Engineering Mathematics 2B

Module 17: Statistics - Point Estimators

Nick Polydorides

School of Engineering

THE UNIVERSITY *of* EDINBURGH

Module 17 contents

Motivation

- Calculation and estimation

Theory

- Point statistical estimators
- Maximum likelihood estimator
- Properties of the estimator

Outcomes

Statistics: Extracting information from data

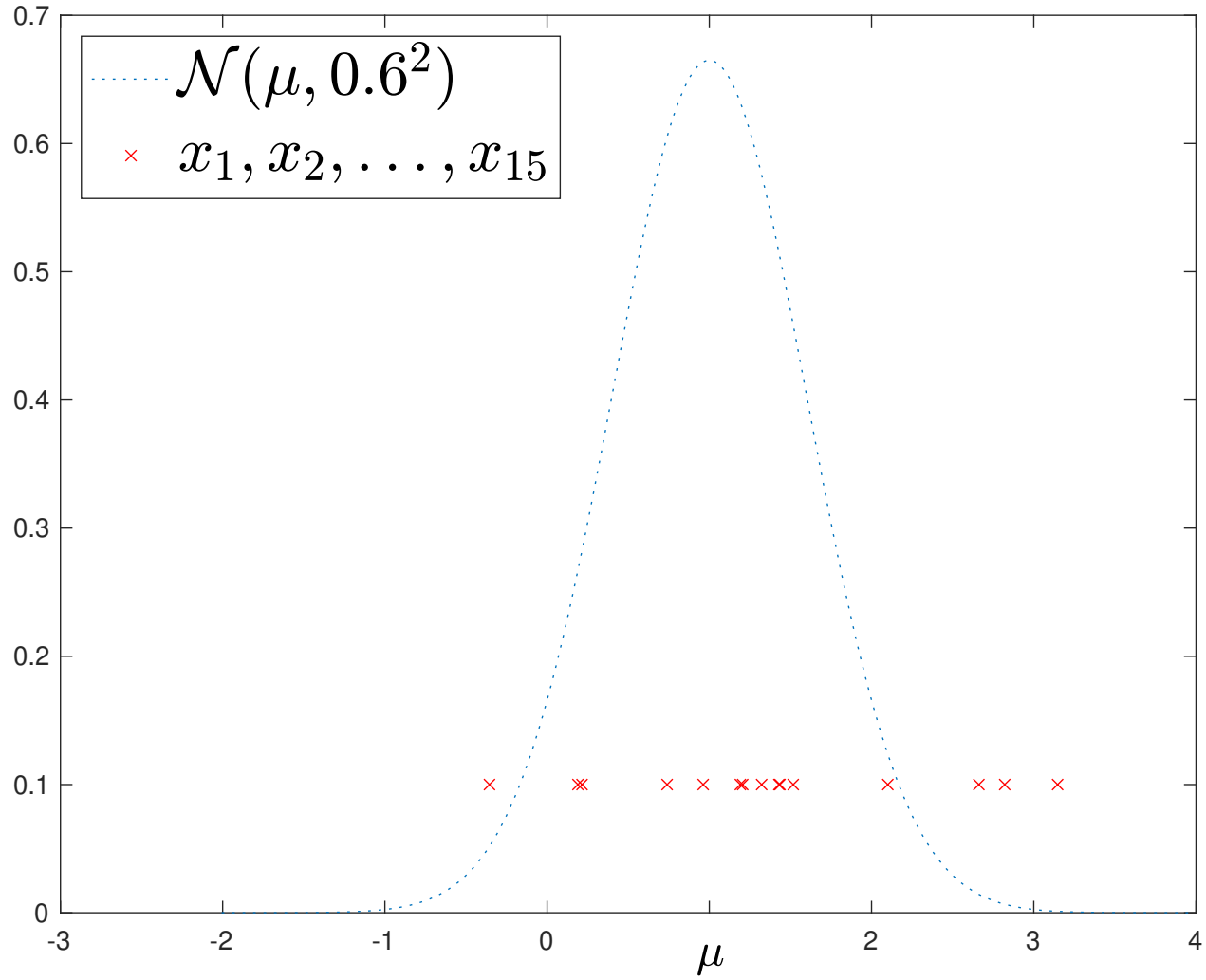
The goal in statistics:

"To estimate population models and their parameters from random variable realisations"

A classical paradigm: Estimate the mean of a random variable X that is distributed with $\mathcal{N}(\mu, \sigma^2)$ from data $x_1, \ldots, x_n$.

