Statistics
Collection of Formulas

Contents

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

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

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

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

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

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

Phi-Coefficient

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

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

Cramér's V

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

Range: $0 \le V \le 1$

Contingency Coefficient C

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

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

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 $0 \le C_{corr} \le 1$

Odds-Ratio

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

Range: $0 \le OR < \infty$

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: $-1 \le \gamma \le 1$

Kendall's τ_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: $-1 \le \tau_b \le 1$

Kendall's/Stuart's τ_c

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

Range: $-1 \le \tau_c \le 1$

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: $-1 \le \rho \le 1$

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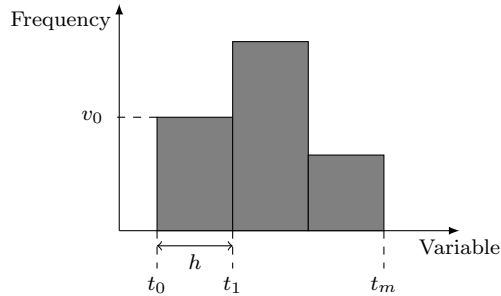$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ or $s_{yy} = \frac{S_{yy}}{n}$

Range: $-1 \le r \le 1$

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)

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:

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

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

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) $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
$\forall A_i \in \mathcal{A}, i = 1, ..., \infty$ 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 $A_i, ..., A_n$ complete decomposition of Ω into pairwise disjoint events

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

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

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 independent $\Leftrightarrow P(A \cap B) = P(A) + P(B)$

X, Y 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

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

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 $f(\cdot)$:**

  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

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

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

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

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

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

Marginal Density

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

Conditional Density

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

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:
>
> $$Var(Y) = \int (y - \mu_Y)^2 f(y)dy$$
> $$= \int (y - \mu_Y)^2 f(y|x)f(x)dydx$$
> $$= \int (y - \mu_{Y|x} + \mu_{Y|x} - \mu_Y)^2 f(y|x)f(x)dydx$$
> $$= \int (y - \mu_{Y|x})^2 f(y|x)f(x)dydx +$$
> $$\int (\mu_{Y|x} - \mu_Y)^2 f(y|x)f(x)dydx +$$
> $$2 \int (y - \mu_{Y|x})(\mu_{Y|x} - \mu_Y)f(y|x)f(x)dydx$$
> $$= \int Var(Y|x)f(x)dx + \int (\mu_{Y|x} - \mu_Y)^2 f(x)dx$$
> $$= E_X(Var(Y|X)) + Var_X(E(Y|X))$$

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 μ and σ^2

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

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

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

Multivariate Normal   symmetric with μ_i and Σ

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

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

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

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 μ with unknown (estimated) variance and ν 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

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

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

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

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

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

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

Exponential   Time between Poisson events

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

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

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

Chi-Squared   squared standard normal random variables with ν degrees of freedom

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

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

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

2.4.3   Exponential Family

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

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

with $h(y) \geq 0$, $t(y)$ vector of the canonical statistic, θ 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 $\frac{\partial^2 \kappa(\theta)}{\partial \theta^2} = \text{Var}(t(Y))$

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**

2.5   Limit Theorems

Law of Large Numbers

Central Limit Theorem
$$Z_n \xrightarrow{d} \text{N}(0,\sigma^2)$$
with $Z_n = \sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}}$ and Y_i i.i.d. with expectation 0 and variance σ^2

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

$$M_{Z_n}(t) = E\left(e^{t(Y_1+Y_2+...+Y_n)/\sqrt{n}}\right)$$

$$= E\left(e^{tY_1/\sqrt{n}} \cdot e^{tY_2/\sqrt{n}} \cdot ... \cdot e^{tY_n/\sqrt{n}}\right)$$

$$= E\left(e^{tY_1/\sqrt{n}}\right) E\left(e^{tY_2/\sqrt{n}}\right) ... E\left(e^{tY_n/\sqrt{n}}\right)$$

$$= M_Y^n(t/\sqrt{n})$$

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

$$\left.\frac{\partial K_{Z_n}(t)}{\partial t}\right|_{t=0} = \frac{n}{\sqrt{n}} \left.\frac{\partial K_Y(t)}{\partial t}\right|_{t=0} = \sqrt{n}\mu$$

$$\left.\frac{\partial^2 K_{Z_n}(t)}{\partial t^2}\right|_{t=0} = \frac{n}{n} \left.\frac{\partial^2 K_Y(t)}{\partial t^2}\right|_{t=0} = \sigma^2$$

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

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 θ, $\mathcal{L}(\theta, \theta) = 0$ holds

