Maximum likelihood Estimation

Content

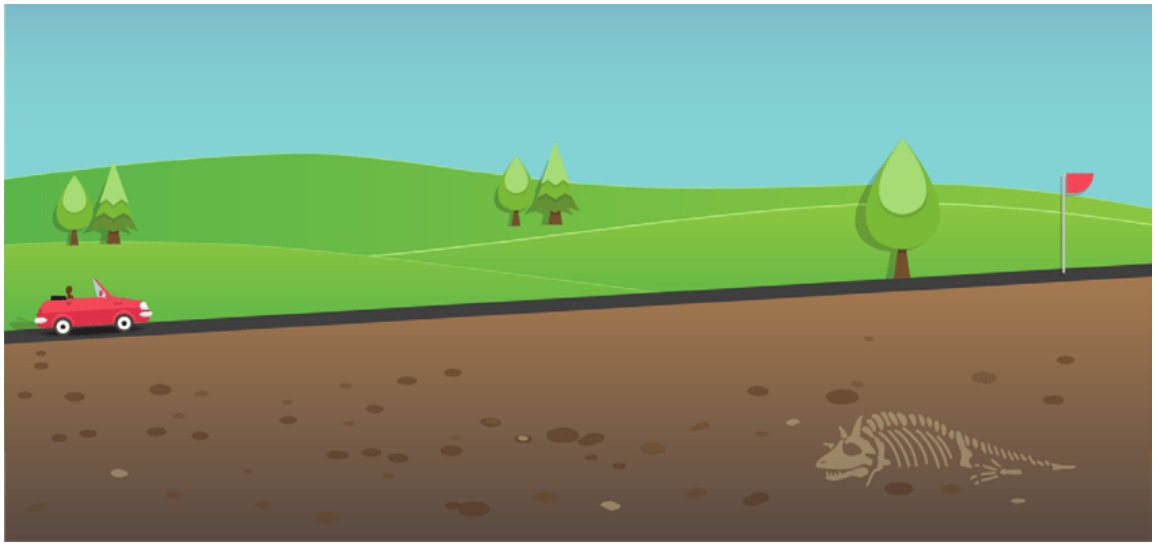
- Introduction to MLE methods
  - Introduction
  - Model estimation
  - Maximum likelihood method
  - Other methods
  - Comparing estimators
- Likelihood ratio test and applications

INTRODUCTION

Models: Deterministic vs Stochastic

Deterministic

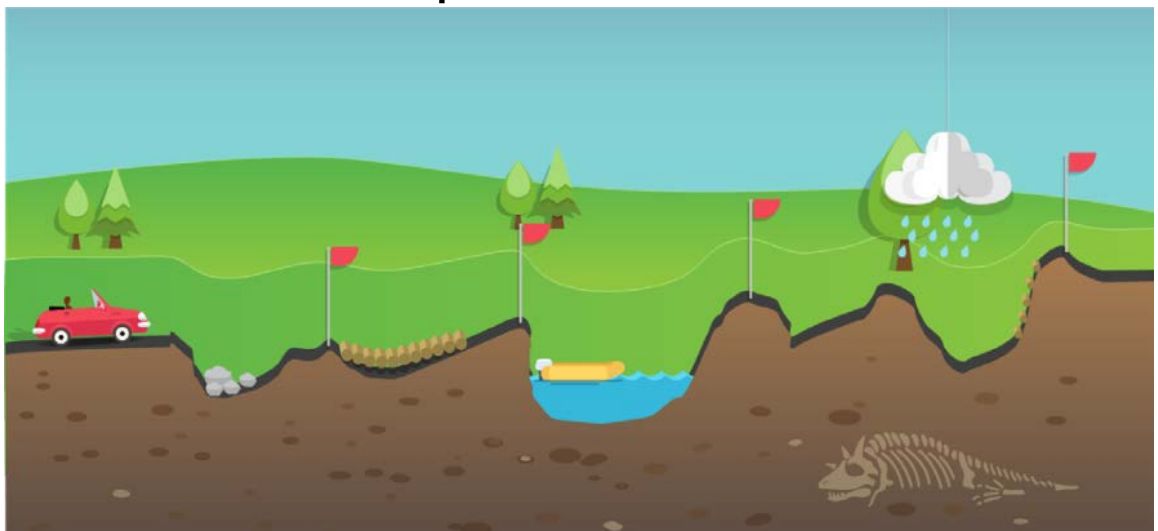
- *Lack of free will, the opposite of random.*
- A **Deterministic Model** imply the calculation of a future event **exactly**, without uncertainty (the involvement of randomness). If something is deterministic, you have all of the data necessary to predict (determine) the exact outcome.



