Statistics for Computer Sciences

Lecture 04 to Lecture 09
Probabilistic and Statistical Models

Stanislav Katina[1]

[1]Institute of Mathematics and Statistics, Masaryk University
Honorary Research Fellow, The University of Glasgow

December 1, 2015

Probabilistic and Statistical Models

Random variable, random vector, data, individuals

- **random variable** and **random vector**
  - *random variable* X is a function from a **sample space** to a set of real numbers $X : \mathcal{Y} \to \mathbb{R}$ (a set of all possible outcomes)
  - 2-dimensional *random vector* $(X_1, X_2)^T : \mathcal{Y} \to \mathbb{R}^2$
  - k-dimensional *random vector* $(X_1, X_2, \ldots, X_k)^T : \mathcal{Y} \to \mathbb{R}^k$
- **data** – *data vector* and *data matrix* – the elements of a vector and the rows of a matrix are measured on ***individuals*** (***statistical units***)
  - *data* as realisations of X – n-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, where n is a *sample size*
  - *data* as realisations of $(X_1, X_2)^T$ – $(n \times 2)$-dimensional matrix with rows $(x_{i1}, x_{i2})^T$, $i = 1, 2, \ldots, n$ and columns $\mathbf{x}_1$ and $\mathbf{x}_2$
  - *data* as realisations of $(X_1, X_2, \ldots, X_k)^T$ – $(n \times k)$-dimensional matrix with rows $(x_{i1}, x_{i2}, \ldots, x_{ik})^T$, $i = 1, 2, \ldots, n$ and columns $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_k$

Probabilistic and Statistical Models

Model

- based on **probabilistic sampling principles**, the *individuals* are sampled from a *population*
- **attribute** – a specific value of a variable
- with certain precision, data are **measured** on individuals
- **descriptive statistics** – describing and summarising data
- **inferential statistics** (**statistical inference**) – inferring (drawing conclusions) about random variable ***based on a model fitted to data***
- $\mathcal{F}$ is a **set of models** (probabilistic or statistical)
  - X is characterised by a model $F(\cdot)$, $F \in \mathcal{F}$
  - $(X_1, X_2)^T$ is characterised by a model $F^{(2)}(\cdot)$, $F \in \mathcal{F}$
  - $(X_1, X_2, \ldots, X_k)^T$ is characterised by a model $F^{(k)}(\cdot)$, $F \in \mathcal{F}$
- **parameter** – a numerical quantity that characterises a model – one-dimensional parameter θ, k-dimensional vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^T$

Probabilistic and Statistical Models

Distribution function, probability and density function

- useful assumption – $X_i, i = 1, 2, \ldots, n$, are **independently identically distributed** random variables
- **distribution function**
  - discrete random variable

$$F_X(x) = \Pr(X \leq x) = \sum_{i : x_i \leq x} \Pr(X = x_i),$$

where $\sum_{i=1}^{k(\infty)} p_i = 1$, $\Pr(X = x_i) = p_i = f_X(x_i) = f(x_i), \forall x_i$, where p_i is **probability mass function**; $\{x_i, p_i\}_{i=1}^{k(\infty)}$, $k \in \mathbb{N}^+$

  - continuous random variable

$$F_X(x) = \int_{-\infty}^{x} f(t)\, dt, f(x) \geq 0,$$

where $\int_{-\infty}^{\infty} f(x)\, dx = 1$, $f_X(x) = f(x) = \frac{\partial}{\partial x} F_X(x)$ is **density function**

Probabilistic and Statistical Models

Parametric and non-parametric model

- Θ is a **parametric space**, the **support** of $F(\cdot;\boldsymbol{\theta})$ is $\mathcal{Y}_{\boldsymbol{\theta}} \subseteq \mathbb{R}^n$ (the smallest set, where the distribution function is defined); sample space $\mathcal{Y} = \cup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{Y}_{\boldsymbol{\theta}}$
- $\mathcal{F}$ as **a parametric set of distribution functions**

$$\mathcal{F} = \left\{ F(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k \right\},$$

- $\mathcal{F}$ as **a parametric set of probability or density functions**

$$\mathcal{F} = \left\{ f(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k \right\}$$

- $\mathcal{F}$ as **non-parametric set**

$$\mathcal{F} = \{\text{a set of all density functions}\},$$

alternatively, probability or distribution function can be used

Probabilistic and Statistical Models

Reading of mathematical notation

- the term "**probability model**" is often reduced to "**distribution**"
- "Random variable X is distributed as $F(x)$" or "random variable X is characterised by distribution $F(x)$", notation $X \sim F_X(x)$; symbol "$\sim$" means "asymptotically", "for sufficiently large n" (notation $X \sim f_X(x)$ is used very rarely)
- "Random variable X is distributed as random variable Y" or "Random variable X and Y are identically distributed" (notation $X \sim Y$ or $F_X(x) \sim F_Y(y)$
- the term "**statistical model**" is often reduced to "**model**" (usually referred as ***causal statistical model*** or ***model of causal dependence***)
- "Y depends on X", where X is **independent variable** and Y is **dependent variable** (notation $Y|X$)

Probabilistic and Statistical Models

Reading of mathematical notation

- "X is normally distributed with parameters μ and σ^2", notation $X \sim N(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2)^T$
- "$\mathbf{X} = (X_1, X_2)^T$ is characterised by bivariate normal distribution with parameters μ_1, μ_2, σ_1^2, σ_2^2 and ρ", notation $X \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$
- "$\mathbf{X} = (X_1, X_2, \ldots, , X_k)^T$ is characterised by multivariate normal distribution with parameters μ_1, μ_2, …, μ_k, σ_1^2, σ_2^2, …, σ_k^2, and $\rho_{1,2}$, …, $\rho_{k-1,k}$, ", notation $X \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta} = (\mu_1, \mu_2, \ldots, \mu_k, \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2, \rho_{1,2}, \ldots, \rho_{k-1,k})^T$
- "X is binomially distributed with parameter p", notation $X \sim Bin(N, p)$, where $\theta = p$
- "X is characterised by distribution with parameter λ", notation $X \sim Poiss(\lambda)$, where $\theta = \lambda$
- "$\mathbf{X} = (X_1, X_2, \ldots, , X_k)^T$ is multinomially distributed with parameter $\mathbf{p}$", notation $\mathbf{X} \sim Mult_k(N, \mathbf{p})$, where $\boldsymbol{\theta} = \mathbf{p}$

Probabilistic and Statistical Models

Measures of normal distribution

- "X is normally distributed with parameters μ and σ^2", notation $X \sim N(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2)^T$
- Random variable Z (Z-transformation) $\Pr(Z = \frac{X-\mu}{\sigma} < x_{1-\alpha}) = 1 - \alpha, Z \sim N(0, 1)$
- Rule "$90 - 95 - 99$" $\Pr(a \le X \le b) = 1 - \alpha$, where $1 - \alpha = 0.90, 0.95$ and 0.99, $a = \mu - x_{1-\alpha/2}\sigma$ and $b = \mu + x_{1-\alpha/2}\sigma$
- Rule "$68.27 - 95.45 - 99.73$" $\Pr(a \le X < b) = \Pr(X < b) - \Pr(X < a) = F_X(b) - F_X(a)$, where $a = \mu - k\sigma$, $b = \mu + k\sigma$, $k = 1, 2$ and 3

Probabilistic and Statistical Models
Approximation of binomial distribution by normal distribution

Definition (approximation of binomial distribution by normal distribution)

If random variable X is binomially distributed with parameter p, $X \sim Bin(N,p)$, where $\theta = p$, then if $Np > 5$ a $Nq > 5$, where $q = 1 - p$, then the distribution of random variable X can be approximated by normal distribution, $X \sim N(Np, Npq)$, where $\boldsymbol{\theta} = (Np, Npq)^T$.

Table: Examples of minimal N for fixed p

p	0.1	0.2	0.3	0.4	0.5
q	0.9	0.8	0.7	0.6	0.5
N	51	26	17	13	11

Probabilistic and Statistical Models
Approximation of binomial distribution by normal distribution

Example

Let Pr(male) $= 0.515$ and Pr(female) $= 0.485$. Let X if the frequency of males and Y frequency of females. Assuming that $X \sim Bin(N,p)$, (a) $\Pr(X \leq 3)$, if $N = 5$, (b) $\Pr(X \leq 5)$, if $N = 10$ and (c) $\Pr(X \leq 25)$, if $N = 50$. Compare the results with normal approximation $X \sim N(Np, Npq)$.

Solution

(a) $E[X] = Np = 5 \times 0.515 = 2.575, E[Y] = 5 \times 0.485 = 2.425$,
$\Pr(X \leq 3) = \sum_{k \leq 3} \binom{5}{k} 0.515^k 0.485^{5-k} = 0.793$,
$\Pr(X \leq 3) = 0.648, N(5 \times 0.515, 5 \times 0.515 \times 0.485)$.
(b) $E[X] = 10 \times 0.515 = 5.15, E[Y] = 10 \times 0.485 = 4.85$,
$\Pr(X \leq 5) = \sum_{k \leq 5} \binom{10}{k} 0.515^k 0.485^{10-k} = 0.586$,
$\Pr(X \leq 5) = 0.462, N(10 \times 0.515, 10 \times 0.515 \times 0.485)$.
(c) $E[X] = 50 \times 0.515 = 25.75, E[Y] = 50 \times 0.485 = 24.25$,
$\Pr(X \leq 25) = \sum_{k \leq 25} \binom{50}{k} 0.515^k 0.485^{50-k} = 0.471$,
$\Pr(X \leq 25) = 0.416, N(50 \times 0.515, 50 \times 0.515 \times 0.485)$.