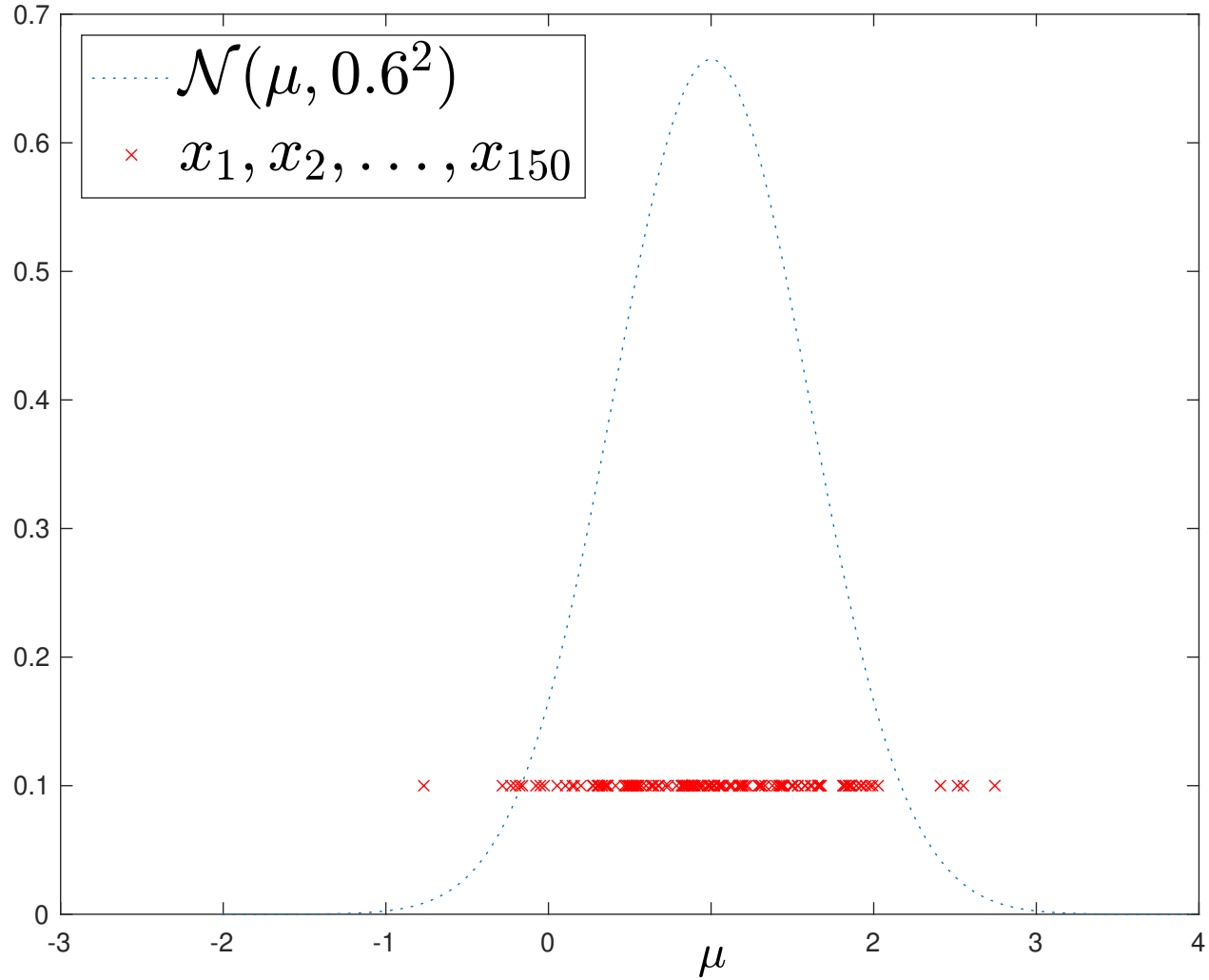
- The model: (=the mechanism that generates the data): Normal distribution. (assumed or known)
- The model parameters: mean (unknown), variance (known)
- The data: $x_1, \ldots, x_n$ (known - not random).

What's the μ in $\mathcal{N}(\mu, \sigma^2)$ of these data?