Models: Deterministic vs Stochastic

Stochastic

- *Having a random probability distribution or pattern that may be analyzed statistically but may not be predicted precisely.*
- A **Stochastic Model** has the capacity to **handle uncertainties** in the inputs applied. Stochastic models possess some inherent randomness --> the same set of parameter values and initial conditions will lead to an ensemble of different outputs.



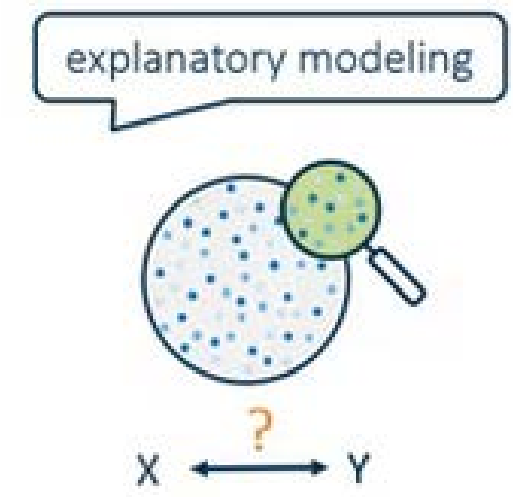
Models: Explanatory vs Predictive

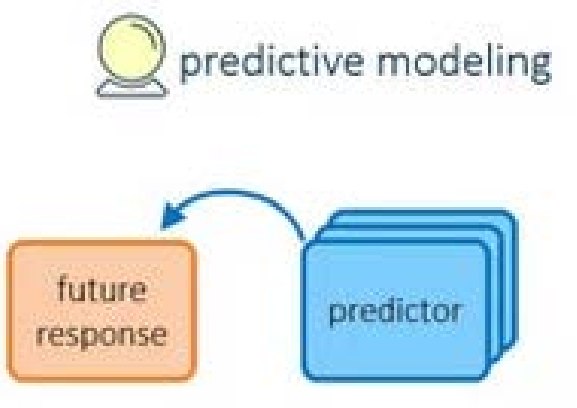
(Generalizations, true in most situations)

- **Explanatory models** are retrospective
- **Predictive models** are prospective

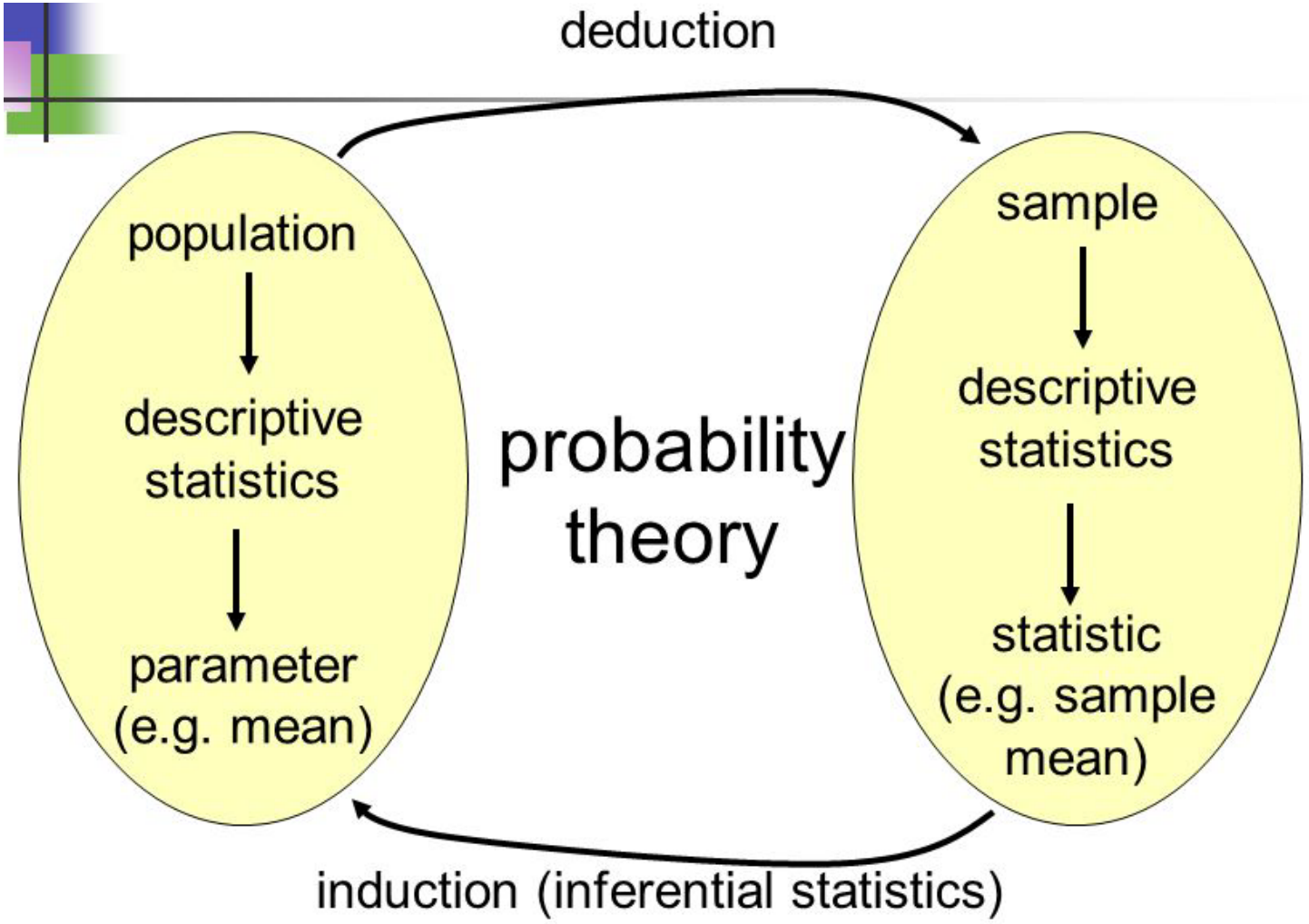
- **Explanatory models** focus on minimizing bias (fits the data closely)
- **Predictive models** focus on minimizing both bias and estimation variance (accurately predict new cases)