Probabilistic and Statistical Models
Approximation of binomial distribution by normal distribution

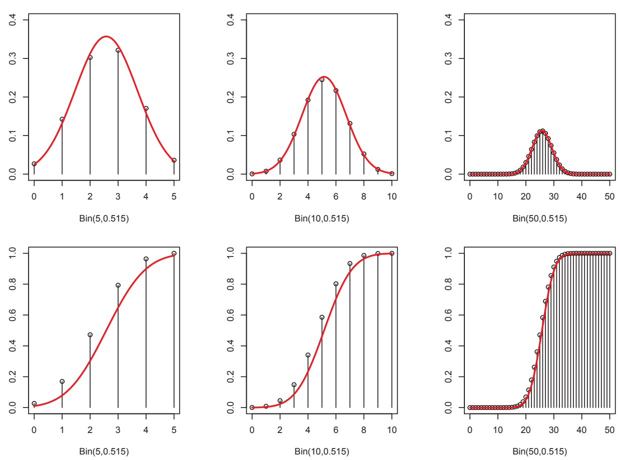

Figure: Probability function of binomial distribution superimposed by the density function of normal distribution ($p = 0.515$; $N = 5, 10$ and 50)

Probabilistic and Statistical Models
Approximation of binomial distribution by normal distribution

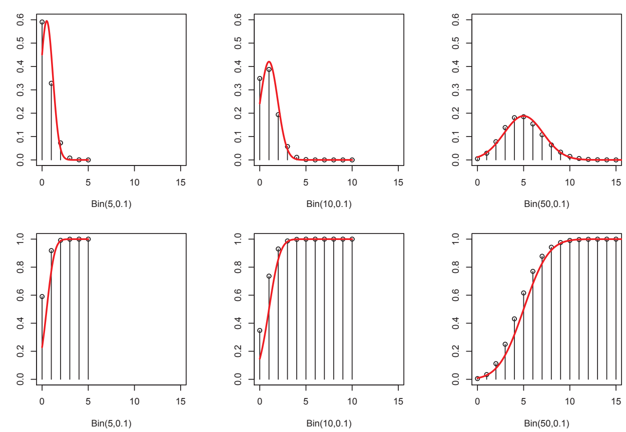

Figure: Distribution function of binomial distribution superimposed by the distribution function of normal distribution ($p = 0.515$; $N = 5, 10$ and 50)

Definition (normal distribution)

Random variable is **normally distributed** with parameters μ and σ, i.e. $X \sim N(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ and density is defined as $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}, \sigma > 0$.

Definition (standardised normal distribution)

Random variable is **normally distributed** with parameters $\mu = 0$ and $\sigma = 1$, i.e. $X \sim N(0, 1)$, where $\boldsymbol{\theta} = (0, 1)^T$ and density is defined as $\phi(x) = f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in \mathbb{R}, \sigma > 0$.

Parameter μ is called **mean** of X and σ^2 the **variance** of X.

Definition (bivariate normal distribution)

Random vector $(X, Y)^T$ is **normally distributed** with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, i.e. $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (\mu_1, \mu_2)^T \text{ a } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

$\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$, $(x, y)^T \in \mathbb{R}^2$, $\mu_j \in \mathbb{R}^1$, $\sigma_j^2 > 0, j = 1, 2$, $\rho \in \langle -1, 1 \rangle$; density is defined as

$$f(x, y) = \frac{1}{A} \exp\left\{ -\frac{1}{B} \left\{ \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right\} \right\},$$

where $A = 2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}$, $B = 2(1-\rho^2)$.

Definition (bivariate standardised normal distribution)

Random vector $(X, Y)^T$ is **normally distributed** with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, i.e. $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (0, 0)^T \text{ a } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

$\boldsymbol{\theta} = (0, 0, 1, 1, \rho)^T$, $(x, y)^T \in \mathbb{R}^2$, $\rho \in \langle -1, 1 \rangle$; density is defined as

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{ -\frac{x^2 + 2\rho xy + y^2}{2(1-\rho^2)} \right\}.$$

Let $x = x_1$, $y = x_2$ and $\mathbf{x} = (x_1, x_2)^T$. Then the density can be rewritten into matrix form:

$$f(\mathbf{x}) = \frac{1}{2\pi(\det(\boldsymbol{\Sigma}))^{1/2}} \exp\left\{ -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} \right\}.$$

Let $(X_1, X_2, \ldots, X_k)^T \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{x}$ is k-dimensional vector, then the density is equal to

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}(\det(\boldsymbol{\Sigma}))^{1/2}} \exp\left\{ -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} \right\}.$$

Marginal distributions of:

- bivariate normal distribution – $X_j \sim N(\mu_j, \sigma_j^2), j = 1, 2, \ldots, k$
- standardised bivariate normal distribution – $X_j \sim N(0, 1), j = 1, 2, \ldots, k$

Probabilistic and Statistical Models

Bivariate normal distribution – simulation

Simulation of pseudo-random numbers from bivariate normal distribution:

1. let $X_1 \sim N(0,1)$ and $X_2 \sim N(0,1)$
2. then $(Y_1, Y_2)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $Y_1 = \sigma_1 X_1 + \mu_1$ and $Y_2 = \sigma_2(\rho X_1 + \sqrt{1-\rho^2} X_2) + \mu_2$

Example

Sumulate pseudo-random numbers from bivariate normal distribution, where $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$.
(a) $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1, \rho = 0$; (1) $n = 50$ and (2) $n = 1000$;
(b) $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1, \rho = 0.5$; (1) $n = 50$ and (2) $n = 1000$;
(c) $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1.2, \rho = 0.5$; (1) $n = 50$ and (2) $n = 1000$.

Probabilistic and Statistical Models

Mixture of two bivariate normal distribution

The mixture of two univariate normal distribution is defined as follows: $pN(\mu_1, \sigma_1^2) + pN(\mu_2, \sigma_2^2)$, where $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^T$

The mixture of two bivariate normal distribution is defined as follows: $pN_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-p)N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\theta} = (\mu_{11}, \mu_{12}, \sigma_{11}^2, \sigma_{12}^2, \rho_1, \mu_{21}, \mu_{22}, \sigma_{21}^2, \sigma_{22}^2, \rho_2)^T$

Probabilistic and Statistical Models

Binomial distribution

Jacob Bernoulli (1655–1705) – one of the founding fathers of probability theory.

Definition (binomial distribution)

Let N be number of independent identical (random) *Bernoulli trials* X_i, where $X_i = 1$ is a **success** (event occurred) and $X_i = 0$ is a **failure** (event did not occur), $i = 1, 2, \ldots, N$. Then **probability of success** $\Pr(X_i = 1) = p$ and **probability of failure** $\Pr(X_i = 0) = 1 - p$. Number of successes $X = \sum_{i=1}^{N} X_i$.
The probability that random variable X is equal to $x = n$ (realisation) is defined as $\Pr(X = x) = \binom{N}{x} p^x (1-p)^{N-x}$, for $x = 0, 1, 2, \ldots, N$.
Expected value of X is defined as $E[X] = \sum_{x=0}^{N} x \Pr(X = x) = \sum_{x=0}^{N} x \binom{N}{x} p^x (1-p)^{N-x} = Np$.
Variance of X is defined as $Var[X] = \sum_{x=0}^{N} (x - E[X])^2 \Pr(X = x) = \sum_{x=0}^{N} (x - Np)^2 \binom{N}{x} p^x (1-p)^{N-x} = Np(1-p)$.

Probabilistic and Statistical Models

Binomial distribution

Reading: Random variable X is binomially distributed with parameters N an p, where $\theta = p$.
Notation: $X \sim Bin(N, p), \theta = p$
Do we need to change it? YES.
Why? Due to generalisation.

Equivalently, $\mathbf{X} \sim Bin(N, p, 1-p)$, where $\mathbf{X} = (X_1, X_2)^T$, $\boldsymbol{\theta} = (p, 1-p)^T$, X_1 is **number of successes**, $X_2 = N - X_1$ is **number of failures**, $X_1 \sim Bin(N, p)$ and $X_2 \sim Bin(N, 1-p)$.
Then d

- $E[X_1] = Np$, $E[X_2] = N(1-p)$,
- $Var[X_2] = Np(1-p) = Var[X_1]$ is independent of p,
- $Cov[X_1, X_2] = -Np(1-p)$ and
- $Cor[X_1, X_2] = -1$.

Finally, $\mathbf{n} = (n_1, n_2)^T$ a $\mathbf{p} = (p_1, p_2)^T$, $p_1 = p$ and $p_2 = 1 - p$.
Then $\boldsymbol{\theta} = \mathbf{p}$.

Probabilistic and Statistical Models

Binomial distribution

Example (number of boys)

Number of boys X in families with N children is binomially distributed, i.e. $X \sim Bin(N, p)$, where $N = 12$, number of families $M = 6115$ (Geissler 1889). Calculate theoretical frequencies $m_{n,E}$. You know that $p = \frac{\sum_{n=0}^{N} nm_n}{NM} = 0.5192$ (weighted average; average of number of families weighted by number of boys).

Table: Observed and theoretical frequencies ($m_{n,O}$ and $m_{n,E}$) of families with n boys (O = observed, E = expected, theoretical)

n	0	1	2	3	4	5	6	7	8	9	10	11	12
$m_{n,O}$	3	24	104	286	670	1033	1343	1112	829	478	181	45	7
$m_{n,E}$	1	12	72	258	628	1085	1367	1266	854	410	133	26	2

Probabilistic and Statistical Models

Multinomial distribution

Definition (multinomial distribution)

Let N be number of independent identical (random) trials and in each of them $J \geq 2$ distinct possible outcomes can occur, where $X_{ji} = 1$ is a **success** (event occurred) and $X_{ji} = 0$ is a **failure** (event did not occur), $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, J$. Number of successes $X_j = \sum_{i=1}^{N} X_{ji}$, $N = \sum_{j=1}^{J} X_j$. Then **probability of success** of i-th outcome in j-th trial is equal to $\Pr(X_{ji} = 1) = p_j$ (**cell probabilities**) and **probability of failure** in j-th trial is equal to $\Pr(X_{ji} = 0) = 1 - p_j$. Let $\mathbf{X} = (X_1, X_2, \ldots, X_J)^T$. The probability that random variables X_j are equal to $x_j = n_j$ is defined as

$$\Pr(X_1 = x_1, \ldots, X_J = x_J) = \frac{N!}{x_1! \ldots x_J!} p_1^{x_1} p_2^{x_2} \ldots p_J^{x_J} = \frac{N!}{\prod_j x_j!} \prod_{j=1}^{J} p_j^{x_j}.$$

Probabilistic and Statistical Models

Multinomial distribution

Expected value of $\mathbf{X}$ is a vector defined as $E[\mathbf{X}] = N\mathbf{p}$.
Covariance matrix of $\mathbf{X}$ is defined as

$$Var[\mathbf{X}] = N\left(\operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T\right),$$

where

$$(Var[\mathbf{X}])_{ij} = \begin{cases} Np_j(1 - p_j)) & \text{if } i = j \\ -Np_i p_j & \text{if } i \neq j \end{cases}$$

Marginal distributions are binomial, i.e. $X_j \sim Bin(N, p_j)$. Then

- $E[X_j] = Np_j$,
- $Var[X_j] = Np_j(1 - p_j)$
- $Cov[X_i, X_j] = -Np_i p_j$
- $Cor[X_i, X_j] = (-p_i p_j) / \sqrt{p_i(1 - p_i) p_j (1 - p_j)}$

Probabilistic and Statistical Models

Multinomial distribution

Reading: Random vector $\mathbf{X}$ is multinomially distributed with parameters N and $\mathbf{p}$, where $\theta = \mathbf{p}$.
Notation: $\mathbf{X} \sim Mult_J(N, \mathbf{p})$.
If $J = 2$, then $Bin(N, p) \approx Mult_2(N, \mathbf{p})$
Realisation of one trial $\mathbf{x}_{ij}$ could be $(1, 0, \ldots, 0)^T$ or $(0, 1, \ldots, 0)^T$.