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

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

As θ 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

$$R(t(.), \theta) = \mathrm{E}_\theta \left(\mathcal{L}(t(Y_1, ..., Y_n), \theta) \right)$$

$$= \int_{-\infty}^{\infty} \mathcal{L}(t(Y_1, ..., Y_n), \theta) \prod_{i=1}^{n} f(y_i; \theta) dy_i$$

Minimax Approach

The risk still depends ton the true parameter θ. Tentative estimation: Choose θ, so that 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)

$$MSE(t(.), \theta) = \mathrm{E}_\theta \left(\{t(Y) - \theta\}^2 \right)$$

$$= \mathrm{Var}_\theta \left(t(Y_1, ..., Y_n) \right) + Bias^2((t(.); \theta)$$

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

> Proof:
> Let $\mathcal{L}(t, \theta) = (t - \theta)^2$
> $$R(t(.), \theta) = \mathrm{E}_\theta(\{t(Y) - \theta\}^2)$$
> $$= \mathrm{E}_\theta(\{t(Y) - \mathrm{E}_\theta(t(Y)) + \mathrm{E}_\theta(t(Y)) - \theta\}^2)$$
> $$= \mathrm{E}_\theta(\{t(Y) - \mathrm{E}_\theta(t(Y))\}^2) + \mathrm{E}_\theta(\{\mathrm{E}_\theta(t(Y)) - \theta\}^2)$$
> $$+ 2\mathrm{E}_\theta(\{t(Y) - \mathrm{E}_\theta(t(Y))\}\{\mathrm{E}_\theta(t(Y)) - \theta\})$$
> $$= \mathrm{Var}_\theta(t(Y_1, ..., Y_n)) + Bias^2((t(.); \theta) + 0$$

Cramér-Rao Inequality

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

> Proof:
> For unbiased estimates: $\theta = \mathrm{E}_\theta(\hat{\theta}) = \int t(y) f(y; \theta) dy$
> $$1 = \int t(y) \frac{\partial f(y; \theta)}{\partial \theta} dy$$
> $$= \int t(y) \frac{\partial \log f(y; \theta)}{\partial \theta} f(y; \theta) dy$$
> $$= \int t(y) s(y; \theta) f(y; \theta) dy$$
> $$= \int (t(y) - \theta) (s(\theta; y) - 0) f(y; \theta) dy \quad \begin{array}{c} \text{1. Bartlett equation} \\ \mathrm{E}_\theta(s(\theta; y)) = 0 \end{array}$$
> $$= \mathrm{Cov}_\theta(t(Y); s(\theta; Y))$$
> $$\geq \sqrt{\mathrm{Var}_\theta(t(Y))} \sqrt{\mathrm{Var}_\theta(s(\theta; Y))} \quad \text{Cauchy-Schwarz}$$
> $$= \sqrt{MSE(t(Y); \theta)} \sqrt{I(\theta)}$$

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 ≥ 0 otherwise.

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

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

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

- $\theta \in \mathbb{R}^p$

- $f(.;\theta)$ Fisher-regulär:

  - $\{y : f(y; \theta > 0)\}$ unabhängig von θ

  - Möglicher Parameterraum Θ 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** $L(\theta; y_1, ..., y_n)$: $\prod_{i=1}^{n} f(y_i; \theta)$

- **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** $I(\theta)$: $-\mathrm{E}_\theta \left(\frac{\partial s(\theta; Y)}{\partial \theta} \right)$

- **beobachtete Fisher-Information** $I_{obs}(\theta)$:
  $-\mathrm{E}_\theta \left(\frac{\partial s(\theta; y)}{\partial \theta} \right)$

Eigenschaften der Score-Funktion

erste Bartlett-Gleichung:

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

Proof:
$$1 = \int f(y; \theta) dy$$
$$0 = \frac{\partial 1}{\partial \theta} = \int \frac{\partial f(y; \theta)}{\partial \theta} dy = \int \frac{\partial f(y; \theta)/\partial \theta}{f(y; \theta)} f(y; \theta) dy$$
$$= \int \frac{\partial}{\partial \theta} \log f(y; \theta) f(y; \theta) dy = \int s(\theta; y) f(y; \theta) dy$$

zweite Bartlett-Gleichung:

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

Proof:
$$0 = \frac{\partial 0}{\partial \theta} = \frac{\partial}{\partial \theta} \int \frac{\partial}{\partial \theta} \log f(y; \theta) f(y; \theta) dy \qquad \text{siehe oben}$$
$$= \int \left(\frac{\partial^2}{\partial \theta^2} \log f(y; \theta) \right) f(y; \theta) dy$$
$$+ \int \frac{\partial \log f(y; \theta)}{\partial \theta} \frac{\partial f(y; \theta)}{\partial \theta} dy$$
$$= \mathrm{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \log f(Y; \theta) \right)$$
$$+ \int \frac{\partial \log f(y; \theta)}{\partial \theta} \frac{\partial \log f(y; \theta)}{\partial \theta} f(y; \theta) dy$$
$$\Leftrightarrow \mathrm{E}_\theta \left(s(\theta; Y) s(\theta; Y) \right) = \mathrm{E}_\theta \left(-\frac{\partial^2}{\partial \theta^2} \log f(Y; \theta) \right)$$
Bartletts zweite Gleichung gilt dann, weil $\mathrm{E}\left(s(\theta; Y)\right) = 0$

ML-Schätzer

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

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

$$\hat{\theta} \overset{a}{\sim} \mathrm{N}\left(\theta, I^{-1}(\theta)\right)$$

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

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

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:
$$0 = s(\hat{\theta}_{ML}; y) \overset{Taylor}{\underset{Series}{\approx}} s(\theta; y) + \frac{\partial s(\theta; y)}{\partial \theta}(\hat{\theta}_{ML} - \theta) \Leftrightarrow$$
$$\hat{\theta}_{ML} \approx \theta - \left(\frac{\partial s(\theta; y)}{\partial \theta} \right)^{-1} s(\theta; y) \approx \theta - I^{-1}(\theta) s(\theta; y)$$
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 δ 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 $\theta_{(t+1)} := \theta_{(t)} + \hat{\mathrm{Var}}_{\theta_{(t)}}(t(Y))^{-1} \mathrm{E}_{\theta_{(t)}}(t(Y))$ 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 $2 \cdot lr(\theta, \hat{\theta}) \overset{a}{\sim} \chi_1^2$

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

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

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

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

3.4 Sufficiency und Consistency

Statistic

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

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

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

Suffizienz

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

Minimalsuffizienz:

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

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

Neyman-Kriterium:

$t(Y_1, ..., Y_n)$ suffizient $\Leftrightarrow f(y; \theta) = h(y)g(t(y); \theta)$

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

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)	α (type I error)
H_1	β (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

$$p - value \leq \alpha \Leftrightarrow \text{``}H_0\text{''}$$

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

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

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

Test for μ, unknown σ^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 $\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n}(Y_i - \bar{Y})^2}$

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

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

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

As $\hat{\theta}$ converges to θ_0 under H_0, it can also be used to calculate the variance: $I^{-1}(\hat{\theta}$.

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

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

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

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

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

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 α, the Neyman Pearson Test is the most powerful test for comparing two estimates for θ.

> Proof:
>
> Decision rule of the NP-Test: $\varphi^* = \begin{cases} 1 & if \frac{f(y;\theta_0)}{f(y;\theta_1)} \leq e^c \\ 0 & \text{otherwise} \end{cases}$
>
> Need to show: $P(\varphi(Y)=1|\theta_1) \leq P(\varphi^*(Y)=1|\theta_1) \, \forall \varphi$
>
> $P(\varphi^*=1|\theta_1) - P(\varphi=1|\theta_1) =$
>
> $= \int \{\varphi^*(y) - \varphi(y)\} f(y;\theta_1) dy$
>
> $\geq \frac{1}{e^c} \int_{\varphi^*=1} \{\varphi^*(y) - \varphi(y)\} f(y;\theta_0) dy \quad f(y;\theta_1) \geq \frac{f(y;\theta_0)}{e^c}$
>
> $+ \frac{1}{e^c} \int_{\varphi^*=0} \{\varphi^*(y) - \varphi(y)\} f(y;\theta_0) dy \quad f(y;\theta_1) \leq \frac{f(y;\theta_0)}{e^c}$
>
> $= \frac{1}{e^c} \int \{\varphi^*(y) - \varphi(y)\} f(y;\theta_0) dy = 0$
>
> 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)
$H_0: X_i \sim F_0 \quad vs. \quad H_1: X_i \sim F \neq F_0$

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

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 χ_{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)
$H_0: X_i \sim F_0 \quad vs. \quad H_1: X_i \sim F \neq F_0$

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

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

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

$$g(Y;\theta) \text{ pivotal}$$

$$\Leftrightarrow$$

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

Approximative Pivotal Statistic
$H_0: X_i \sim F$ pivotal $\quad vs. \quad H_1: X_i \sim F$ not pivotal

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

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

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

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

4.5 Multiple Tests

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

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

$$FDR = \text{E}\left(\frac{\#\text{"}H1\text{"}|H_0}{\#\text{"}H1\text{"}} \right)$$
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 $\text{E}(e|x) = 0$

- **Constant Variation** $Var(Y|x) = \sigma^2$

For the normal linear model:

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

Attributes of the Regression Line
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x})$$
$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$
$$= y_i - (\bar{y} + \hat{\beta}_1(x_i - \bar{x}))$$

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

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

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

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

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

The β-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
$$Y = X\hat{\beta} + e$$
$$\hat{Y} = X\hat{\beta}$$

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

Assumptions

- **Independent Observations** $y_1, ... y_n$ 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) $Y|x_{1:p} \sim N(X\beta, \sigma^2)$

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

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

The estimates are the same as for the OLS procedure.