- **Explanatory models** data is used to estimate the model
- **Predictive models** data is divided into training set (to estimate the model) and validation set (assess the model performance)



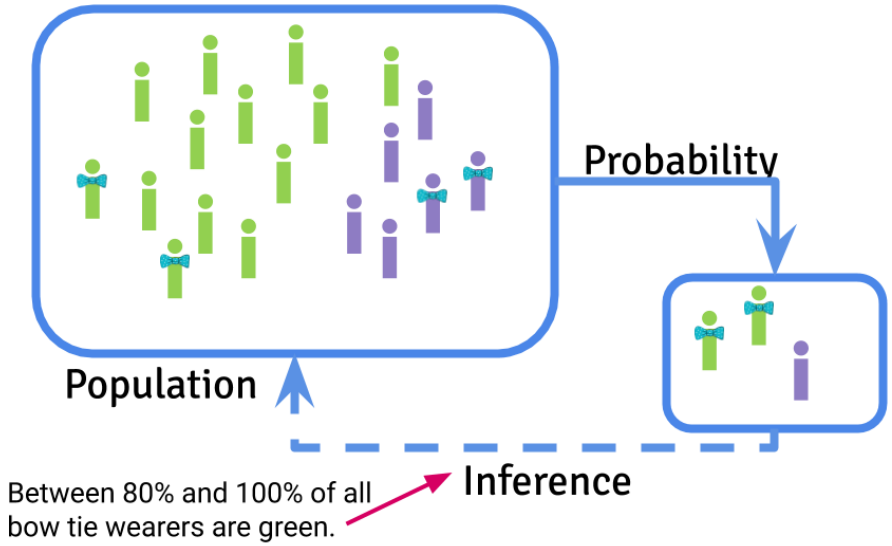


Statistical Induction and Deduction principle



Statistical Inference

- GOAL: to report sources of "signal" in a data set, while documenting and accounting for both systematic and "random" sources of errors.
- Using probability to extract a sample from a population, take measurements on that sample, and use the samples to infer something about the characteristics of the population *on average*.
- Example: Proportion of bow tie wearers in thee population using a sample



Population and Sample

- A population is the entire group that you want to draw conclusions about
- A sample is the specific group that you will collect data from
- The size of the sample is always less than the total size of the population
- In statistics/research:
  - the populations are often assumed infinite
  - a population does not always refer to people

Examples

- A survey
  - To know the % of households that recycles the trash in Catalonia
    - Population: all households in Catalonia
    - Sample: 100 selected individuals from the population

- A laboratory cooking experiment realization
  - We wish to know the temperature our cookies dough in a particular time of cooking.
  - We measure the temperature in a laboratory experiment, and we repeat the measurements many times, say 250 times. We could use the mean of all these measurements to estimate the temperature of the dough at a particular time
    - Population: all experiments we could possibly perform (conceptual, and infinite!)
    - Sample: 250 measurements.