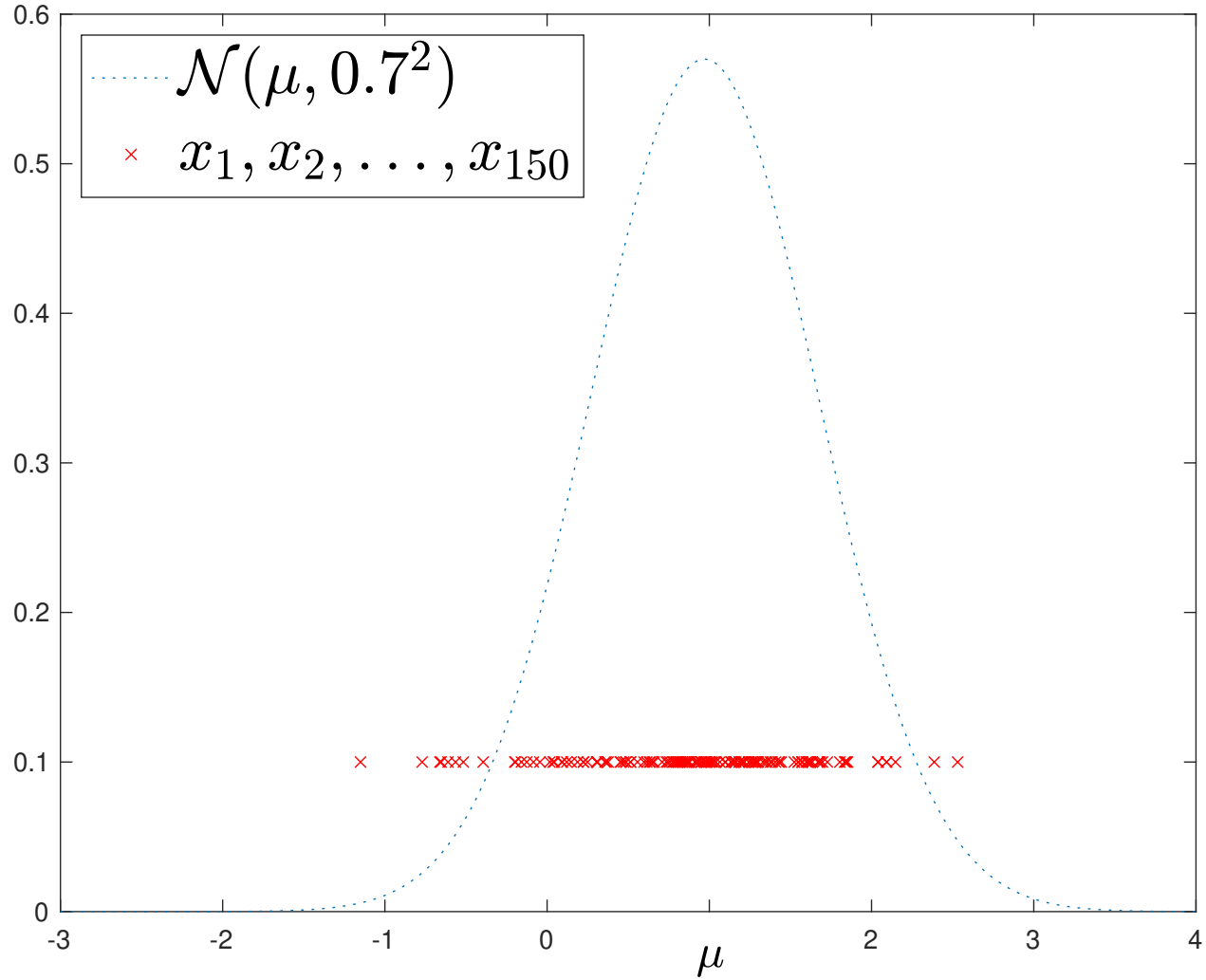
The unknown population mean is $\mu = 1$. The data mean is 1.37.

What's the μ in $\mathcal{N}(\mu, \sigma^2)$ of these data?



The unknown population mean is $\mu = 1$. The data mean is 0.9904.

What's the μ in $\mathcal{N}(\mu, \sigma^2)$ of these data? (Eye Test)



The unknown population mean is μ =??? The data mean is 0.9180.

Statistics: Extracting information from data

The underlying assumption in statistics is that we cannot sample the whole population.

E.g. *How many people in the UK have the flu today?* This can only be addressed in a statistical context as it is impractical to screen every individual in the UK.

Regarding answering such questions we must distinguish between:

- **Estimation**: Answering a question using a subset of the necessary information. We know partly the model for the data. Our answer will always have some uncertainty.
- **Calculation**: Answering a question using all necessary information. We know completely the model for the data. Our answer is deemed exact, i.e. with absolute certainty.