Example (number of individuals with certain blood type)

Number of individuals $\mathbf{X} = (X_1, X_2, X_3, X_4)^T$ with certain blood group is multinomially distributed following Hardy-Wienberg equilibrium, i.e. $\mathbf{X} = (X_1, X_2, X_3, X_4)^T \sim Mult_4(N, \mathbf{p})$, where $N = 500$ (Katina et al. 2015). Calculate theoretical frequencies $n_{E,j}$.

attributes (groups)	0	A	B	AB
$n_{O,j}$	209	184	81	26
$n_{E,j}$	210	183	80	27

Example (number of individuals with certain socioeconomic status, political philosophy and political affiliation)

Number of individuals $X_1, \ldots, X_8$ with socioeconomic status, political philosophy and political affiliation is multinomially distributed, i.e. $\mathbf{X} = (X_1, \ldots, X_8)^T \sim Mult_8(N, \mathbf{p})$, where $\mathbf{p} = (p_1, p_2, \ldots, p_8)^T$ and $N = 50$ (Christensen 1990). Calculate (a) $Var[X_1]$, (b) $Var[X_3]$, (c) $Cov\,[X_1, X_3]$ and (d) $Corr\,[X_1, X_3]$.

Table: 2×4 contingency table of probabilities p_j

	D-Li	D-C	R-Li	R-C	total
H	0.12	0.12	0.04	0.12	0.4
Lo	0.18	0.18	0.06	0.18	0.6
total	0.30	0.30	0.10	0.30	1.0

Notation: (1) socioeconomic status (high – H, low – Lo), (2) political philosophy (democrat – D, republican – R) a (3) political affiliation (liberal – Li, conservative – C). Then X_1 (H-D-Li), X_2 (H-D-C), X_3 (H-R-Li), X_4 (H-R-C), X_5 (Lo-D-Li), X_6 (Lo-D-C), X_7 (Lo-R-Li) a X_8 (Lo-R-C).

Solution:

$Var[X_1] = 50 \times 0.12 \times (1 - 0.12) = 5.28$
$Var[X_3] = 50 \times 0.04 \times (1 - 0.04) = 1.92$
$Cov\,[X_1, X_3] = -50 \times 0.12 \times 0.04 = -0.24$
$Cor\,[X_1, X_3] = -0.24/\sqrt{5.28 \times 1.92} = -0.075$

What are the expected frequencies?

Table: 2×4 contingency table of frequencies X_j

	D-Li	D-C	R-Li	R-C
H	6	6	2	6
Lo	9	9	3	9

Example (number of individuals with certain eye and hair colour)

Let $\mathbf{X} = (X_1, X_2, \ldots, X_{12})^T$ be random vector of number of individuals with certain eye colour (with levels blue Bl, green Gr, brown Br) and hair color (with levels blond Blo, light-brown LB, black Ble, red R), where X_1 means Bl-Blo, X_2 means Bl-LB, X_3 means Bl-Ble, X_4 means Bl-R, X_5 means Gr-Blo, X_6 means Gr-LB, X_7 means Gr-Ble, X_8 means Gr-R, X_9 means Br-Blo, X_{10} means Br-LB, X_{11} means Br-Ble and X_{12} means Br-R. Let $\mathbf{X} \sim Mult_{12}(N, \mathbf{p})$, where $N = 6800$ (Yule and Kendall 1950).

Table: 3×4 contingency table of frequencies n_j

	Blo	LB	Ble	R	row sums
Bl	1768	807	189	47	2811
Gr	946	1387	746	53	3132
Br	115	438	288	16	857
column sums	2829	2632	1223	116	6800

Definition (product-multinomial distribution)

Let N_k be number of independent identical (random) trials and in each of them $J \geq 2$ distinct possible outcomes can occur, where $X_{kji} = 1$ is a **success** (event occurred) and $X_{kji} = 0$ is a **failure** (event did not occur), $i = 1, 2, \ldots, N_k$, $k = 1, 2, \ldots, K$, $j = 1, 2, \ldots, J$. Number of successes $X_{kj} = \sum_{i=1}^{N_k} X_{kji}$ and $\sum_{k=1}^{K} N_k = N$. Then **probability of success** of kj-th outcome in i-th trial is equal to $\Pr(X_{kji} = 1) = p_{kj}$ (**cell probabilities**) and **probability of failure** of kj-th outcome in i-th trial is equal to $\Pr(X_{kji} = 0) = 1 - p_{kj}$. Let $\mathbf{X}_k = (X_{k1}, X_{k2}, \ldots, X_{kJ})^T$ si multinomially distributed with parameters N_k and $\mathbf{p}_k$, i.e. $\mathbf{X}_k \sim Mult_J\,(N_k, \mathbf{p}_k)$, kde $\boldsymbol{\theta}_k = \mathbf{p}_k$ a $\mathbf{p}_k = (p_{k1}, p_{k2}, \ldots, p_{kJ})^T$. Let realisations of $\mathbf{X}_k$ be $\mathbf{x}_k$. The $x_{kj} = n_{kj}$ and $\mathbf{n}_k = (n_{k1}, n_{k2}, \ldots, n_{kJ})^T$. Additionally, $\mathbf{X}_k$ are independent.

The probability that random variables X_{kj} are equal to $x_{kj} = n_{kj}$ (for all j and k) is defined as

$$\Pr(X_{kj} = x_{kj}, \forall k, j) = \prod_{k=1}^{K} \Pr(X_{kj} = x_{kj}, \forall j).$$

The probability that random variables X_{kj} are equal to $x_{kj} = n_{kj}$ (for all j) is defined as

$$\Pr(X_{kj} = x_{kj}, \forall j) = \left(N_k! / \prod_{j=1}^{J} x_{kj}! \right) \prod_{j=1}^{J} p_{kj}^{x_{kj}}.$$

Then

$$\Pr(X_{kj} = x_{kj}, \forall k, j) = \prod_{k=1}^{K} \left(\left(N_k! / \prod_{j=1}^{J} x_{kj}! \right) \prod_{j=1}^{J} p_{kj}^{x_{kj}} \right).$$

Reading: Random matrix $\mathbf{X}$ is product-multinomially distributed with parameters $\mathbf{N} = (N_1, N_2, \ldots, N_K)^T$ and $\mathbf{P}$ with the rows $\mathbf{p}_k$, where $\theta_k = \mathbf{p}_k, k = 1, 2, \ldots, K$.
Notation: $\mathbf{X} \sim ProdMult_K(\mathbf{N}, \mathbf{p})$.
If $K = 1$, then $Mult_J(N, \mathbf{p}) \approx ProdMult_1(N, \mathbf{p})$
Realisation of one trial $\mathbf{x}_{kij}$ could be $(1, 0, \ldots, 0)^T$ or $(0, 1, \ldots, 0)^T$.
Then

- **expected frequencies** are equal to $N_k p_{kj}$,
- within each $\mathbf{X}_k$, **variances** $Var[X_{kj}]$, **covariances** $Cov[X_{kj}]$ and **correlations** $Cor[X_{kj}]$ are calculated as for multinomial distribution,
- between $\mathbf{X}_k$, e.g. $Cov[\mathbf{X}_1, \mathbf{X}_2]$, $k = 1, 2$, are zeroes due to independence of $\mathbf{X}_k$

Example (number of individuals with certain blood type)

Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^T$, where $\mathbf{X}_1 = (X_{11}, X_{12}, X_{13}, X_{14})^T$ is number of individuals in Košice (Slovakia) with certain blood group, $\mathbf{X}_2 = (X_{21}, X_{22}, X_{23}, X_{24})^T$ is number of individuals in Prague (Czech Republic) with certain blood group. $\mathbf{X}$ is product-multinomially distributed, i.e. $\mathbf{X} \sim ProdMult_2(\mathbf{N}, \mathbf{P})$, where $\mathbf{N} = (N_1, N_2)^T$, where $N_1 = 500$ and $N_2 = 400$ (Katina et al. 2015). Calculate theoretical frequencies $n_{E,kj}$. **Question**: What are the probabilities of having particular blood group in Prague and Košice?

Table: Observed frequencies of particular blood group in Prague an Košice

attributes (groups)	0	A	B	AB
$n_{1j} = n_{\text{Prague},j}$	209	184	81	26
$n_{2j} = n_{\text{Košice},j}$	138	147	84	31

Example (number of individuals with certain socioeconomic status, political philosophy and political affiliation)

Number of individuals $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^T$ with socioeconomic status, political philosophy and political affiliation is product-multinomially distributed, i.e. $\mathbf{X} \sim ProdMult_2(\mathbf{N}, \mathbf{P})$, where $\mathbf{X}_1 = (X_{11}, X_{12}, X_{13}, X_{14})^T$ are number of individuals with high socioeconomic status, $\mathbf{X}_2 = (X_{21}, X_{22}, X_{23}, X_{24})^T$ number of individuals with low socioeconomic status, $\mathbf{p}_k = (p_{1|k}, p_{2|k}, \ldots, p_{J|k})^T$, $p_{kj} = p_{j|k} = \frac{n_{jk}}{n_k}$, $k = 1, 2$, $\mathbf{N} = (N_1, N_2)^T$, $N_1 = 30$, $N_2 = 20$ (Christensen 1990). Calculate (a) probabilities $p_{j|k}$, (b) expected frequencies, (c) $Var[X_{3|1}]$, (d) Cov and (e) $Cov\left[X_{1|1}, X_{3|2}\right]$.

Notation: (1) socioeconomic status (high – H, low – Lo), (2) political philosophy (democrat – D, republican – R) a (3) political affiliation (liberal – Li, conservative – C). Then $X_{11} = X_{1|1}$ (H-D-Li), $X_{12} = X_{2|1}$ (H-D-C), $X_{13} = X_{3|1}$ (H-R-Li), $X_{14} = X_{4|1}$ (H-R-C), $X_{21} = X_{1|2}$ (Lo-D-Li), $X_{22} = X_{2|2}$ (Lo-D-C), $X_{23} = X_{3|2}$ (Lo-R-Li) a $X_{24} = X_{4|2}$ (Lo-R-C).

Solution:

Table: 2×4 contingency table of probabilities $p_{j|k}$

	D-Li	D-C	R-Li	R-C	total
H	0.3	0.3	0.1	0.3	1.0
Lo	0.3	0.3	0.1	0.3	1.0

Table: 2×4 contingency table of frequencies n_{kj}

	D-Li	D-C	R-Li	R-C	total
H	9	9	3	9	30
Lo	6	6	2	6	20

$Var(X_{3|1}) = 30 \times 0.1 \times (1 - 0.1) = 2.7$.
$Cov\left[X_{1|2}, X_{3|2}\right] = -20 \times 0.3 \times 0.1 = -0.6$,
$Cov\left[X_{1|1}, X_{3|2}\right] = 0$, due to the independence of $\mathbf{X}_1$ and $\mathbf{X}_2$.