Parameters, estimators

The characteristic of the population you are estimating is called a *parameter* and you use a *statistical estimate* to try to guess what that parameter might be. You can then use the information you have about sources of uncertainty to infer how accurate and precise you think your estimate will be.

- Parameters are fixed, unknown quantities that specify the population.
- Any number you compute using some sample of data is called a statistic.
- Statistics are random variables whereas parameters are not (unless you are a Bayesian).
- Estimators are statistics that are used to estimate the unknown parameters.

Notation:

- $\hat{\mu} = \overline{x}$ with μ the true population mean.
- $\hat{\sigma} = s$ with σ the true population standard deviation.

MODEL ESTIMATION

Point estimate and Interval estimate

Statistical models have unknown parameters that need to be estimated

- Point estimation: Estimation done with a single value computed from a sample
- Interval estimation: Estimation done with a range of plausible values computed from a sample

Examples:

- A sample mean is a point estimate of a population mean
- An interval estimate gives you a range of values where the parameter is expected to lie. A confidence interval is the most common type of interval estimate.

Methods for obtaining point estimators

- Maximum likelihood (ML) method
- Method of moments (MM)
- Bayesian methods

MAXIMUM LIKELIHOOD METHOD

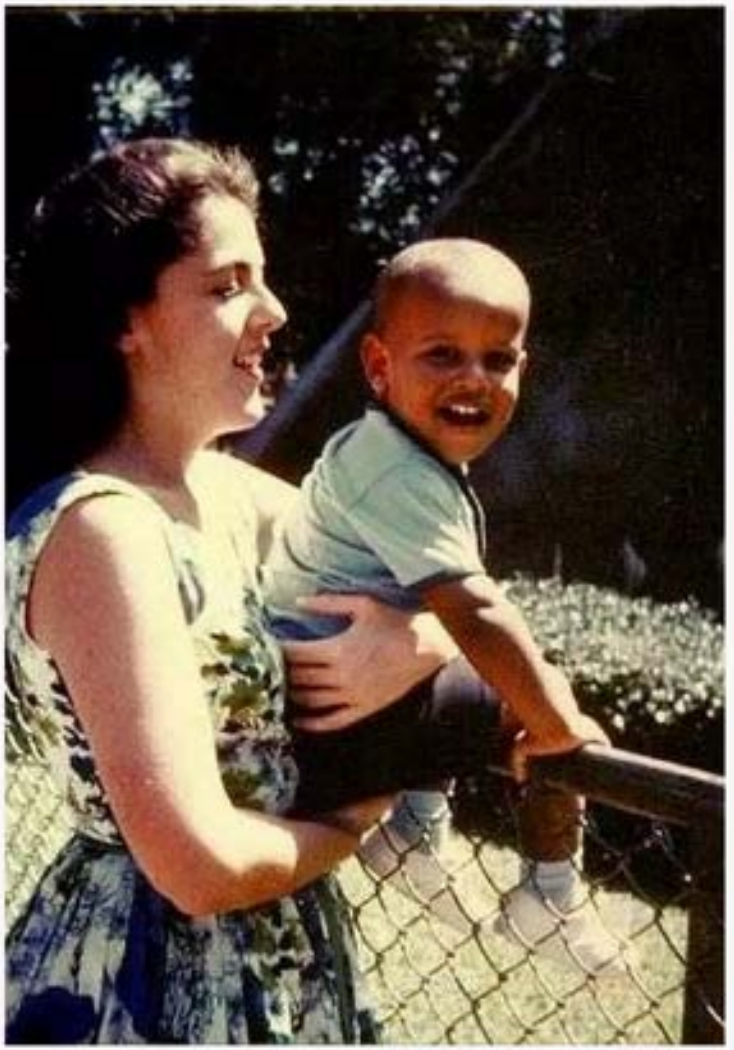
Maximum likelihood method (MLM)

Objective

- To introduce the idea of working out the most likely cause of an observed result by considering the likelihood of each of several possible causes and picking the cause with the highest likelihood

Who is the daddy?

- Here is a picture of a boy and his mother…



Source: https://www.slideshare.net/maxchipulu/the-method-of-maximum-likelihood

Which of the following men, do you think, is the child's father?