β is the **B**est **L**inear **U**nbiased **E**stimator

> Proof:
> 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 σ^2 is biased.

> Proof:
> $H := X(X^T X)^{-1} X^T$ hat matrix; $HH = H = H^T$ (idempotent)
> $$E((Y - X\hat{\beta})^T(Y - X\hat{\beta})) = E((Y^T(I_n - H)^T((I_n - H)Y)$$
> $$= E(tr(Y^T(I_n - H)Y)$$
> $$= E(tr((I_n - H)YY^T)$$
> $$= tr((I_n - H)E(YY^T))$$
> $$= tr((I_n - H)E(X\beta\beta^T X^T + \sigma I_n))$$
> $$= \sigma^2 tr((I_n - H))$$
> $$= \sigma^2(n - p)$$

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

with s an unbiased estimator

5.2 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

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

or more general:

$$f_{post}(\theta|X) = \frac{f(X|\theta) \cdot f_\theta(\theta)}{\int f(X|\tilde\theta)f_\theta(\tilde\theta)d\tilde\theta}$$

$$= C \cdot f(X|\theta) \cdot f_\theta(\theta) \quad \text{choose C so that } \int f(\theta|X) = 1$$

$$\propto f(X|\theta) \cdot f_\theta(\theta)$$

Point Estimates

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

$$\hat\theta_{postmode} = \underset{\vartheta}{\arg\max} f_\theta(\vartheta, x)$$

$$\hat\theta_{Bayesrisk} = \underset{t(.)}{\arg\min} R_{Bayes}(t(.))$$

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

$$\hat\theta_{postBayesrisk} = \underset{t(.)}{\arg\min} R_{postBayes}(t(.)|y)$$

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

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

$f_\theta(\theta) = const.$ 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 $E(KL(f_\theta(.), f_{post}(., x))$

Empirical Bayes
Let the prior depend on a hyper-parameter: $f_\theta(\theta, \gamma)$
Choose γ 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
$\pi \sim \text{Be}(\alpha, \beta)$	$\text{Bin}(n, \pi)$	$\text{Be}(\alpha+k, \beta+n-k)$
$\mu \sim \text{N}(\gamma, \tau^2)$	$\text{N}(\mu, \sigma^2)$	$\text{N}(.,.) \overset{n\to\infty}{\longrightarrow} \text{N}(\bar y, \frac{\sigma^2}{n})$
$\sigma^2 \sim \text{IG}(\alpha, \beta)$	$\text{N}(\mu, \sigma^2)$	$\text{IG}(\alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum_{i=1}^n(y_i-\mu)^2)$
$\lambda \sim \text{Ga}(\alpha, \beta)$	$\text{Po}(\lambda)$	$\text{Ga}(\alpha+n\bar y, \beta+n)$

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

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

with the one-dimenional $J_P := -\frac{\partial^2 l_{(n)}(\theta,y)}{\partial\theta^2} - \frac{\partial^2 \log f\theta(\theta)}{\partial\theta^2}$ 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: $\int e^{\left\{\sum_{i=1}^n \log f(y_i|\theta)+\log f_\theta(\theta)\right\}}d\theta =$
> $\int e^{\{l(\theta;y)+\log f_\theta(\theta)\}}d\theta \overset{TS}{\approx} \int e^{(l_P(\hat{\theta}_P)-\frac{1}{2}J_P(\hat{\theta}_P)(\vartheta-\hat{\theta}_P)^2)}d\vartheta$
> 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 α 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 θ^* from proposal distribution $q\left(\theta|\theta_{(t)}^*\right)$

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

This sequence has a stationary distribution for $n \to \infty$.
Choice of q: trade-off between exploring Θ and reaching a high α.
Burn-in and thinning out give $i.i.d.$ samples from $f_{post}(\theta|X)$.

Gibbs Sampling: For high dimensions α is close to zero.
Sample from the marginal distributions seperately:
$$\theta_i^* \sim f_{\theta_i|y,\theta\setminus\theta_i}\left(\theta_i^*|y,\theta_{t^*,i}\right)$$
with $\theta_{t^*,i}$ most recent estimates without θ_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_θ to independence: $q_\theta(\theta) = \prod_{k=1}^p q_k(\theta_k)$
Update each component iteratively. Works well for big p.