Calculation Vs Estimation

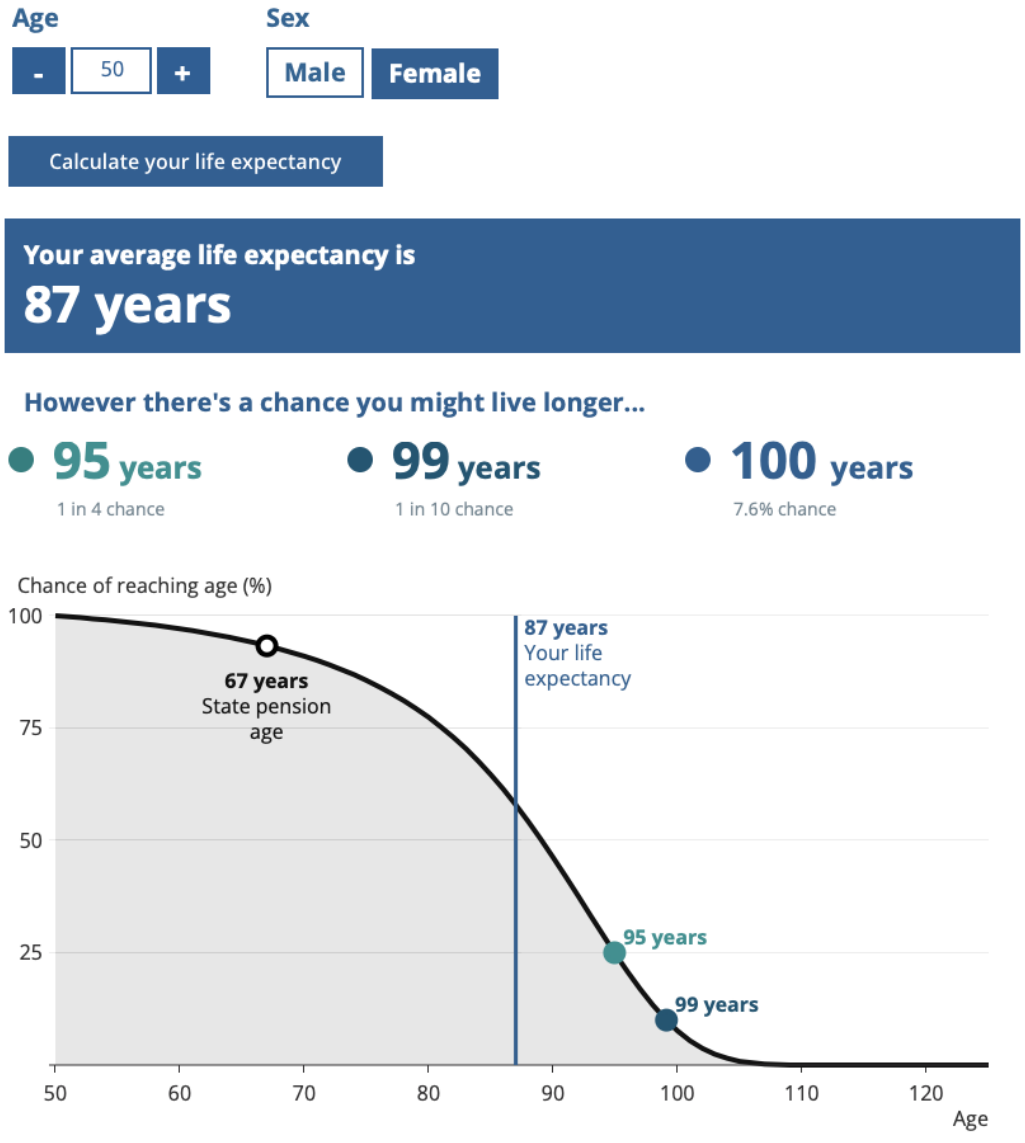
"*A homogeneous metallic sphere was measured to be of radius 5 cm and mass 3 kg. The sphere is dropped from a height of 10 m where it is at rest. Calculate its velocity when hitting the ground.*"

This is a calculation problem because the models are known from physics, i.e. Newton's laws of motion, e.g. $F = mg$ and $g = \dot{v}$, $v = s/t$, etc

"*Consider a 50 year old female born and living in Scotland. The person is in full time employment and has no medical conditions. Estimate how long will they live.*"

There are no deterministic laws of biology that can be used here. Only statistical data. Besides, the answer will depend on many other aspects of lifestyle.

Life expectancy calculator



Courtesy: www.ons.gov.uk

Statistical estimators

A **statistic** is a quantity calculated from the sampled data. An **estimator** for the parameter θ of the distribution is a statistic that approximates the true value of the parameter.

A statistic is always a random variable, as it depends on the particular sample. Consequently all parameter estimators are random or uncertain. (recall Sum_n and $\bar{X}_n$ are random).

Types of statistics

1. **Point estimators** (point statistics) are single valued quantities estimated from the data.

2. **Interval estimators** (range statistics) are closed intervals whose limits can be estimated from the data.

Notice: An estimate is the realisation value of an estimator which is random.

Mathematical framework

Say we have observed the realisations of n identical variables

$$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$$

where

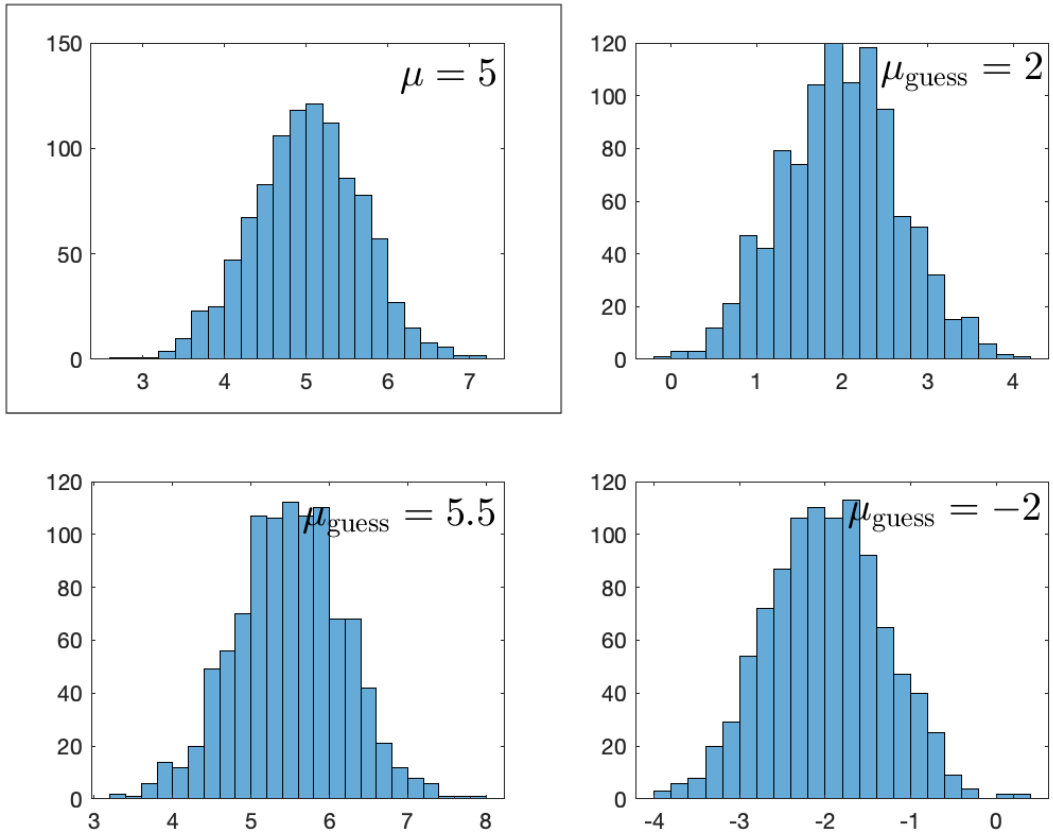
$$X_i \sim p_{X_i}(x_i), \quad \text{with} \quad \mathbb{E}[X_i] = \mu.$$

