Statistical Inference I and II

Likelihood function

Stanislav Katina[1]

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

November 13, 2018

Definition (likelihood function)

For a statistical model $\mathcal{F}$ where we expect the data $x \in \mathbb{R}$ to be observed, the function $L : \boldsymbol{\Theta} \to \mathbb{R}^+ \cup \{0\}$, 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.

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 maximised.***

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** $\frac{\partial^2}{\partial\theta^2} l(\theta|\mathbf{x})$,
- the second derivative is negative at the point of maximum 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 $\widehat{\theta}$, i.e. $\widehat{Var}[\widehat{\theta}] = 1/\mathcal{I}(\widehat{\theta})$.
- Since $X_i, i = 1, 2, \ldots, n$ are independent, $\mathcal{I}(\widehat{\theta}) = ni(\widehat{\theta})$, where $i(\widehat{\theta})$ is a likelihood of one observation.

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 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 maximises the likelihood (or log-likelihood) function for the given value of θ_1, the **profile likelihood function** is defined as

$$L_P(\theta_1|\mathbf{x}) = \max_{\forall\theta_2} L(\boldsymbol{\theta}|\mathbf{x}) = L((\theta_1, \widehat{\theta}_{2|\theta_1})^T|\mathbf{x})$$

and **profile log-likelihood function** as

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

The term "profile" comes about through thinking of the usual (log-)likelihood function as 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.

Profile relative likelihood function is defined as:

$$\mathcal{L}_P(\theta_1|\mathbf{x}) = \frac{L((\theta_1, \widehat{\theta}_{2|\theta_1})^T|\mathbf{x})}{L((\widehat{\theta}_1, \widehat{\theta}_{2|\theta_1})^T|\mathbf{x})}$$

and **profile relative log-likelihood function** as

$$\ln \mathcal{L}_P(\theta_1|\mathbf{x}) = \ln \frac{L((\theta_1, \widehat{\theta}_{2|\theta_1})^T|\mathbf{x})}{L((\widehat{\theta}_1, \widehat{\theta}_{2|\theta_1})^T|\mathbf{x})}.$$

The **estimated likelihood function** is defined as

$$L_e(\theta_1|\mathbf{x}) = L((\theta_1, \widehat{\theta}_2)^T|\mathbf{x})$$

and **estimated log-likelihood function** as

$$l_e(\theta_1|\mathbf{x}) = l((\theta_1, \widehat{\theta}_2)^T|\mathbf{x}).$$

Estimated relative likelihood function is defined as:

$$\mathcal{L}_e(\theta_1|\mathbf{x}) = \frac{L((\theta_1, \widehat{\theta}_2)^T|\mathbf{x})}{L((\widehat{\theta}_1, \widehat{\theta}_2)^T|\mathbf{x})}$$

and **estimated relative log-likelihood function** as

$$\ln \mathcal{L}_e(\theta_1|\mathbf{x}) = \ln \frac{L((\theta_1, \widehat{\theta}_2)^T|\mathbf{x})}{L((\widehat{\theta}_1, \widehat{\theta}_2)^T|\mathbf{x})}.$$

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

Let X be binomially distributed with sample size N and parameter $\theta = p$, 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 of parameter p)

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

Solution

$$S(p) = \frac{\partial}{\partial p} l(p|x) = \frac{x}{p} - \frac{N-x}{1-p}, \text{ where if } S(p) = 0, \text{ then } \widehat{p} = \frac{x}{N}.$$

$$\frac{\partial^2}{\partial p^2} l(p|x) = -\frac{x}{p^2} - \frac{N-x}{(1-p)^2}, \text{ where if}$$

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

If $p = \widehat{p}$, then

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

Example (maximal likelihood estimation of parameter p)

Generate in R pseudo-random variables $X \sim Bin(N, p)$, where $N = 20$. Write R-function to calculate likelihood function $L(p|x)$ of binomial distribution and visualise it for (1) $x = 2, N = 20$, (2) $x = 10, N = 20$ and (3) $x = 18, N = 20$. 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 $p \in (0, 1), x = 18, N = 20$