Notation: (1) socioeconomic status (high – H, low – Lo), (2) political philosophy (democrat – D, republican – R) a (3) political affiliation (liberal – Li, conservative – C). Then X_1 (H-D-Li), X_2 (H-D-C), X_3 (H-R-Li), X_4 (H-R-C), X_5 (Lo-D-Li), X_6 (Lo-D-C), X_7 (Lo-R-Li) a X_8 (Lo-R-C).

Solution:

$Var[X_1] = 50 \times 0.12 \times (1 - 0.12) = 5.28$
$Var[X_3] = 50 \times 0.04 \times (1 - 0.04) = 1.92$
$Cov\,[X_1, X_3] = -50 \times 0.12 \times 0.04 = -0.24$
$Cor\,[X_1, X_3] = -0.24/\sqrt{5.28 \times 1.92} = -0.075$

What are the expected frequencies?

Table: 2×4 contingency table of frequencies X_j

	D-Li	D-C	R-Li	R-C
H	6	6	2	6
Lo	9	9	3	9

Definition (Poisson distribution)

Let X be random variable characterised by Poisson distribution, i.e. $X(\lambda)$, where $\theta = \lambda$. Then

$$\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, \ldots,$$

where $x = n$ is realisation of X. Then $E[X] = \lambda$ and $Var[X] = \lambda$.

Example (Poisson distribution; killing by horse kicks)

Data were published by Russian economist *Ladislaus Bortkiewicz* in his book entitled *Das Gesetz der keinem Zahlen* (The Law of Small Numbers) in 1898. Let X be the number of corpses with certain number of solders killed by horse kicks in the Prussian army within one year (von Bortkiewicz 1898; in 10 different army corps; in 20 years, between 1875 and 1894), n be the number of annual deaths, m_n be the number of army corps with particular number of annual deaths, $M = \sum m_n = 10 \times 20 = 200$. Then $X \sim Poiss(\lambda)$, where $\lambda = \frac{\sum_n nm_n}{\sum_n m_n} = 0.61$ (weighted average; average of number of army corps weighted by number of annual deaths).

Table: Observed and theoretical frequencies ($m_{n,O}$ and $m_{n,E}$) of solders killed by horse with n annual deaths over 20 years

n	0	1	2	3	4	5+
$m_{n,O}$	109	65	22	3	1	0
$m_{n,E}$	109	66	20	4	1	0

Example (Poisson distribution; accidents in the factories)

Let X be the number of workers having an accident in the munition factories in England during First World War (Greenwood and Yule 1920), n be the number of accidents, m_n be the number of workers with particular number of accidents, $M = \sum m_n = 647$. Then $X \sim Poiss(\lambda)$, where $\lambda = \frac{\sum_n nm_n}{\sum_n m_n} = 0.47$ (weighted average; average of number of workers weighted by number of accidents).

Table: Observed and theoretical frequencies ($m_{n,O}$ and $m_{n,E}$) of workers with n accidents

n	0	1	2	3	4	≥ 5
$m_{n,O}$	447	132	42	21	3	2
$m_{n,E}$	406	189	44	7	1	0

1. multinomial distribution – example – **number of individuals with certain eye and hair color**: Are the rows and columns of a contingency table independent?
   - Are the frequencies of individuals with certain eye color (with levels blue, green, brown) independent of hair color (with levels blond, light-brown, black, red)?
2. product-multinomial distribution: Are the vectors of frequencies the same in each row? Are the vectors of frequencies independent of the row index?
   - example – **number of individuals with certain socioeconomic status, political philosophy and affiliation** – Are the vectors of frequencies of individuals (D-Li, D-C, R-Li, R-C) the same for each level of socioeconomic status (high and low)?
   - example – **blood groups** – Is the distribution of the blood groups (0, A, B, AB) the same in Prague and Košice?

3. binomial distribution – example – **number of boys**:
   - Is the probability of number of boys in the families with 12 boys binomial?
   - Is the probability of having a boy in the family equal to 0.5?
4. Poisson distribution:
   - example – **killing by horse kick** – Is the distribution of number of corpses with certain number of solders killed by horse kick Poisson?
   - example – **accidents in the factories** – Is the distribution of number of workers having an accident Poisson?

Probabilistic and Statistical Models

Types of contingency tables – multinomial distribution

$1 \times J$ **contingency table of frequencies**

outcome 1	outcome 2	...	outcome J	sum
x_1	x_2	...	x_J	N

$1 \times J$ **contingency table of probabilities**

outcome 1	outcome 2	...	outcome J	sum
p_1	p_2	...	p_J	1

$2 \times J$ **contingency table of frequencies**

	outcome 1	outcome 2	...	outcome J	sum
row 1	x_{11}	x_{12}	...	x_{1J}	N_1
row 2	x_{21}	x_{22}	...	x_{2J}	N_2

$2 \times J$ **contingency table of probabilities**

	outcome 1	outcome 2	...	outcome J	sum
row 1	p_{11}	p_{12}	...	p_{1J}	$p_{1\bullet} \neq 1$
row 2	p_{21}	p_{22}	...	p_{2J}	$p_{2\bullet} \neq 1$

Probabilistic and Statistical Models

Types of contingency tables – multinomial distribution

$K \times J$ **contingency table of frequencies**

	outcome 1	outcome 2	...	outcome J	sum
row 1	x_{11}	x_{12}	...	x_{1J}	N_1
row 2	x_{21}	x_{22}	...	x_{2J}	N_2
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
row K	x_{K1}	x_{K2}	...	x_{KJ}	N_K

$K \times J$ **contingency table of probabilities**

	outcome 1	outcome 2	...	outcome J	sum
row 1	p_{11}	p_{12}	...	p_{1J}	$p_{1\bullet} \neq 1$
row 2	p_{21}	p_{22}	...	p_{2J}	$p_{2\bullet} \neq 1$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
row K	p_{K1}	p_{K2}	...	p_{KJ}	$p_{K\bullet} \neq 1$

Probabilistic and Statistical Models

Types of contingency tables – product-multinomial distribution

$1 \times J$ **contingency table of frequencies** ($\approx$ multinomial distribution)

outcome 1	outcome 2	...	outcome J	sum
x_1	x_2	...	x_J	N

$1 \times J$ **contingency table of probabilities** ($\approx$ multinomial distribution)

outcome 1	outcome 2	...	outcome J	sum
p_1	p_2	...	p_J	1

$2 \times J$ **contingency table of frequencies** ($\approx$ multinomial distribution)

	outcome 1	outcome 2	...	outcome J	sum
group 1	x_{11}	x_{12}	...	x_{1J}	N_1
group 2	x_{21}	x_{22}	...	x_{2J}	N_2

$2 \times J$ **contingency table of probabilities**

	outcome 1	outcome 2	...	outcome J	sum
group 1	$p_{1\|1}$	$p_{2\|1}$	...	$p_{J\|1}$	1
group 2	$p_{1\|2}$	$p_{2\|2}$	...	$p_{J\|2}$	1

Probabilistic and Statistical Models

Types of contingency tables – product-multinomial distribution

$K \times J$ **contingency table of frequencies** ($\approx$ multinomial distribution)

	outcome 1	outcome 2	...	outcome J	sum
group 1	x_{11}	x_{12}	...	x_{1J}	N_1
group 2	x_{21}	x_{22}	...	x_{2J}	N_2
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
group K	x_{K1}	x_{K2}	...	x_{KJ}	N_K

$K \times J$ **contingency table of probabilities**

	outcome 1	outcome 2	...	outcome J	sum
group 1	$p_{1\|1}$	$p_{2\|1}$	...	$p_{J\|1}$	1
group 2	$p_{1\|2}$	$p_{2\|2}$	...	$p_{J\|2}$	1
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
group K	$p_{1\|K}$	$p_{2\|K}$	...	$p_{J\|K}$	1

Probabilistic and Statistical Models

Data structure for $1 \times J$ contingency table – multinomial distribution

	outcome 1	outcome 2	...	outcome J	sum
$\mathbf{x}_1$	1	0	...	0	1
$\mathbf{x}_2$	0	1		0	1
$\mathbf{x}_3$	0	1		0	1
$\mathbf{x}_4$	1	0	...	0	1
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
$\mathbf{x}_{N-1}$	0	0	...	1	1
$\mathbf{x}_N$	1	0	...	0	1
sum$=\mathbf{x}$	x_1	x_2	...	x_J	N

- sum of each row is one
- sum of all row sums is N
- sum of each column is x_j, where $j = 1, 2, \ldots, J$
- sum of all x_j, $j = 1, 2, \ldots, J$, is N
- $\mathbf{x} = \mathbf{n}$

Probabilistic and Statistical Models

Data structure for $K \times J$ contingency table – (product-)multinomial distribution

	outcome 1	outcome 2	...	outcome J	sum
$\mathbf{x}_{k1}$	1	0	...	0	1
$\mathbf{x}_{k2}$	0	1		0	1
$\mathbf{x}_{k3}$	0	1		0	1
$\mathbf{x}_{k4}$	1	0	...	0	1
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
$\mathbf{x}_{k,N_k-1}$	0	0	...	1	1
$\mathbf{x}_{k,N_k}$	1	0	...	0	1
sum$=\mathbf{x}_k$	x_{k1}	x_{k2}	...	x_{kJ}	N_k

- sum of each row is one
- sum of all row sums is N_k
- sum of each column is x_{kj}, where $j = 1, 2, \ldots, J$
- sum of all x_{kj}, $j = 1, 2, \ldots, J$, is N
- $\mathbf{x}_k = \mathbf{n}_k$, where $k = 1, 2, \ldots, K$

Probabilistic and Statistical Models

Assignments in R

Assignment **number of boys**:

1. Draw probability mass function of number of boys in the families with 12 children?
2. What are the probabilities of having n boys in the family ($n = 1, 2, \ldots, 12$)? What is the probability of having eight or more boys in the family? What is the probability of having five to seven boys in the family?

Assignment **killing by horse kick**:

1. Draw probability mass function of number of corpses with certain number of solders killed by horse kick?
2. What are the probabilities of having n annual deaths ($n = 0, 1, 2, 3, 4, 5+$)? What is the probability of having one or less annual deaths?

Assignment **accidents in the factories**:

1. Draw probability mass function of number of workers having an accident?
2. What are the probabilities of having n accidents ($n = 0, 1, 2, 3, 4, 5+$)? What is the probability of having two or more accidents?

