhunen-Loève expansion: $U \Lambda^{\frac{1}{2}} Z_. \sim \mathrm{N}(0, U \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} U^T) = \mathrm{N}(0, \Sigma)$ with $Z_. \sim \mathrm{N}(0, \mathbb{1})$, therefore $\tilde{Y}_. = \tilde{V} \tilde{\Lambda}^{\frac{1}{2}} \tilde{Z}_.$
With spectral decomposition: $Y = V \Lambda^{\frac{1}{2}} U^T$ (for column-centered $Y$)

# 10 Missing/Deficient Data

**Missing Completely at Random (MCAR)** independent

$$P(R_i|Y_i) = P(R_i)$$

with $R_{ij} = \begin{cases} 0 & \text{if } Y_{ij} \text{ missing} \\ 1 & \text{otherwise} \end{cases}$ and $R_i = (R_{i1}, ..., R_{iq})$

A complete case analysis will lead to unbiased results.

**Missing at Random (MAR)** depends on observed variables

$$P(R_i|Y_i) = P(R_i|Y_{iO_i})$$

with $O_i = \{j : R_{ij} = 1\}$ and $M_i = \{j : R_{ij} = 0\}$

Complete case analysis $P(Y|X, Z)$:

- only response $Y_i$ MAR: unbiased
- only covariate $X_i$ MAR: biased
  Asymptotically unbiased with *inverse probability weighting*:
  1. Estimate $\pi(y_i, z_i) = P(R_{X_i}{=}1|y_i, z_i)$
  2. Use weighted score function $\hat{s}_w(\theta) = \sum_{i=1}^n \frac{R_{X_i}}{\hat{\pi}} s_i(\theta)$
- both MAR: biased and $\pi(y_i, z_i)$ can not be estimated due to missing $Y_i$

**Missing Not at Random (MNAR)**

$$P(R_i|Y_i) \neq P(R_i|Y_{iO_i})$$

Can not be corrected to be unbiased.

**EM Algorithm** replace $y_{iM}$ by $E(Y_{iM}|y_{iO})$

Expectation Step:

$$Q(\theta, \theta_{(t)}) = \sum_{i=1}^n \int l_i(\theta) f(y_{iM}|y_{iO}; \theta_{(t)}) dy_{iM}$$

Maximization Step:

$$\frac{\partial Q(\theta, \theta_{(t)})}{\partial \theta} = s(\theta, \theta_{(t)}) \stackrel{!}{=} 0$$

**Louis' Formula for Variance Estimates in EM Settings**

$$J_O(\theta) = \sum_{i=1}^n \{\mathrm{E}(J_i(\theta)|y_{iO}) - \mathrm{E}(s_i(\theta)s_i(\theta)|y_{iO}) + s_{iO}(\theta)s_{iO}(\theta)\}$$

**Multiple Imputation** EM but considers estimation variability

1. Create $K$ complete datasets by simulating missing data $\sim f_{post}(y_{iM}|y_{iO})$
2. Fit $K$ models $Y_i \sim f(y|\theta)$
3. Rubin's Rule: $\hat{\theta}_{MI} = \frac{1}{K}\sum_{k=1}^K \hat{\theta}^*_{(k)}$;
   $\widehat{\mathrm{Var}}(\hat{\theta}_{MI}) = \hat{V} + (1+\frac{1}{K})\bar{B}$ with $\hat{V} = \frac{1}{K}\sum_{k=1}^K I^{-1}(\hat{\theta}^*_{(k)})$ and $\bar{B} = \frac{1}{K-1}\sum_{k=1}^K (\hat{\theta}^*_{(k)} - \hat{\theta}_{MI})(\hat{\theta}^*_{(k)} - \hat{\theta}_{MI})^T$

**Estimate Accuracy**

$$\hat{\mu}_g - \mu_g = \rho_{R_g} \times \sigma_g \times \sqrt{\frac{N-n}{n}}$$

with $\rho_{R_g}$ data quality (correlation between $R_j$ and $g(Y_j)$), $\sigma_g$ variability, and $\sqrt{\frac{N-n}{n}}$ data quantity; $g$ some known function

- MCAR: $MSE(\hat{\mu}_g) = \frac{1}{N-1} \times \sigma_g^2 \times \frac{N-n}{n}$
- MNAR: $MSE(\hat{\mu}_g) = \mathrm{E}(\rho^2_{R_g}) \times \sigma_g^2 \times \frac{N-n}{n}$
  $n_{eff} = \frac{\frac{n}{N}}{1-\frac{n}{N}} \frac{1}{\mathrm{E}(\rho^2_{R_g})}$