Consider the case where the mean μ is unknown. Say we denote it by θ.

It seems reasonable to search for a θ that is likely to yield another set of n data similar to those observed (under the true μ). In other words, we would like to fix the parameter θ such that

$$\text{Likelihood}(X_1 = x_1, \ldots, X_n = x_n | \theta) \rightarrow \text{maximum}$$

Data likelihood



Histograms (data bins) for a thousand data samples: top right as generated with true mean $\mu = 5$, and then three other guesses. By visual inspection the prediction with $\mu_{\text{guess}} = \theta = 5.5$ overlaps more with the actual data.

Mathematical framework

Mathematically we need the likelihood function for all observed data $X_1 = x_1, \ldots, X_n = x_n$ as

$$L(x_1, \ldots, x_n \mid \theta) = p_{X_1,\ldots,X_n}(x_1, \ldots, x_n \mid \theta)$$

where $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ is the **joint probability density of $X_1, \ldots, X_n$ evaluated at the observed values $x_1, \ldots, x_n$.**

If the variables are independent then

$$L(x_1, \ldots, x_n \mid \theta) = p_{X_1,\ldots,X_n}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} p_{X_i}(x_i|\theta)$$

and we can express the likelihood from the parametrically known $P_{X_i}(x_i|\theta)$. Moreover, if the rv are identical then $P_{X_i}(x_i|\theta) = P_X(x|\theta)$ for all $i = 1, \ldots, n$.

The maximum likelihood estimator (MLE)

The MLE defines a useful estimator by picking the **point estimator that maximises the data likelihood**.

A realisation of the ML estimator (the ML estimate) can be computed by solving the optimisation problem

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(x_1, \ldots, x_n \mid \theta),$$

where Θ is the set of admissible values for the unknown parameter θ, e.g. the set of reals $\mathbb{R}$.

A **maximum likelihood estimator (MLE)** exists whenever the above optimisation problem has a solution. If the likelihood is concave then it has a maximum allowing to locate it by solving the first-order optimality condition

$$\boxed{\frac{\partial}{\partial \theta} L(x_1, \ldots, x_n \mid \theta) = 0}$$

MLE for the Binomial

We observe n iid Bernoulli trials with probability of success p, and collect data $X = x$ the number of successes. We would like the MLE of $\theta := p$. The likelihood function is

$$L(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad 0 < \theta < 1 \doteq \Theta$$

The log-likelihood function is

$$\log L(x|\theta) = \log \binom{n}{x} + x \log \theta + (n-x) \log(1-\theta)$$

Setting $\frac{\mathrm{d}}{\mathrm{d}\theta} \log L(x|\theta) = 0$ yields

$$\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \quad \Rightarrow \quad \hat{\theta}_{\mathrm{MLE}} = \frac{x}{n}$$

and this is at a maximum as the second derivative there is negative.

MLE for the Exponential

Consider iid $X_1, X_2, \ldots, X_n$ where $X_i \sim p_{X_i}(x_i) = \lambda \mathrm{e}^{-\lambda x_i}$, $0 \leq x_i < \infty$, and $\lambda > 0$. With data $X_1 = x_1, \ldots, X_m = x_m$, we would like the MLE of $\lambda := \theta$. The likelihood function is

$$L(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \theta \mathrm{e}^{-\theta x_i} = \theta^n \mathrm{e}^{-\theta(x_1+\cdots+x_n)}, \quad \theta > 0$$

The log-likelihood function is

$$\log L(x_1, \ldots, x_n|\theta) = n \log \theta - \theta \sum_{i=1}^{n} x_i$$

Setting $\frac{\mathrm{d}}{\mathrm{d}\theta} \log L(x_1, \ldots, x_n|\theta) = 0$ yields

$$\frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \quad \hat{\theta}_{\mathrm{MLE}} = \frac{n}{\sum_{i=1}^{n} x_i}$$

which is positive by virtue of $x_i > 0$ and the point is at a maximum as the second derivative is negative.