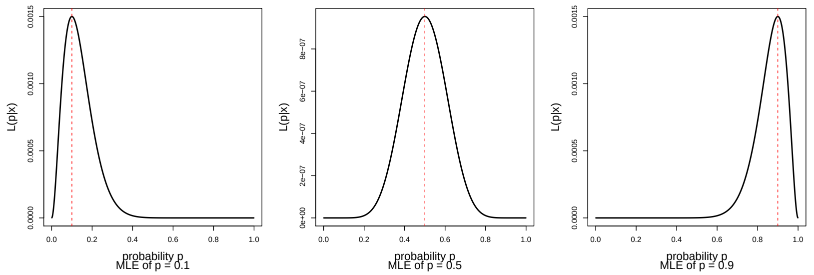

Figure: Likelihood functions of binomial distribution $X \sim Bin(N, p)$, where $N = 20$

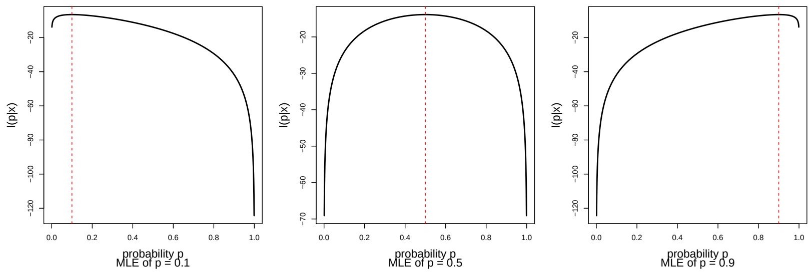

Figure: Log-likelihood functions of binomial distribution $X \sim Bin(N, p)$, where $N = 20$

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

Let $\mathbf{X}$ be multinomially distributed with sample size N and parameter $\boldsymbol{\theta} = \mathbf{p}$, 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

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

Example (maximum-likelihood estimation of parameter $\mathbf{p}$)

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

Solution

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

Then

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

$$(S(\mathbf{p}))_j = \frac{\partial}{\partial p_j} l(\mathbf{p}|\mathbf{x}) = \frac{n_j}{p_j} - \frac{n_J}{p_J}, \text{ where if } (S(\mathbf{p}))_j = 0, \text{ then } \widehat{p}_j = \frac{n_j}{N},$$

where $(S(\mathbf{p}))_j$ are the elements of $S(\mathbf{p})$. Then

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

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

Then

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

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

Then

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

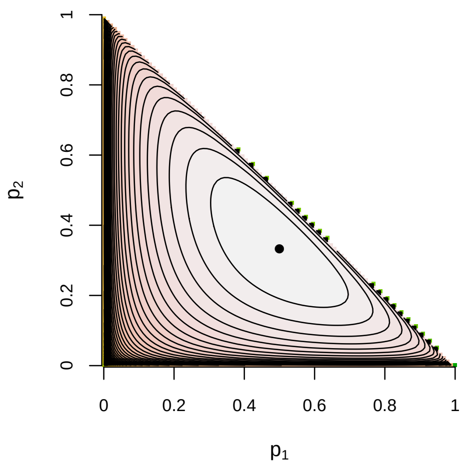

Figure: Log-likelihood function of multinomial (trinomial) distribution

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

Example (maximum-likelihood estimation)

Let X be distributed as 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 estimation of parameter 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)
$l(\lambda|\mathbf{x}) = \sum_n n m_n \ln\lambda - \lambda\sum_n m_n$, where $\lambda \in (0, 2)$

Probabilistic and Statistical Models
Likelihood function of Poisson distribution

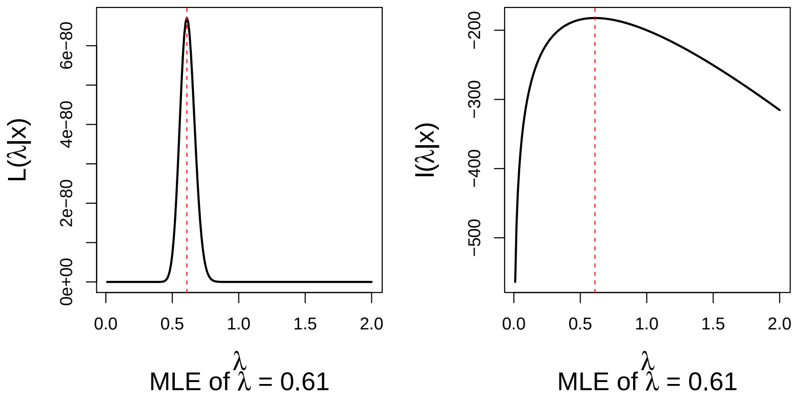

Figure: Likelihood function $L(\lambda|\mathbf{x})$ (left) and log-likelohood function $l(\lambda|\mathbf{x})$ of Poisson distribution $X \sim Poiss(\lambda)$ for horse kick data