Probabilistic and Statistical Models

Assignments in R

Assignment **number of individuals with certain socioeconomic status, political philosophy and affiliation**:

1. What is the number of all 2×4 contingency table with $N = 50$?
   $\binom{n+k-1}{k} = \binom{57}{7} = \binom{57}{50} = 264385836$
   `choose(57,50)=choose(57,7)`
2. What is the probability of getting the following 2×4 contingency table?

	D-Li	D-C	R-Li	R-C
H	5	7	4	6
Lo	8	7	3	10

$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_8 = x_8) = \frac{50!}{5!7!4!6!8!7!3!10!} 0.12^5 0.12^7 0.04^4 0.12^6 0.18^8 0.18^7 0.06^3 0.18^{10} = 2.332506 \times 10^{-6}$

```
n<-c(5,7,4,6,8,7,3,10)
p<-c(.12,.12,.04,.12,.18,.18,.06,.18)
dmultinom(x=n,prob=p) ## 2.332506e-06
```

Probabilistic and Statistical Models

Assignments in R

Assignment **number of individuals with certain socioeconomic status, political philosophy and affiliation**:

3. What is the most probable 2×4 contingency table and what is the probability of getting it?

	D-Li	D-C	R-Li	R-C
H	6	6	2	6
Lo	9	9	3	9

$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_8 = x_8) =$
$\frac{50!}{6!6!2!6!9!9!3!9!} 0.12^6 0.12^6 0.04^2 0.12^6 0.18^9 0.18^9 0.06^3 0.18^9 =$
1.020471×10^{-5}
$4.375\times$ more than in (2)

```
n<-c(6,6,2,6,9,9,3,9)
p<-c(.12,.12,.04,.12,.18,.18,.06,.18)
dmultinom(x=n,prob=p) ## 1.020471e-05
```

4. Draw probability mass function of number of possible 2×4 contingency tables with $N = 50$?

Probabilistic and Statistical Models

Likelihood function

Definition (likelihood function)

For a statistical model $\mathcal{F}$ where we expect the data $x \in \mathbb{R}$ to be observed, the function $L : \mathbf{\Theta} \to \mathbb{R}^+ \cup \{0\}$, called **likelihood function** (**likelihood**), is defined as

$$L(\theta|\mathbf{x}) = L(\theta, \mathbf{x}) = c(\mathbf{x}) f(\theta, \mathbf{x}),$$

where $c \in \mathbb{R}$ is independent of θ,

$$f(\theta|\mathbf{x}) = f(\theta, \mathbf{x}) = \prod_{i=1}^{n} f(x_i, \theta).$$

Likelihood $L(\theta|\mathbf{x})$ is used when describing a function of a parameter given an outcome.
Density (probability mass function) $f(x_i, \theta) = f(x_i|\theta)$ is used when describing a function of the outcome given a fixed parameter value.

Probabilistic and Statistical Models

Likelihood function

The **natural logarithm of the likelihood function**, called the **log-likelihood**, is defined as

$$\ln(L(\theta|\mathbf{x})) = l(\theta|\mathbf{x}) = \ln c + \ln(f(\theta|\mathbf{x})).$$

- The log-likelihood, is more convenient to work with.
- We are searching for the maximum of likelihood function.
- Because the logarithm is **a monotonically increasing function**, ***the logarithm of a function achieves its maximum value at the same points as the function itself***. Hence the log-likelihood can be used in place of the likelihood in finding the maximum.
- ***Finding the maximum of a function involves taking the (partial) derivative of a function, equaling it to zero, and solving for the parameter being maximized.***

Probabilistic and Statistical Models

Likelihood function

Definition (maximum-likelihood estimate)

The estimate of a parameter θ, $\widehat{\theta}_{ML} = \widehat{\theta}$, called **maximum-likelihood estimate** (**MLE**), is a value which maximises the likelihood function, i.e.

$$\widehat{\theta}_{ML} = \arg\max_{\forall\theta} L(\theta|\mathbf{x}) = \arg\max_{\forall\theta} l(\theta|\mathbf{x}).$$

The process of maximisation is called **maximum-likelihood estimation**:

- **the first derivative of log-likelihood function** (**score function**) $S(\theta) = \frac{\partial}{\partial\theta} l(\theta|\mathbf{x})$,
- **likelihood equations** (**score equations**) $S(\theta) = 0$,
- **the second derivative of log-likelihood function** $\frac{\partial^2}{\partial\theta^2} l(\theta|\mathbf{x})$,
- in the maximum is the second derivative negative and ***the curvature*** in $\widehat{\theta}$ is equal to **Fisher information** $\mathcal{I}(\widehat{\theta}) = -\frac{\partial^2}{\partial\theta^2} l(\theta|\mathbf{x})|_{\theta=\widehat{\theta}}$.

- The curvature is inversely related to the variance of θ, i.e. $\widehat{Var[\widehat{\theta}]} = 1/\mathcal{I}(\widehat{\theta})$.
- Since $X_i, i = 1, 2, \ldots, n$ are independent, $\mathcal{I}(\widehat{\theta}) = ni(\widehat{\theta})$.

Ronald Aylmer Fisher (1890–1962) – English statistician, wrote in 1925:

> *What has now appeared is that the mathematical concept of probability is inadequate to express our mental confidence or diffidence in making such inferences, and that the mathematical quantity which appears to be appropriate measuring our order of preference among different possible populations, does not in fact obey the* ***laws of probability****. To distinguish it from probability, I have used the term "****likelihood****" to designate this quantity.*

Let $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$, where s θ_1 is the **parameter of interest** and θ_2 a **nuisance parameter**. In some cases, the separation into two such components can be achieved after suitable reparametrisation.

If $\widehat{\theta}_{2|\theta_1}$ denotes the value of θ_2 which maximizes the likelihood (or log-likelihood) function for the given value of θ_1, we define **profile likelihood function**

$$L_P(\theta_1|\mathbf{x}) = L((\theta_1, \widehat{\theta}_{2|\theta_1})^T|\mathbf{x})$$

and **profile log-likelihood function**

$$l_P(\theta_1|\mathbf{x}) = l((\theta_1, \widehat{\theta}_{2|\theta_1})^T|\mathbf{x})$$

The term "profile" comes about through thinking of the usual (log-)likelihood function a s a hill being observed from a viewpoint with abscissa $\theta_2 = \infty$, so that, for any fixed θ_1, only the highest value $L((\theta_1, \widehat{\theta}_{2|\theta_1})^T|\mathbf{x})$ or $l((\theta_1, \widehat{\theta}_{2|\theta_1})^T|\mathbf{x})$ is seen.

Definition (likelihood and log-likelihood function of binomial distribution)

Let X be binomially distributed with parameters N and $\theta = p$, i.e. $X \sim Bin(N, p)$. Realisations of X be $x = n$. Then the **likelihood function** is equal to

$$L(p|x) = \prod_{i=1}^{N} \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i} = p^x (1-p)^{N-x} \prod_{i=1}^{N} \binom{N}{x_i}.$$

Since the product of binomial coefficients is not important in likelihood maximisation, only the ***kernel*** (often called likelihood as well) is used. Then

$$L(p|x) \approx p^x (1-p)^{N-x}.$$

The **log-likelihood function** is equal to

$$l(p|x) = x \ln p + (N-x) \ln (1-p).$$

Example (maximum-likelihood estimation)

Let X be binomially distributed with parameters N and $\theta = p$, i.e. $X \sim Bin(N, p)$. Derive $\widehat{p}$ and $\widehat{Var[\widehat{p}]}$.

Solution (partial)

$$S(p) = \frac{\partial}{\partial p} l(p|x) = \frac{x}{p} - \frac{N-x}{1-p},$$

$$\frac{\partial^2}{\partial p^2} l(p|x) = -\left(N\widehat{p}\right)/p^2 - N\left(1-\widehat{p}\right)/(1-p)^2.$$

Then

$$\widehat{p} = \frac{x}{N} \text{ and } \widehat{Var[\widehat{p}]} = \frac{\widehat{p}(1-\widehat{p})}{N}.$$

Probabilistic and Statistical Models

Likelihood function of binomial distribution

Example (maximal likelihood estimates of p)

Generate in R pseudo-random variables $X \sim Bin(N, p)$, where $N = 20$. Write R-function to calculate (1) likelihood function $L(p|x)$ of binomial distribution, where $x = 2, N = 20$, (2) likelihood function $L(p|x)$ of binomial distribution, where $x = 10, N = 20$ and (3) likelihood function $L(p|x)$ of binomial distribution, where $x = 18, N = 20$. Repeat the same for log-likelihood function. Calculate also $\widehat{p}$ using function `optimize()`. Draw all three functions in three side-by-side windows with highlighted maxima.

Solution (partial)

$l(p|x) = p^x\,(1-p)^{N-x}$, where $p \in (0,1), x = 2, N = 20$

$l(p|x) = p^x\,(1-p)^{N-x}$, where $p \in (0,1), x = 10, N = 20$

$l(p|x) = p^x\,(1-p)^{N-x}$, where $p \in (0,1), x = 18, N = 20$

Probabilistic and Statistical Models

Likelihood function of binomial distribution

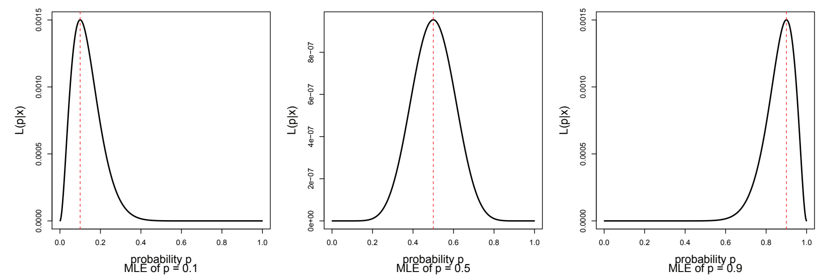

Figure: Likelihood functions of binomial distribution $X \sim Bin(N, p)$, where $N = 20$

Probabilistic and Statistical Models

Likelihood function of binomial distribution

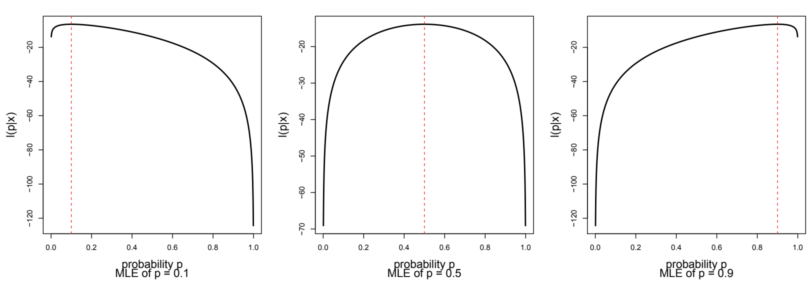

Figure: Log-likelihood functions of binomial distribution $X \sim Bin(N, p)$, where $N = 20$

Probabilistic and Statistical Models

Likelihood function of multinomial distribution

Definition (likelihood and log-likelihood function of multinomial distribution)

Let $\mathbf{X}$ be multinomially with parameters N and $\boldsymbol{\theta} = \mathbf{p}$, i.e. $\mathbf{X} \sim Mult_J\,(N, \mathbf{p})$. Realisations of X_j be $x_j = n_j$. Then the (kernel of) **likelihood function** is equal to

$$L(\mathbf{p}|\mathbf{x}) = \prod_{i=1}^{N} \frac{N!}{\prod_{j=1}^{J} x_j!} \prod_{j=1}^{J} p_j^{x_{ji}} \approx \prod_{j=1}^{J} p_j^{x_j}$$

and the **log-likelihood function** is equal to

$$l\,(\mathbf{p}|\mathbf{x}) = \sum_{j=1}^{J} x_j \ln p_j .$$

Probabilistic and Statistical Models

Likelihood function of multinomial distribution

Example (maximum-likelihood estimation)

Let $\mathbf{X}$ be multinomially with parameters N and $\theta = \mathbf{p}$, i.e. $\mathbf{X} \sim Mult_J(N, \mathbf{p})$. Derive $\widehat{\mathbf{p}}$ and $\widehat{Var}[\widehat{\mathbf{p}}]$.