The logic of Maximum likelihood

- The child appears to be of mixed race parentage
- His mother is white
- Therefore, of the four possible daddies, daddy A is the least likely
- Daddy C is the next least likely because of the child's appearance
- This leaves B or D; but which one did you go for?
- If only we had more information, e.g. blood types or DNA or history of the two men, we could be more exact in our view of which of the two is more likely....
- But we could never be certain that anyone's daddy is the real daddy! We simply accept the most likely choice. Every daddy is an MLM daddy!

Solution

- The child in the picture is President Barack Obama and his mother
- Daddy D is the actor Sydney Poitier
- C is the actor Bruce Lee
- A is Professor Alan Agresti- Statistics Icon
- B is Barack Obama, Sr,

Example

Suppose you flip a coin independently 10 times, and you see

HTTTHHTHHH

What is your estimate of the probability the coin comes up heads?

a) $2/5$
b) $1/2$
c) $3/5$
d) $55/100$

Maximum likelihood estimation

Idea: We obtained the results we got.

High probability events happen more often than low probability events.

So, guess the rules that maximize the probability of the events we saw (relative to other choices of the rules).

Since that event happened, might as well guess the set of rules for which that event was most likely.

See Statistical models and stochastic processes. Course 2022-2023 notes.

File T1_MLEstimation.pdf

Slide 10 to 24

Two parameters 25-28

Special cases and some definitions and properties: 30-41

Maximum Likelihood Estimation

- Formally, we are trying to estimate a parameter of the experiment (here: the probability of a coin flip being heads).
- The likelihood of an event E given a parameter θ is

$\mathcal{L}(E;\theta)$, this is $\mathbb{P}(E)$ when the experiment is run with θ

We will use the notation $\mathbb{P}(E;\theta)$ for probability when run with parameter θ where the semicolon means "extra rules" rather than conditioning

- We will choose $\hat{\theta} = \text{argmax}_\theta \, \mathcal{L}(E;\theta)$
- argmax is the argument that produces the maximum so the θ that causes $\mathcal{L}(E;\theta)$ to be maximized.

Notation comparison

- $\mathbb{P}(X|Y)$ probability of X, conditioned on the **event** Y having happened (Y is a subset of the sample space)
- $\mathbb{P}(X;\theta)$ probability of X, where to properly define our probability space we need to know the extra piece of information θ. Since θ is not an event, this is not conditioning
- $\mathcal{L}(X;\theta)$ the likelihood of event X, given that an experiment was run with parameter θ. Likelihoods do not have all the properties we associate with probabilities (e.g. they do not all sum up to 1) and this is not conditioning on an event (θ is a parameter/rule of how the event could be generated).

MLE

Maximum Likelihood Estimator

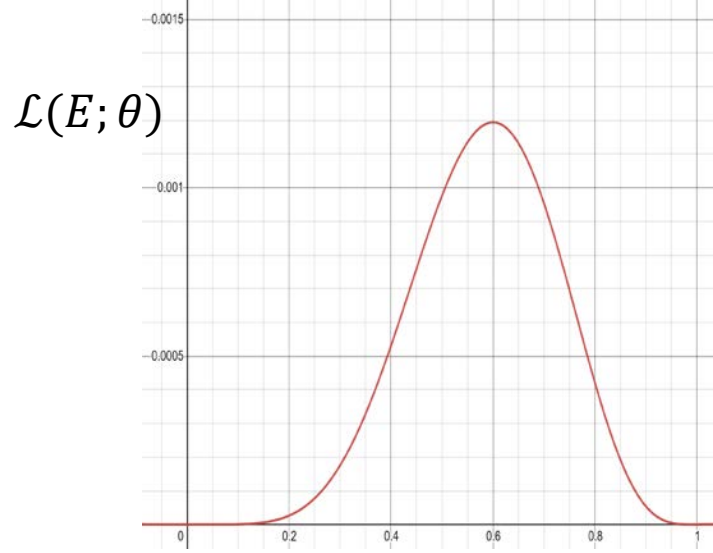
The maximum likelihood estimator of the parameter θ is:

$$\hat{\theta} = \text{argmax}_{\theta} \ \mathcal{L}(E; \theta)$$

θ is a variable, $\hat{\theta}$ is a number (or formula given the event).

We will also use the notation $\hat{\theta}_{\text{MLE}}$ if we want to emphasize how we found this estimator.

The Coin Example



$\mathcal{L}(\text{HTTTHHTHHH}\,;\theta) = \theta^6(1-\theta)^4$

Where is θ maximized?

How do we usually find a maximum?