Probabilistic and Statistical Models
Assignments in R

Assignment **number of boys**:

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

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

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

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

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

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

Example (maximum-likelihood estimation of parameters μ and σ^2)

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

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

Example (maximal likelihood estimation of parameters 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)

$l_P(\mu|\mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma_\mu^2 - \frac{1}{2\sigma_\mu^2} \left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2\right)$, where $\mu \in (2, 6), \sigma_\mu = 1$;

$l_P(\sigma^2|\mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{\sum_{i=1}^{n}(x_i - \mu_\sigma)^2}{2\sigma^2}$, where $\mu_\sigma = 4, \sigma \in (0.5, 1.5)$;

$l(\boldsymbol{\theta}|\mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}$, where $\mu \in (2, 6)$ and $\sigma \in (0.5, 1.5)$.

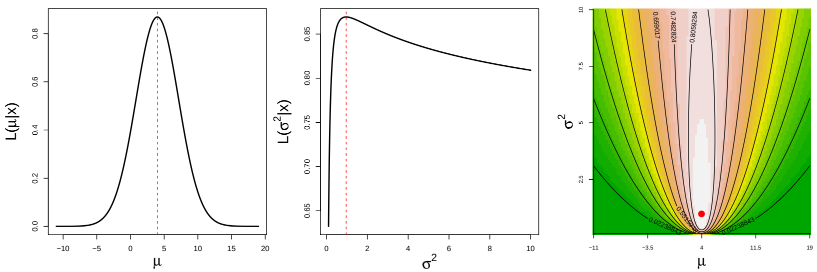

Figure: Profile likelihood functions (left, middle) and likelihood function (right) of normal distribution $X \sim N(\mu, \sigma^2)$, where $\mu = 4, \sigma^2 = 1$ and $n = 1000$; all functions multiplied by suitable constant

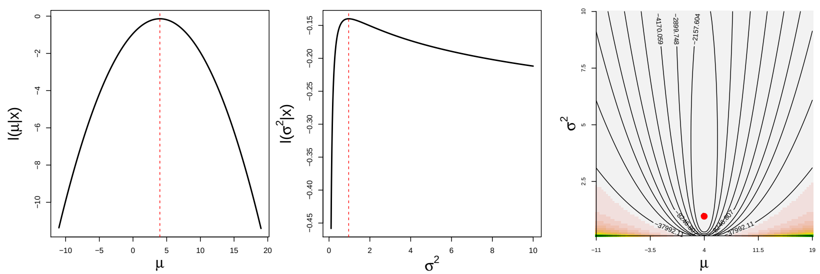

Figure: Profile log-likelihood functions (left, middle) and log-likelihood function (right) of normal distribution $X \sim N(\mu, \sigma^2)$, where $\mu = 4, \sigma^2 = 1$ and $n = 1000$; all functions are multiplied by suitable constant

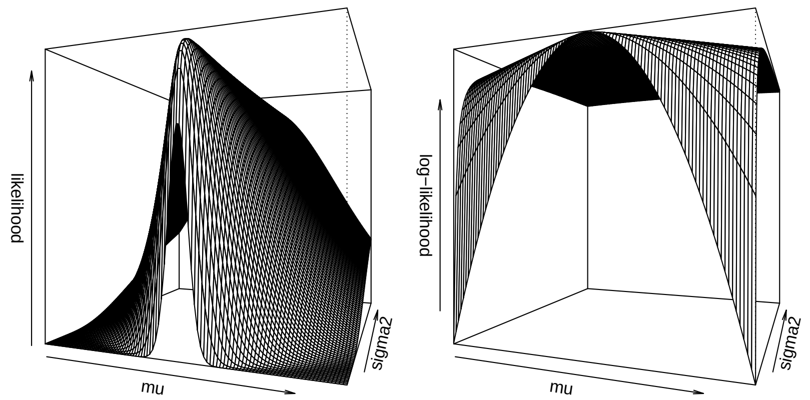

Figure: Likelihood (left) and log-likelihood (right) function of normal distribution $X \sim N(\mu, \sigma^2)$, where $\mu = 4, \sigma^2 = 1$ and $n = 1000$; all functions are multiplied by suitable constant