Solution (partial)

Let $p_J = 1 - \sum_{j=1}^{J-1} p_j$ and $\mathbf{p} = (p_1, p_2, \ldots, p_{J-1})^T$

Then

$$l(\mathbf{p}|\mathbf{x}) = \sum_{j=1}^{J-1} n_j \ln p_j + n_J \ln(1 - \sum_{j=1}^{J-1} p_j)$$

and

$$(S(\mathbf{p}))_j = \frac{\partial}{\partial p_j} l(\mathbf{p}|\mathbf{x}) = \frac{n_j}{p_j} - \frac{n_J}{p_J}$$

as the elements of $S(\mathbf{p})$). Then

$$\mathcal{I}(\mathbf{p}) = -\frac{\partial}{\partial \mathbf{p}} S(\mathbf{p}) = \text{diag}\left(\frac{n_1}{p_1^2}, \frac{n_2}{p_2^2}, \ldots, \frac{n_{J-1}}{p_{J-1}^2}\right) + \frac{n_J}{p_J^2}\mathbf{1}\mathbf{1}^T.$$

Probabilistic and Statistical Models

Likelihood function of multinomial distribution

$$\mathcal{I}(\widehat{\mathbf{p}}) = N\left(\text{diag}\left(\frac{1}{\widehat{p}_1}, \frac{1}{\widehat{p}_2}, \ldots, \frac{1}{\widehat{p}_{J-1}}\right) + \frac{\mathbf{1}\mathbf{1}^T}{\widehat{p}_J}\right).$$

Then

$$\mathcal{I}(\widehat{\mathbf{p}}) = N\begin{pmatrix} \frac{1}{\widehat{p}_1} + \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_J} & \cdots & \frac{1}{\widehat{p}_J} \\ \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_2} + \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_J} & \cdots & \frac{1}{\widehat{p}_J} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_J} & \cdots & \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_{J-1}} + \frac{1}{\widehat{p}_J} \end{pmatrix}.$$

$$\widehat{Var}[\widehat{\mathbf{p}}] = \mathcal{I}^{-1}(\widehat{\mathbf{p}}) = \frac{1}{N}\left(\text{diag}\left(\widehat{\mathbf{p}}\right) - \widehat{\mathbf{p}}\widehat{\mathbf{p}}^T\right).$$

Then

$$\widehat{Var}[\widehat{\mathbf{p}}] = \frac{1}{N}\begin{pmatrix} \widehat{p}_1(1-\widehat{p}_1) & -\widehat{p}_1\widehat{p}_2 & \cdots & -\widehat{p}_1\widehat{p}_{J-1} \\ -\widehat{p}_2\widehat{p}_1 & \widehat{p}_2(1-\widehat{p}_2) & \cdots & -\widehat{p}_2\widehat{p}_{J-1} \\ \vdots & \vdots & \vdots & \vdots \\ -\widehat{p}_{J-1}\widehat{p}_1 & -\widehat{p}_{J-1}\widehat{p}_2 & \cdots & \widehat{p}_{J-1}(1-\widehat{p}_{J-1}) \end{pmatrix}.$$

Probabilistic and Statistical Models

Likelihood function of multinomial distribution

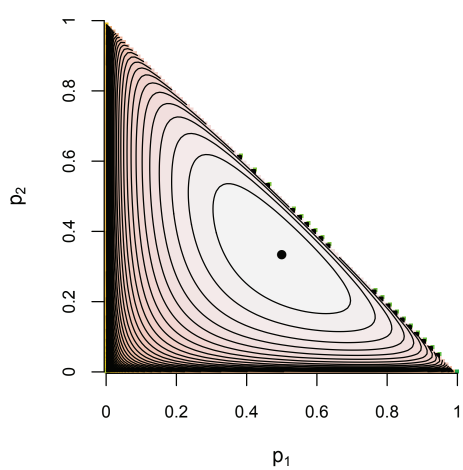

Figure: Log-likelihood function of multinomial (trinomial) distribution

Probabilistic and Statistical Models

Likelihood function of Poisson distribution

Definition (likelihood and log-likelihood function of Poisson distribution)

Let X be distributed as Poisson with parameter $\theta = \lambda$, i.e. $X \sim Poiss(\lambda)$. Realisations of X_j be $x_j = n_j$. Then the (kernel of) **likelihood function** is equal to

$$L(\lambda|\mathbf{x}) = \prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \approx \lambda^{\sum_{i=1}^{N} x_i} e^{-N\lambda}$$

and the **log-likelihood function** is equal to

$$l(\lambda|\mathbf{x}) = \sum_{i=1}^{N} x_i \ln \lambda - N\lambda.$$

In general, $L(\lambda|\mathbf{x}) = \prod_n p_n^{m_n}$, where $p_n = \Pr(X = n) = e^{-\lambda}\lambda^n/n!$ and $l(\lambda|\mathbf{x}) = -\lambda \sum_n m_n + \sum_n n m_n \ln \lambda$.

Example (maximum-likelihood estimation)

Let X be distributed Poisson with parameter $\theta = \lambda$, i.e. $X \sim Poiss(\lambda)$. Derive $\widehat{\lambda}$ and $\widehat{Var}[\widehat{\lambda}]$.

Solution (partial)

$$S(\lambda) = \frac{\partial}{\partial \lambda} l\,(\lambda|\mathbf{x}) = \frac{\sum_{i=1}^{N} x_i}{\lambda} - N,$$

$$\frac{\partial^2}{\partial \lambda^2} l\,(\lambda|\mathbf{x}) = -\frac{\sum_{i=1}^{N} x_i}{\lambda^2}.$$

Then

$$\widehat{\lambda} = \frac{\sum_{i=1}^{N} x_i}{N} = \overline{x} \text{ and } \widehat{Var}[\widehat{\lambda}] = \frac{\overline{x}}{N}$$

In general notation, $\widehat{\lambda} = \frac{\sum_n n m_n}{\sum_n m_n}$.

Example (maximal likelihood estimates of λ)

Write R-function to calculate likelihood function $L(\lambda|x)$ and log-likelihood function $l(\lambda|x)$ of Poisson distribution $X \sim Poiss(\lambda)$ for horse kick data. Calculate also $\widehat{\lambda}$ using function `optimize()`. Draw both functions in two side-by-side windows with highlighted maximum.

Solution (partial)

$l(\lambda|\mathbf{x}) = -\lambda \sum_n m_n + \sum_n n m_n \ln \lambda$, where $\lambda \in (0, 2)$

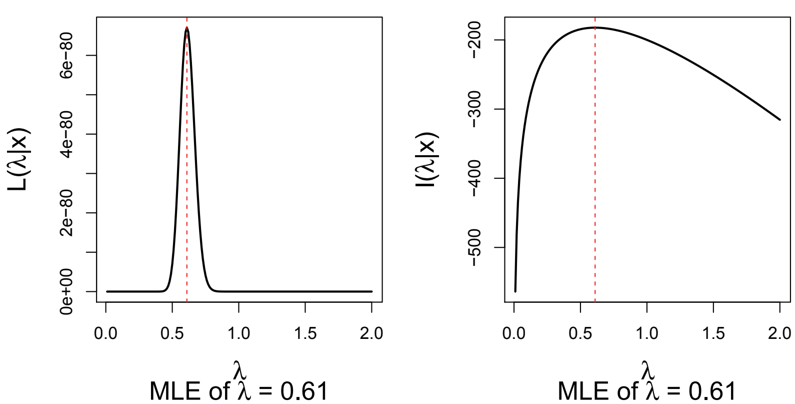

Figure: Likelihood function $L(\lambda|\mathbf{x})$ (left) anf log-likelohood function $l(\lambda|\mathbf{x})$ of Poisson distribution $X \sim Poiss(\lambda)$ for horse kick data

Assignment **number of boys**:

Calculate $\widehat{p}$ (the probability of having a boy in a family) and $\widehat{Var}[\widehat{p}]$ (the variance probability of having a boy in a family).

Assignment **killing by horse kick**:

Calculate $\widehat{\lambda}$ (the mean number of annual deaths) and $\widehat{Var}[\widehat{\lambda}]$ (the variance of mean number of annual deaths).

Assignment **accidents in a factory**:

Calculate $\widehat{\lambda}$ (the mean number of accidents in a factory) and $\widehat{Var}[\widehat{\lambda}]$ (the variance of mean number of accidents in a factory).

Probabilistic and Statistical Models

Assignments in R

Assignment **blood groups**:
In Prague and Košice, calculate $\widehat{\mathbf{p}}$ (the probabilities of having certain blood group in particular city) and $\widehat{Var[\widehat{\mathbf{p}}]}$ (the covariance matrix of probability of having certain blood group in particular city).

Assignment **eye and hair color**:
Calculate $\widehat{\mathbf{p}}$ (the probabilities of having certain eye and hair color) and $\widehat{Var[\widehat{\mathbf{p}}]}$ (the covariance matrix of probability of having certain eye and hair color).

Probabilistic and Statistical Models

Likelihood function of normal distribution

Definition (likelihood and log-likelihood function of normal distribution)

Let X be distributed normally with parameter $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, i.e. $X \sim N(\mu, \sigma^2)$. Realisations of X_i be x_i. Then the **likelihood function** is equal to

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)\right) \end{aligned}$$

and the **log-likelihood function** is equal to

$$l(\boldsymbol{\theta}|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 .$$

Probabilistic and Statistical Models

Likelihood function of normal distribution

Example (maximum-likelihood estimation)

Let X be distributed normally with parameter $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, i.e. $X \sim N(\mu, \sigma^2)$. Derive $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\sigma}^2)^T$ and $\widehat{Var[\widehat{\boldsymbol{\theta}}]} = \widehat{\boldsymbol{\Sigma}}$.