Computing the MLE of the Gaussian mean

We have data $x_1, \ldots, x_n$ from iid $X_1, \ldots, X_n$ where $X_i \sim \mathcal{N}(\mu, \sigma^2)$, and we want the MLE of μ. In effect,

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(x_1, \ldots, x_n \mid \theta), \quad \Theta \doteq \mathbb{R}$$

As in this case θ is a single parameter, to find the **maximum of the likelihood function** if it exists, we set

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\theta} L(x_1, \ldots, x_n \mid \theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\Big\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\theta})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\theta})\Big\} \\ &= 0 \end{aligned}$$

where $\boldsymbol{x} = \begin{bmatrix} x_1 & \ldots & x_m \end{bmatrix}^\top$, $\boldsymbol{\theta} = \begin{bmatrix} \theta & \ldots & \theta \end{bmatrix}^\top$ and $\Sigma = \sigma^2 I$, for I the $n \times n$ identity matrix.

Computing the MLE of the Gaussian mean

Taking the constant (θ-independent) term out of the exponential we simplify as

$$\frac{\mathrm{d}}{\mathrm{d}\theta} L(\cdot \mid \theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \Big\{ \text{constant} \cdot \exp\Big\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\theta})^\top (\boldsymbol{x} - \boldsymbol{\theta}) \Big\} \Big\} = 0$$

and recall that we still have just a single unknown parameter to estimate, i.e. $\boldsymbol{\theta}$ is a vector padded with θ.

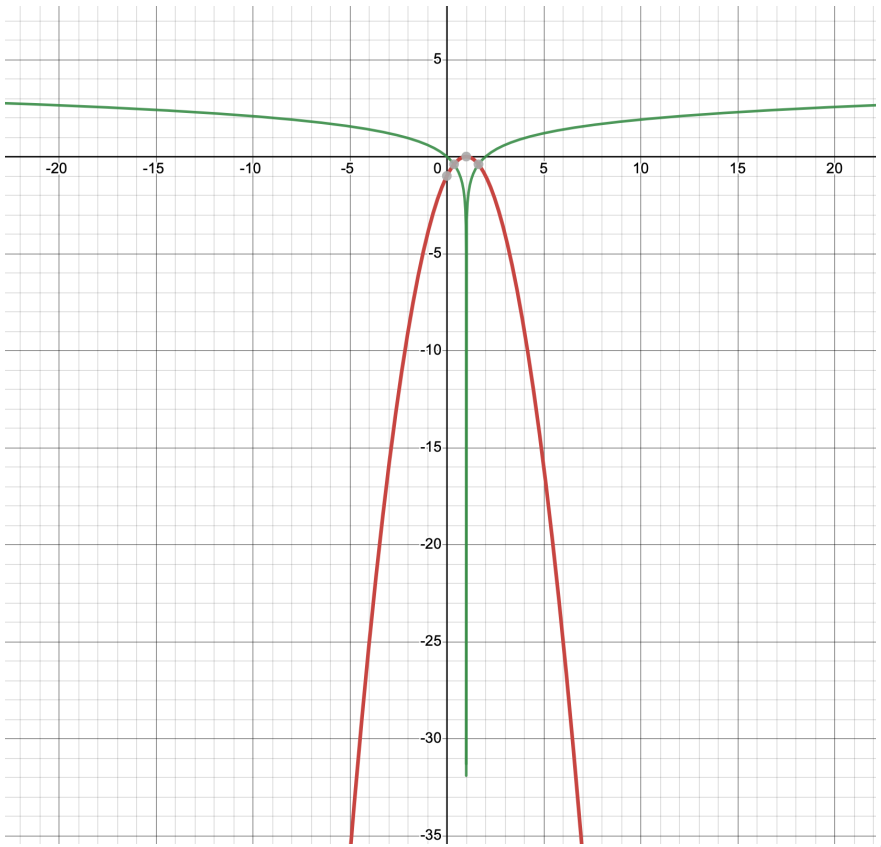
Notice: $\hat{\theta}_{\mathrm{MLE}}$ is also the **minimum of the negative logarithmic likelihood**

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \mathbb{R}} L(\cdot \mid \theta) = \arg\max_{\theta \in \mathbb{R}} \log L(\cdot \mid \theta) = \arg\min_{\theta \in \mathbb{R}} - \log L(\cdot \mid \theta),$$

and log simplifies the exponential term in L.

Converting maximisation to minimisation

If a function $L(\theta)$ is concave with $\frac{\mathrm{d}L}{\mathrm{d}\theta}(\theta)$ attaining a zero at a point, then this point is a global maximum, say at $\theta = \theta_{\max}$. In effect, $-L(\theta)$ is convex with a global minimum at $\theta = \theta_{\max}$. The natural logarithm of a function, if defined, preserves the *location* of the minima and maxima, e.g. $f(\theta) = -(\theta - 1)^2$ (red), and $\log(-f(\theta))$ (green).