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 r 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 $\widehat{\theta}$ is 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 log-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.*

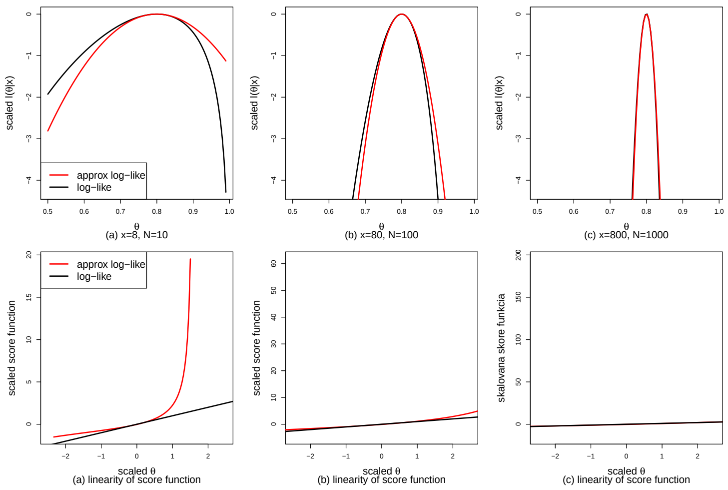

Figure: Relative binomial log-likelihood, its quadratic approximation (top) and linearity of score function (bottom)

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**

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

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 $\mathcal{I}(\theta^{(i-1)}) \neq 0$, for $i = 1, 2, \ldots$

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.

Definition (multivariate Newton-Raphson method)

Having **quadratic approximation of log-likelihood function** about $\boldsymbol{\theta}_0$

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

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**

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

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 $\mathcal{I}(\boldsymbol{\theta})$ is regular, i.e. $\det\left(\mathcal{I}(\boldsymbol{\theta}^{(i-1)})\right) \neq \mathbf{0}$, for $i = 1, 2, \ldots$
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**

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

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 $\boldsymbol{\theta}_1^{(i-1)}$, $\boldsymbol{\theta}_2^{(i)}$, $\boldsymbol{\theta}_3^{(i)}$.

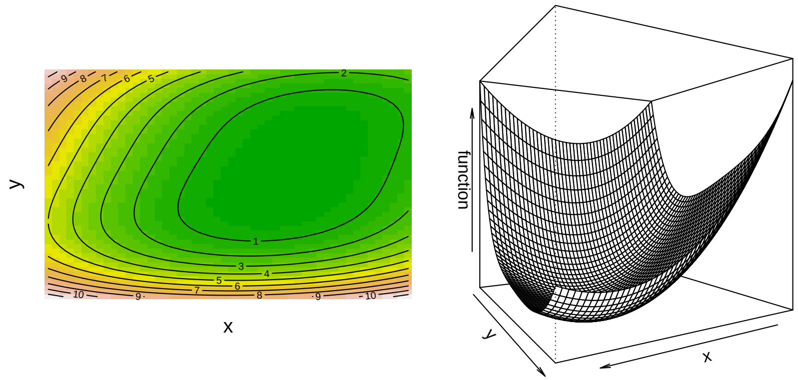

Figure: Demonstration of Nelder-Mead method of minimising the function $((x-y)^2 + (x-2)^2 + (y-3)^4)/10$, number of iterations is 49

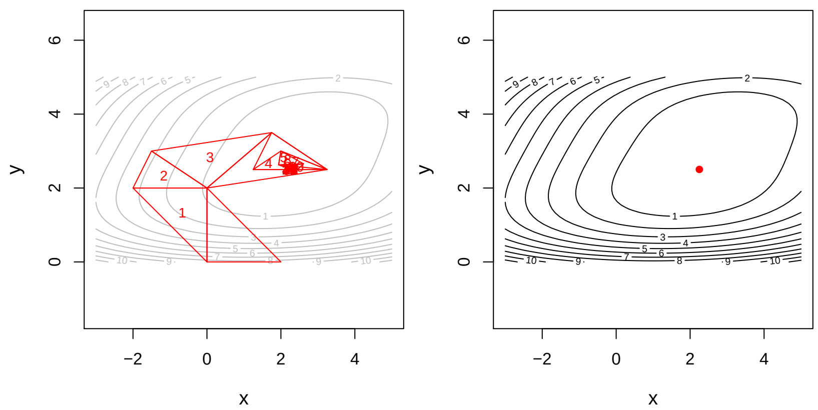

Figure: Demonstration of Nelder-Mead method of minimising the function $((x-y)^2 + (x-2)^2 + (y-3)^4)/10$, number of iterations is 49