Calculus!!

$$\frac{d}{d\theta}\theta^6(1-\theta)^4 = 6\theta^5(1-\theta)^4 - 4\theta^6(1-\theta)^3$$

Set equal to 0 and solve

$$6\hat{\theta}^5(1-\hat{\theta})^4 - 4\hat{\theta}^6(1-\hat{\theta})^3 = 0 \Rightarrow 6(1-\hat{\theta}) - 4\hat{\theta} = 0 \Rightarrow -10\hat{\theta}$$
$$= -6 \Rightarrow \hat{\theta} = \frac{3}{5}$$

The Coin Example

- For this problem, θ must be in the closed interval $[0,1]$. Since $\mathcal{L}()$ is a continuous function, the maximum must occur at and endpoint or where the derivative is 0.

- Evaluate $\mathcal{L}(\cdot;0) = 0, \mathcal{L}(\cdot;1) = 0$
- at $\theta = 0.6$ we get a positive value, so $\theta = 0.6$ is the maximizer on the interval $[0,1]$.

Maximizing a function

CLOSED INTERVALS

- Set derivative equal to 0 and solve.
- Evaluate likelihood at endpoints and any critical points.
- Maximum value must be maximum on that interval.

SECOND DERIVATIVE TEST

- Set derivative equal to 0 and solve.
- Take the second derivative. If negative everywhere, then the critical point is the maximizer.

A Math Trick

- We are going to be taking the derivative of products a lot.
- The product rule is not fun. There has to be a better way… **Take the log!**
- $\ln(a \cdot b) = \ln(a) + \ln(b)$
- We do NOT need the product rule if our expression is a sum!

- Can we still take the max? $\ln()$ is an increasing function, so
  $\text{argmax}_\theta \ln(\mathcal{L}(E;\theta)) = \text{argmax}_\theta \mathcal{L}(E;\theta)$

Coin flips is easier

- $\mathcal{L}(\text{HTTTHHTHHH}; \theta) = \theta^6(1-\theta)^4$
- $\ln(\mathcal{L}(\text{HTTTHHTHHH}; \theta) = 6\ln(\theta) + 4\ln(1-\theta)$
- $\frac{d}{d\theta}\ln(\mathcal{L}(\cdot)) = \frac{6}{\theta} - \frac{4}{1-\theta}$
- Set to 0 and solve:
- $\frac{6}{\hat{\theta}} - \frac{4}{1-\hat{\theta}} = 0 \Rightarrow \frac{6}{\hat{\theta}} = \frac{4}{1-\hat{\theta}} \Rightarrow 6 - 6\hat{\theta} = 4\hat{\theta} \Rightarrow \hat{\theta} = \frac{3}{5}$
- $\frac{d^2}{d\theta^2} = \frac{-6}{\theta^2} - \frac{4}{(1-\theta)^2} < 0$ everywhere, so any critical point must be a maximum.

What about continuous random variables?

- Can NOT use probability, since the probability is going to be 0.

  -->Can use the density!

- It is supposed to show relative chances, that is all we are trying to find anyway.

$$\mathcal{L}(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

Continuous Example

- Suppose you get values $x_1, x_2, \ldots x_n$ from independent draws of a normal random variable $\mathcal{N}(\mu, 1)$ (for μ unkown)
- We will also call these "realizations" of the random variable.

- $\mathcal{L}(x_i; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right)$
- $\ln(\mathcal{L}(x_i; \mu)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2$

Finding $\hat{\mu}$

- $\ln(\mathcal{L}) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2$
- $\frac{d}{d\mu}\ln(\mathcal{L}) = \sum_{i=1}^{n} x_i - \mu$
- Setting $\mu = 0$ and solving:

$$\sum_{i=1}^{n} x_i - \hat{\mu} = 0 \Rightarrow \sum_{i=1}^{n} x_i = \hat{\mu} \cdot n \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Check using the second derivative test:

$$\frac{d^2}{d\mu^2}\ln(\mathcal{L}) = -n$$

Second derivative is negative everywhere, so log-likelihood is concave down and average of the x_i is a maximizer.

Summary

Given: an event E (usually n i.i.d. samples from a distribution with unknown parameter θ).

1. Find likelihood $\mathcal{L}(E;\theta)$

   Usually $\prod \mathbb{P}(x_i;\theta)$ for discrete and $\prod f(x_i;\theta)$ for continuous

2. Maximize the likelihood. Usually:
   A. Take the log (if it will make the math easier)
   B. Take the derivative
   C. Set the derivative to 0 and solve
3. Use the second derivative test to confirm you have a maximizer

Note: i.i.d means independent, identically distributed

Two Parameter Estimation

Two Parameter Estimation Setup

- We just saw that to estimate μ for $\mathcal{N}(\mu, 1)$ we get:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Now what happens if we know our data is $\mathcal{N}()$ but nothing else. Both the mean and the variance are unknown.