Solution (partial)

$$S_1(\boldsymbol{\theta}) = \frac{\partial}{\partial\mu} l(\boldsymbol{\theta}|\mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu),$$

$$S_2(\boldsymbol{\theta}) = \frac{\partial}{\partial\sigma^2} l(\boldsymbol{\theta}|\mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Then

$$\widehat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\widehat{\mu})^2, \text{ and } \mathcal{I}(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\widehat{\sigma}^4} \end{pmatrix}.$$

Probabilistic and Statistical Models

Likelihood function of normal distribution

Example (maximal likelihood estimates of μ and σ^2)

Generate in R pseudo-random variables $X \sim N(\mu, \sigma^2)$, where $\mu = 4$, $\sigma^2 = 1$ and $n = 1000$. Write R-function to calculate (1) (profile) likelihood function $L_P(\mu|\mathbf{x})$ of normal distribution for generated data X, (2) (profile) likelihood function $L_P(\sigma^2|\mathbf{x})$ of normal distribution for generated data X, and (3) likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ of normal distribution for generated data X, where $\boldsymbol{\theta} = (\mu, \sigma^2)^T$. Repeat the same for log-likelihood function. Calculate also MLEs using functions `optimize()` and `optim()`. Draw all three functions in three side-by-side windows with highlighted maxima.

Solution (partial)
$l_P(\mu|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma_1^2 - \frac{1}{2\sigma_1^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)$,
where $\mu \in (2, 6), \widehat{\sigma}_\mu = 1$;
$l_P(\sigma^2|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{\sum_{i=1}^{n}(x_i-\mu_1)^2}{2\sigma^2}$, where
$\widehat{\mu}_\sigma = 4, \sigma \in (0.5, 1.5)$;
$l(\boldsymbol{\theta}|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$, where $\mu \in (2, 6)$ and
$\sigma \in (0.5, 1.5)$.

Probabilistic and Statistical Models

Likelihood function of normal distribution

```
n <- 1000
# Profile likelihood for mu
sigma.mu <- 1
x <- rnorm(n,4,sigma)
"negloglikemu" <- function(mu){n/2*log(2*pi)
        +n/2*log(sigma.mu^2)+(sum(x^2)-2*mu*sum(x)+n*mu^2)/(2*
             sigma.mu^2)}
OPTmu <- optimize(negloglikemu,c(2,6),maximum=FALSE)
OPTmu$minimum # 3.987524
# Profile likelihood for sigma^2
mu.sigma <- 4
"negloglikesigma" <- function(sigma2){n/2*log(2*pi)
        +n/2*log(sigma2)+sum((x-mu.sigma)^2)/(2*sigma2)}
OPTsigma <- optimize(negloglikesigma,c(0.5,1.5),maximum=FALSE)
OPTsigma$minimum # 0.9630124
# Likelihood for mu and sigma^2
"negloglike" <- function(theta){(n/2)*log(2*pi)
      +(n/2)*log(theta[2])+(1/(2*theta[2]))*sum((x-theta[1])^2)}
OPTboth <- optim(c(3,0.5),negloglike,method="Nelder-Mead",
                   hessian=TRUE)
OPTboth$parameter # 3.9875376 0.9627521
```

Probabilistic and Statistical Models

Likelihood function of normal distribution

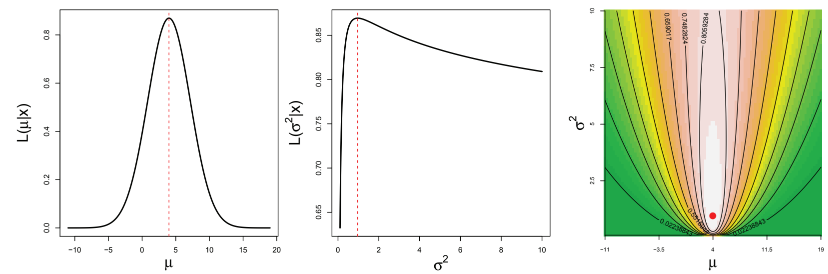

Figure: Profile likelihood functions (left, middle) and likelihood function (right) of normal distribution $X \sim N(\mu, \sigma^2)$, where $\mu = 4, \sigma^2 = 1$ and $n = 1000$; all functions multiplied by suitable constant, here $10^{-4}L(\cdot)$

Probabilistic and Statistical Models

Likelihood function of normal distribution

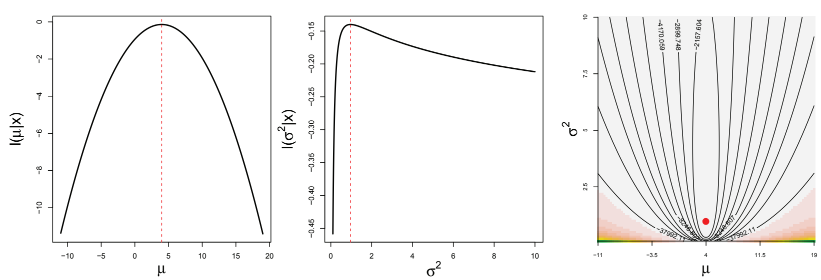

Figure: Profile log-likelihood functions (left, middle) and log-likelihood function (right) of normal distribution $X \sim N(\mu, \sigma^2)$, where $\mu = 4, \sigma^2 = 1$ and $n = 1000$; all functions are multiplied by suitable constant, here $\exp(10^{-4}L(\cdot))$

Probabilistic and Statistical Models

Likelihood function of normal distribution

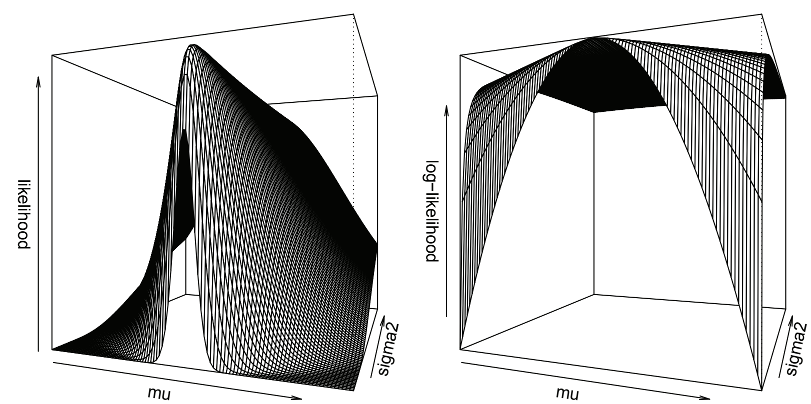

Figure: Likelihood (left) and log-likelihood (right) function of normal distribution $X \sim N(\mu, \sigma^2)$, where $\mu = 4, \sigma^2 = 1$ and $n = 1000$; all functions are multiplied by suitable constant, here $\exp(10^{-4}L(\cdot))$ and $10^{-4}L(\cdot)$

Definition (relative likelihood and log-likelihood function)

Relative likelihood function is defined as

$$\mathcal{L}(\theta|\mathbf{x}) = \frac{L(\theta|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})}$$

and **relative log-likelihood function** as

$$\ln \mathcal{L}(\theta|\mathbf{x}) = \ln \frac{L(\theta|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})}.$$

- It is often useful that likelihood function could be **approximated by a quadratic function**.
- But additionally to the location of maxima of likelihood function, we need **the curvature around maximum**.
- Since the log-likelihood, is more convenient to work with, we need **a quadratic approximation of log-likelihood function**.

Definition (Taylor polynomial of order r)

if a function $g(x)$ has derivatives of order r, that is, $g^{(r)}(x) = \frac{\partial^r}{\partial x^r} g(x)$ exists, then for any constant a, the**Taylor polynomial of order r about a** is

$$T_r(x) = \sum_{j=0}^{r} \frac{g^{(j)}(a)}{j!}(x-a)^j.$$

In practical statistical situations we assume that the **remainder** $g(x) - T_r(x)$ converges to zero as n increases, therefore we are going to ignore it. There are many explicit forms, one of the most useful is

$$g(x) - T_r(x) = \int_a^x \frac{g^{(r+1)}(t)}{r!}(x-t)^r dt.$$

If $g^{(r)}(a) = \frac{\partial r}{\partial x^r} g(x)|_{x=a}$ exists, then

$$\lim_{x \to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0.$$

The **quadratic approximation of log-likelihood function** about $\widehat{\theta}$ defined as

$$l(\theta|\mathbf{x}) \approx l(\widehat{\theta}|\mathbf{x}) + S(\widehat{\theta})(\theta - \widehat{\theta}) - \frac{1}{2}\mathcal{I}(\widehat{\theta})(\theta - \widehat{\theta})^2,$$

The **quadratic approximation of relative likelihood function** about $\widehat{\theta}$ is defined as

$$\ln \mathcal{L}(\theta|\mathbf{x}) = \ln \frac{L(\theta|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})} = l(\theta|\mathbf{x}) - l(\widehat{\theta}|\mathbf{x}) \approx -\frac{1}{2}\mathcal{I}(\widehat{\theta})(\theta - \widehat{\theta})^2.$$

It is often useful to visualise a derivative of the quadratic approximation $S(\theta) \approx -\mathcal{I}(\widehat{\theta})(\theta - \widehat{\theta})$ or $-\mathcal{I}^{-1/2}(\widehat{\theta})S(\theta) \approx \mathcal{I}^{1/2}(\widehat{\theta})(\theta - \widehat{\theta})$, where $-\mathcal{I}^{-1/2}(\widehat{\theta})S(\theta)$ is visualised against $\mathcal{I}^{1/2}(\widehat{\theta})(\theta - \widehat{\theta})$. *If the quadratic approximation is correct, this should be a line with slope equal to one.*

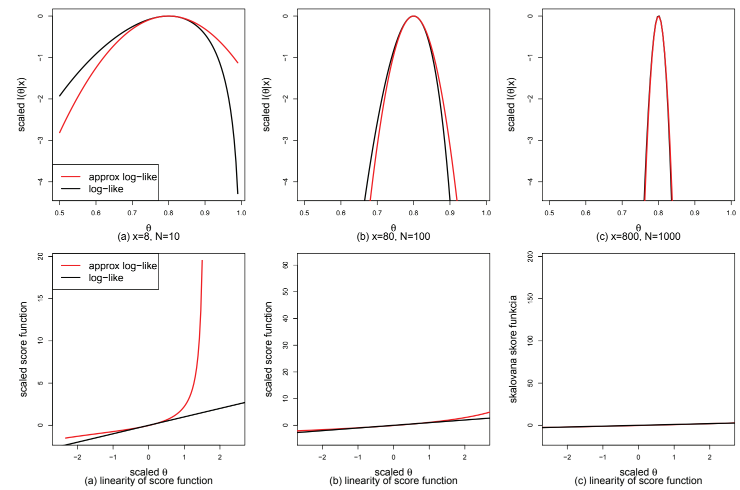

Figure: Relative binomial log-likelihood, its quadratic approximation and linearity of score function

Probabilistic and Statistical Models
Numerical maximisation of likelihood function

Isaac Newton (1643–1727) and **Joseph Raphson** (1648–1715).