Computing the MLE of the Gaussian mean

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \arg\min_{\theta\in\mathbb{R}} -\text{constant}\cdot\exp\Big\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\theta})^\top(\boldsymbol{x}-\boldsymbol{\theta})\Big\}\\ &= \arg\min_{\theta\in\mathbb{R}} -\text{constant}\cdot\frac{1}{2}\big(\boldsymbol{x}^\top\boldsymbol{x}-2\boldsymbol{x}^\top\boldsymbol{\theta}+\boldsymbol{\theta}^\top\boldsymbol{\theta}\big)\\ &= \arg\min_{\theta\in\mathbb{R}} -\text{constant}\cdot\frac{1}{2}\Big(\sum_{j=1}^n x_j^2-2\theta\sum_{j=1}^n x_j+n\theta^2\Big)\end{aligned}$$

Setting the derivative of the parenthesis above to zero yields

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\Big(-\theta\sum_{j=1}^n x_j\Big)+\frac{\mathrm{d}}{\mathrm{d}\theta}\Big(\frac{1}{2}n\theta^2\Big)=0\;\Rightarrow\;-\sum_{j=1}^n x_j+n\theta=0,$$

$$\therefore\quad\hat{\theta}_{\text{MLE}}=\frac{1}{n}\sum_{j=1}^n x_j.$$

The MLE of the Gaussian mean is equal to the data mean.

How "good" is the estimator?

How do we know how good (accurate) our estimator, MLE or otherwise, is?

If we have multiple estimators how do we choose the "best" between them?

A principled way is by using the **Mean Squared Error (MSE)**, the expected squared difference between $\hat{\theta}$ and the true θ

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

where the bias of the estimator is

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

and its variance is

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

Bias & Variance

Bias is the **expected** difference between the parameter and its estimator.

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta} - \mathbb{E}[\theta]] = \mathbb{E}[\hat{\theta}] - \theta.$$

Over repeated samples, an unbiased estimator converges to the true value on average. An estimator is said to be **unbiased** if its bias is zero.

Between unbiased estimators, we prefer estimators that have **small variance**. This implies, that under repeated sampling, estimators with lower variance are more likely to be closer to the true θ.

Standard error

The quantity $\sqrt{\mathrm{Var}(\hat{\theta})}$ is called the **standard error (SE)** of the estimator

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})} = \sqrt{\mathbb{E}\big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\big]}$$

and gives the square root of the variance with which $\hat{\theta}$ is distributed.

Example: Let $\hat{\theta}$ be the MLE for the mean μ of X when $X \sim \mathcal{N}(\mu, \sigma^2)$, obtained from a large iid sample $x_1, \ldots, x_n$ as $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$. By CLT, we know $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, and since $\hat{\theta}$ is a realisation of the random variable $\bar{X}_n$, hence $\mathrm{Var}(\hat{\theta}) = \frac{\sigma^2}{n}$. In effect,

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

MLE of Gaussian model parameters

In fitting iid data to a normal distribution we may need to estimate both $\theta_1 := \mu$ and $\theta_2 := \sigma$. Set $\boldsymbol{\theta} = \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}^\top$. Suppose we have iid data $x_1, \ldots, x_n$ from

$$\mathcal{N}(\mu, \sigma^2) = p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}},$$

then like before, the data likelihood conditioned on the parameters is

$$\begin{aligned} L(x_1, \ldots, x_n \mid \boldsymbol{\theta}) &= \text{constant} \cdot \exp\Big\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\theta})^\top (\boldsymbol{x} - \boldsymbol{\theta})\Big\} \\ &= \prod_{i=1}^{n} p_{X_i}(x_i \mid \boldsymbol{\theta}) \\ &= \frac{1}{\left(\sqrt{2\pi}\theta_2\right)^n} \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{\theta_2^2}\right\} \end{aligned}$$

the joint PDF of n Gaussians with mean θ_1 and variance θ_2.

MLE of Gaussian model parameters

The negative log likelihood yields

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^2} -\log L(x_1, \ldots, x_n \mid \boldsymbol{\theta})$$

but now we have to use the gradients in the θ_1 and θ_2 directions separately

$$\frac{\partial}{\partial\theta_1}\left\{-\log\left(\frac{1}{\left(\sqrt{2\pi}\theta_2\right)^n}\exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^n(x_i-\theta_1)^2}{\theta_2^2}\right\}\right)\right\} = 0,$$

and

$$\frac{\partial}{\partial\theta_2}\left\{-\log\left(\frac{1}{\sqrt{2\pi}\theta_2}\exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^n(x_i-\theta_1)^2}{\theta_2^2}\right\}\right)\right\} = 0.$$

MLE of Gaussian

Taking the derivatives of the logs (after some algebra) leads to

$$\hat{\theta}_{1\text{MLE}} = \hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\theta}_{2\text{MLE}} = \hat{\sigma}_{\text{MLE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}_{\text{MLE}}\right)^2}$$

As we do not know θ we can use its sample-based estimator instead, which makes $\hat{\sigma}_{\text{MLE}}$ *biased*. An alternative, unbiased estimator for σ^2 is the sample mean

$$\hat{\sigma}^2_{\text{sample mean}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_{\text{MLE}})^2$$

Notice: Notice that the square root of $\hat{\sigma}^2_{\text{sample mean}}$ is not an unbiased estimator of the population's standard deviation.