Log-likelihood

- Let θ_μ and θ_{σ^2} be the unknown mean and standard deviation of a normal distribution. Suppose we get independent draws $x_1, x_2, \ldots, x_n$.

- $\mathcal{L}\left(x_1, \ldots, x_n; \theta_\mu, \theta_{\sigma^2}\right) =$

  $$\prod_{i=1}^{n} \frac{1}{\sqrt{\theta_{\sigma^2} 2\pi}} \exp\left(-\frac{1}{2} \cdot \frac{\left(x_i - \theta_\mu\right)^2}{\theta_{\sigma^2}}\right)$$

- $$\ln\left(\mathcal{L}\left(x_i; \theta_\mu, \theta_{\sigma^2}\right)\right) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{\theta_{\sigma^2} 2\pi}}\right) - \frac{1}{2} \cdot \frac{\left(x_i - \theta_\mu\right)^2}{\theta_{\sigma^2}}$$

Expectation

Arithmetic is nearly identical to known variance case.

- $\ln\left(\mathcal{L}\left(x_i;\theta_\mu,\theta_{\sigma^2}\right)\right)=\sum_{i=1}^{n}\ln\left(\frac{1}{\sqrt{\theta_{\sigma^2}2\pi}}\right)-\frac{1}{2}\cdot\frac{\left(x_i-\theta_\mu\right)^2}{\theta_{\sigma^2}}$
- $\frac{\partial}{\partial\theta_\mu}\ln(\mathcal{L})=\sum_{i=1}^{n}\frac{(x_i-\theta_\mu)}{\theta_{\sigma^2}}$
- Setting equal to 0 and solving
- $\sum_{i=1}^{n}\frac{(x_i-\widehat{\theta_\mu})}{\theta_{\sigma^2}}=0\Rightarrow$

  $\sum_{i=1}^{n}\left(x_i-\widehat{\theta_\mu}\right)=0\Rightarrow\sum_{i=1}^{n}x_i=n\cdot\widehat{\theta_\mu}\Rightarrow\widehat{\theta_\mu}=\frac{\sum_{i=1}^{n}x_i}{n}$
- $\frac{\partial^2}{\partial\theta_\mu^2}=-\frac{n}{\theta_{\sigma^2}}$
- θ_{σ^2} is an estimate of a variance. It will never be negative (and as long as the draws aren't identical it won't be 0). So, the second derivative is negative, and we really have a maximizer.

Variance

- $\ln\left(\mathcal{L}\left(x_i;\theta_\mu,\theta_{\sigma^2}\right)\right)=\sum_{i=1}^{n}\ln\left(\frac{1}{\sqrt{\theta_{\sigma^2}2\pi}}\right)-\frac{1}{2}\cdot\frac{\left(x_i-\theta_\mu\right)^2}{\theta_{\sigma^2}}$

$$=\sum_{i=1}^{n}-\frac{1}{2}\ln(\theta_{\sigma^2})-\frac{1}{2}\ln(2\pi)-\frac{1}{2}\cdot\frac{\left(x_i-\theta_\mu\right)^2}{\theta_{\sigma^2}}$$

$$=-\frac{n}{2}\ln(\theta_{\sigma^2})-\frac{n\cdot\ln(2\pi)}{2}-\frac{1}{2\theta_{\sigma^2}}\sum_{i=1}^{n}\left(x_i-\theta_\mu\right)^2$$

- $\frac{\partial}{\partial\theta_{\sigma^2}}\ln(\mathcal{L})=-\frac{n}{2\theta_{\sigma^2}}+\frac{1}{2\left(\theta_{\sigma^2}\right)^2}\sum_{i=1}^{n}\left(x_i-\theta_\mu\right)^2$

Variance

- $\frac{\partial}{\partial \theta_{\sigma^2}} \ln(\mathcal{L}) =$

$$-\frac{n}{2\theta_{\sigma^2}} + \frac{1}{2(\theta_{\sigma^2})^2} \sum_{i=1}^{n} (x_i - \theta_\mu)^2$$

$$-\frac{n}{2\widehat{\theta_{\sigma^2}}} + \frac{1}{2(\widehat{\theta_{\sigma^2}})^2} \sum_{i=1}^{n} (x_i - \theta_\mu)^2 = 0$$

- $\Rightarrow -\frac{n}{2}\widehat{\theta_{\sigma^2}} + \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_\mu)^2 = 0$ (multiply by $(\widehat{\theta_{\sigma^2}})^2$)
- $\Rightarrow \widehat{\theta_{\sigma^2}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta_\mu)^2$