Given data $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, the log-likelihood function $L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^n f(x_i, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^T$, must be maximised numerically. One complication is that $l(\boldsymbol{\theta}|\mathbf{x})$ is unbounded. To see how this can happen, fix p, μ_2 and σ_2^2 at any set values, with the exception that p is not equal to 0 or 1. Denote these fixed values by p_*, $\mu_{2,*}$ and $\sigma_{2,*}^2$, respectively. Now, set $\mu_1 = x_i$ for any choice of $i \in \{1, 2, \ldots, n\}$. This leaves only σ_1^2 unspecified, and $\boldsymbol{\theta}_{\sigma_1^2} = (p_*, x_i, \mu_{2,*}, \sigma_1^2, \sigma_{2,*}^2)^T$ can be used to denote the parameter vector with the values of the other parameters fixed as described. When $\boldsymbol{\theta} = \boldsymbol{\theta}_{\sigma_1^2}$, the binormal density function evaluated

$$f(x_i, \boldsymbol{\theta}_{\sigma_1^2}) = \frac{p_*}{\sqrt{2\pi}\sigma_1} + \frac{1 - p_*}{\sqrt{2\pi}\sigma_{2,*}} \exp\left(-\frac{(\mu_1 - \mu_{2,*})^2}{2\sigma_{2,*}^2}\right).$$

Note that $f(x_i, \boldsymbol{\theta}_{\sigma_1^2})$ can be made arbitrarily large by making σ_1 arbitrarily small.

Since $L(\boldsymbol{\theta}_{\sigma_1^2}|\mathbf{x}) = \prod_{i=1}^n f(x_i, \boldsymbol{\theta}_{\sigma_1^2})$, and each $f(x_i, \boldsymbol{\theta}_{\sigma_1^2})$ is bounded away from zero (by virtue of p_*, $\mu_{2,*}$ and $\sigma_{2,*}^2$ being fixed), it follows that $l(\boldsymbol{\theta}_{\sigma_1^2}|\mathbf{x})$ can also be made arbitrarily large.

A further problem is that the parametrisation of the binormal model is not identifiable because **the role of the two distributions in the mixture can be swapped**. That is, the binormal distribution corresponding to parameters $(p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ is the same as that specified by parameters $(1 - p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

The unbounded likelihood and non-identifiability issues can be eliminated by **suitable restriction on the parameter space**. One possibility is to constrain **the ratio of the two standard deviations by requiring** that $0 < c < \frac{\sigma_1^2}{\sigma_2^2} < 1$, where c is some suitably small constant.

In practice, despite the unbounded likelihood and non-identifiability, a **sensible local maximum of the likelihood function** can often be found using **unconstrained numerical optimisation**. This is especially the case if there is **good separation between the two component normal distributions**, and the optimizer is given a starting value of $\boldsymbol{\theta}$ that is somewhere in the general vicinity of the local maximum. Ultimately, it is the shape of the likelihood function in the neighbourhood of this local maximum that is relevant to inference.

The binormal density function is a linear combination of the density functions given by $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributions.

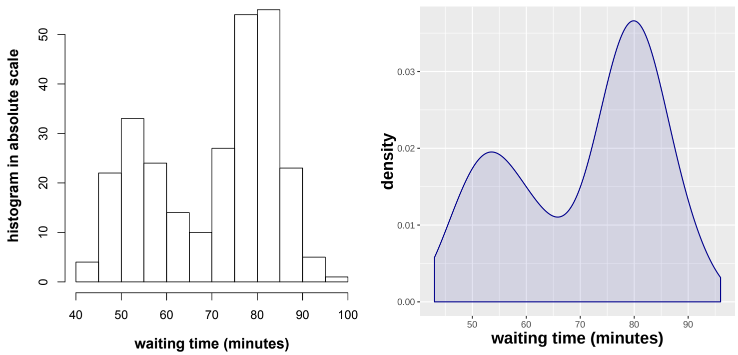

Figure: Mixture of two normal densities – data `faithful`

The histogram of waiting times shows that they look like a combination of (very roughly) 40 % from $N(52, 25)$ distribution and 60 % from $N(80, 25)$ distribution. The corresponding parameter values $\boldsymbol{\theta}^{(0)} = (0.4, 52, 80, 25, 25)^T$ would make good starting values for finding a local MLE using numerical optimisation. To estimate $\boldsymbol{\theta}$, use `optim()` function.