MLE of Gaussian (derivation)

To verify that the $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is at a maximum we can check the eigenvalues of the Hessian, evaluated at that point. (Recall definitions of convex and concave functions from vector calculus)

Criterion: The point $\hat{\boldsymbol{\theta}}_{\text{MLE}} = (\hat{\mu}_{\text{MLE}}, \hat{\sigma}_{\text{MLE}})$ is at a maximum if the Hessian is negative definite, i.e. it has two negative eigenvalues. This is equivalent to requiring its determinant to be positive and either of the diagnoals to be negative.

Recalling the likelihood

$$L(\boldsymbol{x}|\theta_1, \theta_2) = \frac{1}{\left(\sqrt{2\pi}\theta_2\right)^n} \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^n (x_i - \theta_1)^2}{\theta_2^2}\right\}$$

then taking the log gives

$$f(\theta_1, \theta_2) := \log L(\boldsymbol{x}|\theta_1, \theta_2) = -n\log(\sqrt{2\pi}\theta_2) - \frac{1}{2}\frac{\sum_{i=1}^n (x_i - \theta_1)^2}{\theta_2^2}$$

MLE of Gaussian (derivation)

Differentiating we get

$$\frac{\partial f}{\partial \theta_1} = \frac{1}{\theta_2^2}\sum_{i=1}^{n}(x_i - \theta_1), \quad \frac{\partial f}{\partial \theta_2} = -\frac{n}{\theta_2} + \frac{1}{\theta_2^3}\sum_{i=1}^{n}(x_i - \theta_1)^2$$

$$\frac{\partial^2 f}{\partial \theta_1^2} = -\frac{n}{\theta_2^2}, \quad \frac{\partial^2 f}{\partial \theta_2^2} = \frac{n}{\theta_2^2} - \frac{3}{\theta_2^4}\sum_{i=1}^{n}(x_i - \theta_1)^2$$

$$\frac{\partial^2 f}{\partial \theta_1 \partial \theta_2} = -\frac{2}{\theta_2^3}\sum_{i=1}^{n}(x_i - \theta_1) = \frac{\partial^2 f}{\partial \theta_2 \partial \theta_1}$$

Substituting for $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ gives

$$\frac{\partial^2 f}{\partial \theta_1 \partial \theta_2}(\hat{\boldsymbol{\theta}}_{\text{MLE}}) = -\frac{2}{\hat{\sigma}^3_{\text{MLE}}}\sum_{i=1}^{n}(x_i - \hat{\mu}_{\text{MLE}}) = -\frac{2}{\hat{\sigma}^3_{\text{MLE}}}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

which comes up to be zero since $\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i) = 0$.

MLE of Gaussian (derivation)

The Hessian at the MLE is thus

$$H(\hat{\boldsymbol{\theta}}_{\text{MLE}}) = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2_{\text{MLE}}} & 0 \\ 0 & \frac{n}{\hat{\sigma}^2_{\text{MLE}}} - \frac{3\sum_{i=1}^n (x_i - \hat{\mu}_{\text{MLE}})^2}{\hat{\sigma}^4_{\text{MLE}}} \end{pmatrix}$$

Entering the value of $\hat{\sigma}_{\text{MLE}}$ leads to

$$H(\hat{\boldsymbol{\theta}}_{\text{MLE}}) = \begin{pmatrix} -\frac{n^2}{\sum_{i=1}^n (x_i - \hat{\mu}_{\text{MLE}})^2} & 0 \\ 0 & -\frac{2n^2}{\sum_{i=1}^n (x_i - \hat{\mu}_{\text{MLE}})^2} \end{pmatrix}$$

so the determinant of the Hessian is

$$|H(\hat{\boldsymbol{\theta}}_{\text{MLE}})| = \frac{2n^4}{\left(\sum_{i=1}^n (x_i - \hat{\mu}_{\text{MLE}})^2\right)^4} > 0$$

with the two diagonals strictly negative, hence $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is at maximum.

Formulas

- The likelihood function for data $X_1 = x_1, \ldots, X_n = x_n$ is

$$L(x_1, \ldots, x_n \mid \theta) = p_{X_1,\ldots,X_n}(x_1, \ldots, x_n \mid \theta)$$

  where $p_{X_1,\ldots,X_n}$ is the joint density.
- The mean squared error of estimate $\hat{\theta}$

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathrm{Bias}(\hat{\theta})^2 + \mathrm{Var}(\hat{\theta})$$

- The bias of $\hat{\theta}$

$$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

- The variance of $\hat{\theta}$

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

Main outcomes of module 17

You MUST know:

1. The purpose of statistics
2. The MLE point estimator for the parameters of the normal, as well as uniform, Bernoulli and binomial distributions.
3. The MSE, bias and variance of estimators
4. The standard error of an estimator

Good to know:

To plot histograms of data in Python, R or Matlab.