**Measurement Error**

$$U = X - X_m \text{ with } \mathrm{E}(U) = \mu_U \text{ and } \mathrm{Var}(U) = \sigma_U^2$$

with $\mu_U$ systematic error (bias/validity), $\mathrm{Var}(U)$ (reliability)

In Regression Settings:

- **error in Y**: $Y_m = \beta_0 + \beta_1 X + \epsilon + U$ and $\mathrm{E}(Y_m|X) = \beta_0 + \mu_U + \beta_1 X$ leads to biased $\hat{\beta}_0$
- **error in X**: $Y = \beta_0 + \beta_1 X + \epsilon$ and $X_m = X + U$ leads to biased $\hat{\beta}_0$ and $\hat{\beta}_1$, the latter is attenuated by the inverse of reliability ratio $rr = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} = \frac{\sigma^2_{X_m} - \sigma_U^2}{\sigma^2_{X_m}}$
  Getting information about $\sigma_U^2$:
  - **Validation Data** with both $X$ and $X_m$ observed
  - **Replication Data** repeated measures of $X_m$
  - **Assumptions** e. g. $\sigma_U^2 = 0$ (naive estimator)

# 11 Experiment Design

**Omitted Variables**

Regression setting ignoring omitted Variables:

$$\int f_{Y|X,Z,U} f_{Z,U} dzdu = \int \frac{f_{Y,X,Z,U}}{f_{X|Z,U}} dzdu \neq f_{Y|X}$$

with $Z$ observable and $U$ unobservable quantities influencing $Y$

Solutions:

- *Randomization*: randomly assign $X$ and then observe $Y$
- *Balancing*: make $X$ independent of $Z$

**Analysis of Variances (ANOVA)** of one categorical variable

Linear constraint: $\sum_{k=1}^{K} n_k \beta_k = 0$ (usually controlled over $\beta_K$)

$$\hat{\beta}_{0,ML} = \hat{\mu}_{ML} = \sum_{k=1}^{K} \sum_{j=1}^{n_k} \frac{y_{kj}}{n} = \bar{y}_{..}$$

$$\hat{\beta}_{k,ML} = \sum_{j=1}^{n_k} \frac{y_{kj} - \bar{y}_{..}}{n_k} = \bar{y}_{k.} - \bar{y}_{..}$$

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

with

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

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

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

**F-Test**

$$F = \frac{SS_{Explained}/(df_0 - df_X)}{SS_{Residual}/df_X} \sim \mathcal{F}_{df_0 - df_X, df_X}$$

with $df_0 = n-1$ and $df_X = n-K$

**Block Design** account for block effects

$$Y_{kbj} = \mu + \beta_k + \alpha_k + \epsilon_{kbj}$$

Linear constraints: $\sum_{k=1}^{K} n_{k.} \beta_k = 0$ and $\sum_{b=1}^{B} n_{.b} \alpha_b = 0$

For the F-Test: $df_0 = df_Z = n-B$ and $df_{X+Z} = n-K-B+1$

**Latin Squares** Sodoku pattern for more variables

**Instrumental Variable**

$$Y = \beta_0 + X\beta_X + \overbrace{U\beta_U + \epsilon}^{\tilde{\epsilon}} \Rightarrow \frac{\partial Y}{\partial X} = \beta_X + \frac{\partial \tilde{\epsilon}}{\partial X}$$

Construct instrumental variable $Z$: $Cov(Z,\epsilon)=0$, $Cov(Z,X)\neq 0$:

$$\frac{\partial Y}{\partial X}|(U=u) = \frac{\partial Y/\partial Z}{\partial X/\partial Z}$$

i.e. fit two regressions $Y|Z$ and $X|Z$ and set $\hat{\beta}_X = \frac{\hat{\beta}_{YZ}}{\hat{\beta}_{XZ}}$

**Propensity Score**

$$\tau = \mathrm{E}(Y(1)|D=1) - \mathrm{E}(Y(0)|D=1)$$

with $\tau$ average treatment effect on the treated, $Y(1)$ response if treated, $Y(0)$ analogous; $D_i$ indicator if $i$ is influenced by the treatment

$$\mathrm{E}(Y(1)|D=1) - \mathrm{E}(Y(0)|D=0) = \tau + \underbrace{\mathrm{E}(Y(0)|D=1) - \mathrm{E}(Y(0)|D=0)}_{\text{selection bias}}$$

If the selection bias is zero, $D$ and $X$ are unconfounded.

$$\hat{\tau} = \sum_{i=1}^{n} (Y_i(1) - Y_{j(i)}(0))$$

after matching treated individual $i$ with individual $j(i)$ from non-treatment group