Definition (Newton-Raphson method)
Having **quadratic approximation of log-likelihood function** about θ_0

$$l(\theta|\mathbf{x}) \approx l(\theta_0|\mathbf{x}) + S(\theta_0)(\theta-\theta_0) - \frac{1}{2}\mathcal{I}(\theta_0)(\theta-\theta_0)^2$$

or **linear approximation of score function** about θ_0

$$S(\theta) \approx S(\theta_0) - \mathcal{I}(\theta_0)(\theta-\theta_0),$$

the numerical maximisation can be done via **iterative function**

$$\theta_0 + \frac{S(\theta_0)}{\mathcal{I}(\theta_0)}.$$

Probabilistic and Statistical Models
Numerical maximisation of likelihood function

The iterative process is defined as follows:

1. initialisation step – starting point $\theta^{(0)}$, where $\mathcal{I}(\theta^{(0)}) \neq 0$,
2. updating equations – iteration of
$$\theta^{(i)} = \theta^{(i-1)} + \frac{S(\theta^{(i-1)})}{\mathcal{I}(\theta^{(i-1)})},$$
where $\mathcal{I}(\theta^{(i-1)}) \neq 0$, pre $i = 1, 2, \dots$
3. stopping rule based on absolute convergence criteria – until $|l(\theta^{(i)}|\mathbf{x}) - l(\theta^{(i-1)}|\mathbf{x})| < \epsilon$, where the **threshold** ϵ is sufficiently small

Geometrical interpretation: $\theta^{(i)}$ is a crossing point of tangent of score function $S(\cdot)$ in the point $[\theta^{(i-1)}, S(\theta^{(i-1)})]$ with x-axis.

In R:

- `optimize(f,interval,maximum= FALSE, tol,...)`
- Newton-Raphson method is combined here with **golden section method** and **successive parabolic interpolation** to speed up the convergence.

Probabilistic and Statistical Models
Numerical maximisation of likelihood function

Definition (multivariate Newton-Raphson method)
Having **quadratic approximation of log-likelihood function** about θ_0

$$l(\boldsymbol{\theta}|\mathbf{x}) \approx l(\boldsymbol{\theta}_0|\mathbf{x}) + S(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathcal{I}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$$

or **linear approximation of score function** about θ_0

$$S(\boldsymbol{\theta}) \approx S(\boldsymbol{\theta}_0) - \mathcal{I}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0).$$

the numerical maximisation can be done via **iterative function**

$$\boldsymbol{\theta}_0 + (\mathcal{I}(\boldsymbol{\theta}_0))^{-1}S(\boldsymbol{\theta}_0).$$

Probabilistic and Statistical Models
Numerical maximisation of likelihood function

The iterative process is defined as follows:

1. initialisation step – starting point $\boldsymbol{\theta}^{(0)}$, where $\mathcal{I}(\boldsymbol{\theta}^{(0)}) \neq \mathbf{0}$,
2. updating equations – iteration of
$$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)} + (\mathcal{I}(\boldsymbol{\theta}^{(i-1)}))^{-1}S(\boldsymbol{\theta}^{(i-1)}),$$
where $\mathcal{I}(\boldsymbol{\theta}^{(i-1)}) \neq \mathbf{0}$, pre $i = 1, 2, \dots$
3. stopping rule based on absolute convergence criteria – until $|l(\boldsymbol{\theta}^{(i)}|\mathbf{x}) - l(\boldsymbol{\theta}^{(i-1}|\mathbf{x})| < \epsilon$, where the **threshold** ϵ is sufficiently small

In R:

- `optim(par,fn,gr,method,control,hessian =FALSE,...)`
- Newton-Raphson method is often modified – **Fisher scoring method**, **quasi Newton method**, **Broyden-Fletcher-Goldfarb-Shannon (BFGS) method**

Probabilistic and Statistical Models

Numerical maximisation of likelihood $\approx$ minimising negative log-likelihood

Nelder-Mead method (method of simplexes) – the idea of "jumps" across triangles defined by the points $\boldsymbol{\theta}_1^{(i-1)}$, $\boldsymbol{\theta}_2^{(i-1)}$, $\boldsymbol{\theta}_3^{(i-1)}$, where $l(\boldsymbol{\theta}_1^{(i-1)}|\mathbf{x}) < l(\boldsymbol{\theta}_2^{(i-1)}|\mathbf{x}) < l(\boldsymbol{\theta}_3^{(i-1)}|\mathbf{x})$. We are substituting point $\boldsymbol{\theta}_1^{(i-1)}$ with a "better" point $\boldsymbol{\theta}_1^{(i)}$, where $l(\boldsymbol{\theta}_1^{(i)}|\mathbf{x}) < l(\boldsymbol{\theta}_1^{(i-1)}|\mathbf{x})$. Then new point is defined based on ***reflection*** (***point symmetry***), ***contraction*** or ***extrapolation*** (***expansion***) as

1. reflection – $\mathbf{z}_1 = \boldsymbol{\theta}_1^{(i)} = \boldsymbol{\theta}_{23}^{(i-1)} + 1\left(\boldsymbol{\theta}_{23}^{(i-1)} - \boldsymbol{\theta}_1^{(i-1)}\right)$,
2. reflection & expansion – $\mathbf{z}_2 = \boldsymbol{\theta}_1^{(i)} = \boldsymbol{\theta}_{23}^{(i-1)} + 2\left(\boldsymbol{\theta}_{23}^{(i-1)} - \boldsymbol{\theta}_1^{(i-1)}\right)$,
3. contraction A – $\mathbf{z}_3 = \boldsymbol{\theta}_1^{(i)} = \boldsymbol{\theta}_{23}^{(i-1)} + \frac{1}{2}\left(\boldsymbol{\theta}_{23}^{(i-1)} - \boldsymbol{\theta}_1^{(i-1)}\right)$,
4. contraction B – $\mathbf{z}_4 = \boldsymbol{\theta}_2^{(i)} = \boldsymbol{\theta}_1^{(i-1)} + \frac{1}{2}\left(\boldsymbol{\theta}_2^{(i-1)} - \boldsymbol{\theta}_1^{(i-1)}\right)$ and
   $\mathbf{z}_5 = \boldsymbol{\theta}_3^{(i)} = \boldsymbol{\theta}_1^{(i-1)} + \frac{1}{2}\left(\boldsymbol{\theta}_3^{(i-1)} - \boldsymbol{\theta}_1^{(i-1)}\right)$

where $\boldsymbol{\theta}_{23}^{(i-1)} = \frac{\boldsymbol{\theta}_2^{(i-1)}+\boldsymbol{\theta}_3^{(i-1)}}{2}$, i.e. the mid-point of the line defined by the points $\boldsymbol{\theta}_2^{(i-1)}$ and $\boldsymbol{\theta}^{(i-1)}$. If $l(\boldsymbol{\theta}_1^{(i)}|\mathbf{x}) < l(\boldsymbol{\theta}_1^{(i-1)}|\mathbf{x})$ then new triangle is defined with $\boldsymbol{\theta}_1^{(i)}$, $\boldsymbol{\theta}_2^{(i-1)}$, $\boldsymbol{\theta}_3^{(i-1)}$ for (1) to (3). Otherwise new triangle is $\boldsymbol{\theta}_1^{(i-1)}$, $\boldsymbol{\theta}_2^{(i)}$, $\boldsymbol{\theta}_3^{(i)}$.

Probabilistic and Statistical Models

Numerical maximisation of likelihood $\approx$ minimising negative log-likelihood

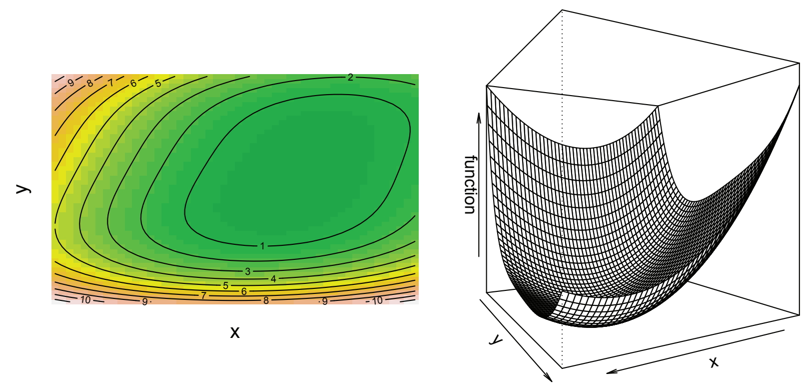

Figure: Demonstration of Nelder-Mead method or minimising the function $((x-y)^2 + (x-2)^2 + (y-3)^4)/10$, number of iterations is 49

Probabilistic and Statistical Models

Numerical maximisation of likelihood $\approx$ minimising negative log-likelihood

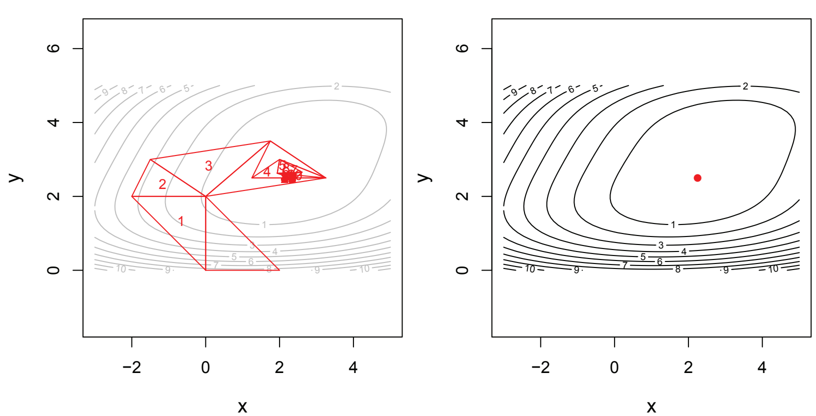

Figure: Demonstration of Nelder-Mead method or minimising the function $((x-y)^2 + (x-2)^2 + (y-3)^4)/10$, number of iterations is 49