The call of `optim()` produced some warning messages (not shown), because it attempted to evaluate negative log-likelihood at parameter values outside of the parameter space (e.g. $\sigma_1, \sigma_2 < 0$). This can be avoided by using `lower` and `upper` bound arguments in the `optim()` call.

Example (Negative binomial 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, $M = \sum m_n = 647$. **Question**: Calculate theoretical frequencies $m_{n,E}$.

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
$m_{n,E}$	446	134	44	15	5	3

Likelihood function is defined as follows

$$L((\alpha, \pi)^T|\mathbf{x}) = \prod_{n=0}^{4} (\Pr(X = n))^{m_n} \left(1 - \sum_{n=0}^{4} \Pr(X = n)\right)^{m_{\geq 5}}$$

and logarithm of likelihood function

$$l((\alpha, \pi)^T|\mathbf{x}) = \sum_{n=0}^{4} m_n \ln \Pr(X = n) + m_{\geq 5} \ln \left(1 - \sum_{n=0}^{4} \Pr(X = n)\right).$$

Using numerical optimisation we get the following estimates $\widehat{\alpha} = 0.84$ and $\widehat{\pi} = 0.64$. Risk ratio $\widehat{\mu} = \frac{1-\widehat{\pi}}{\widehat{\pi}}\widehat{\alpha} = 0.47$.

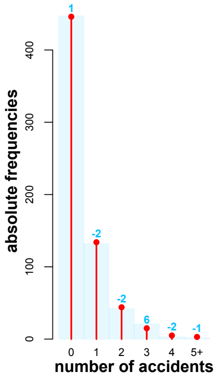

Figure: Comparison of observed and expected frequencies (negative binomial distribution)

Example (ZIP distribution; number of movements of a foetal lamb)

Let X be the number of movements of a foetal lamb in 240 five-second periods (Leroux and Puterman 1992), n be the number of movements, m_n be the number of periods with particular number of movements. **Question**: Calculate theoretical frequencies $m_{n,E}$ using Poisson and ZIP distribution.

Table: Observed and theoretical frequencies ($m_{n,O}$ and $m_{n,E}$) of five-second periods with n movements

n	0	1	2	3	4	5	6	7
$m_{n,O}$	182	41	12	2	2	0	0	1
$m_{n,E}$ (Poisson)	168	60	11	1	0	0	0	0
$m_{n,E}$ (ZIP)	182	37	16	4	1	0	0	0

Likelihood function is defined as follows

$$L((\lambda,p)^T|\mathbf{x}) = (p + (1-p)f(0,\lambda))^{m_0} \prod_{I(n>0)} ((1-p)f(n,\lambda))^{m_n}$$

and logarithm of likelihood function

$$l((\lambda,p)^T|\mathbf{x}) = m_0 \ln(p + (1-p)f(0,\lambda)) + \sum_{I(n>0)} m_n \ln((1-p)f(x,\lambda)).$$

For Poisson model, $\widehat{\lambda} = \frac{\sum_n n m_n}{\sum_n m_n} = \frac{86}{240} = 0.358$. For ZIP model, using numerical optimisation we get the following estimates $\widehat{\lambda} = 0.847$ a $\widehat{p} = 0.577$.

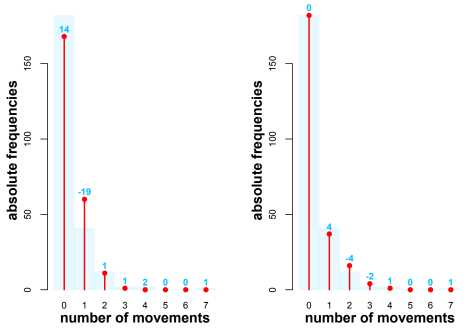

Figure: Comparison of observed and expected frequencies, Poisson distribution (left), ZIP distribution (right)