To get the overall max We'll plug in $\widehat{\theta_\mu}$

Summary

- If you get independent samples $x_1, x_2, \ldots, x_n$ from a $\mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are unknown, the maximum likelihood estimates of the normal is:

$$\widehat{\theta_\mu} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ and } \widehat{\theta_{\sigma^2}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \widehat{\theta_\mu}\right)^2$$

- The maximum likelihood estimator of the mean is the **sample mean** that is the estimate of μ is the average value of all the data points.
- The MLE for the variance is: the variance of the experiment "choose one of the x_i at random"

See file T1_MLEstimation.pdf

Slide 10 to 24

Two parameters 25-28

Special cases and some definitions and properties: 30-41

OTHER METHODS

Other methods for obtaining point estimators

- Method of moments (MM)
- Bayesian methods

Method of moments (MM)

Moments

- The k^{th} moment of a r.v. X is given by

$$\mu_k = E\left(X^k\right)$$

- The k^{th} central moment of a r.v. is given by

$$\mu_k = E\left(X - \mu_1\right)^k$$

- E.g. the variance of X is the second central moment

$$V(X) = E\left(X - E(X)\right)^2$$

Method of moments (MM)

Method of moments

Sample	Population
$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\mu_1 = E(X)$
$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$	$\mu_2 = E(X^2)$
$m_3 = \frac{1}{n}\sum_{i=1}^{n} X_i^3$	$\mu_3 = E(X^3)$
$\vdots$	$\vdots$
$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$	$\mu_k = E(X^k)$

- Equate sample moments to population moments.
- Use as many moments as the number of parameters you need to estimate.
- Write the parameters as a function of the sample moments.

Method of moments (MM)

Method of moments

Let $X \sim U(0, \theta)$. We take a simple random sample of size n.

$$m_1 = \overline{X} \qquad E(X) = \frac{\theta}{2}$$

$$\frac{\hat{\theta}_{MM}}{2} = \overline{X} \rightarrow \hat{\theta}_{MM} = 2\overline{X}$$

Exercise:

Let $X \sim Exp(\lambda)$. We take a simple random sample of size n.

$$f(x, \lambda) = \lambda e^{-\lambda x}$$

Find an estimator $\hat{\lambda}_{MM}$ for λ by using the method of moments.

We know how to estimate parameters such as μ, σ^2, or p. Now we need *general* methods to estimate other parameters

METHOD OF MOMENTS :

From a sample $X_1,\ldots,X_n$, suppose we wish to estimate some unknown parameter θ

If $E(X_i)$ is some function of θ, then we can estimate that function $\mu(\theta)$ using only the statistic $\overline{X}$

Now try to solve for θ as a function of $\bar{x}$, and call the solution $\hat{\theta}$

Example: $X_1,\ldots,X_n$ ~Exponential(λ)

$E(X)=1/\lambda$ (so $\mu(\lambda)=1/\lambda$)

Method of Moments : Solve for λ:

$$\overline{X} = \mu(\hat{\lambda}) \Rightarrow \hat{\lambda} = \frac{1}{\overline{X}}$$

Two Parameter Estimation

Suppose there are two parameters to estimate (θ, λ) and that

$E(X) = g(\theta, \lambda)$ $\quad$ $Var(X) = h(\theta, \lambda)$

$$\text{MME's} \Rightarrow \text{find } \hat{\theta}, \hat{\lambda} \text{ so that}$$

$$\overline{X} = g(\hat{\theta}, \hat{\lambda})$$

$$S^2 = h(\hat{\theta}, \hat{\lambda})$$

Key: solve 2 eqns for 2 unknowns

So $(\hat{\theta}, \hat{\lambda})$ are functions of $(\overline{X}, S^2)$

Method of moments for the Normal distribution

Let $X_1, X_2, \ldots X_n$ be random sample of size n from a $N(\mu, \sigma^2)$ distribution.

$$\mu_1 = E(X) = \mu \qquad \mu_2 = E(X^2) = V(X) + E(X)^2 = \sigma^2 + \mu^2$$

$$m_1 = \overline{X} \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\hat{\mu}_{MM} = \overline{X}$$

$$\hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

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

Generalized Method of Moments

- Method of moments is generally consistent.
- Given *p*(**x**;theta) if we know that the *kth* moment of *x[n]* μ_k is a function of θ as given by Eq.11.

$$\mu_k = E\left(x[n]^k\right) = f(\theta) \qquad (11)$$

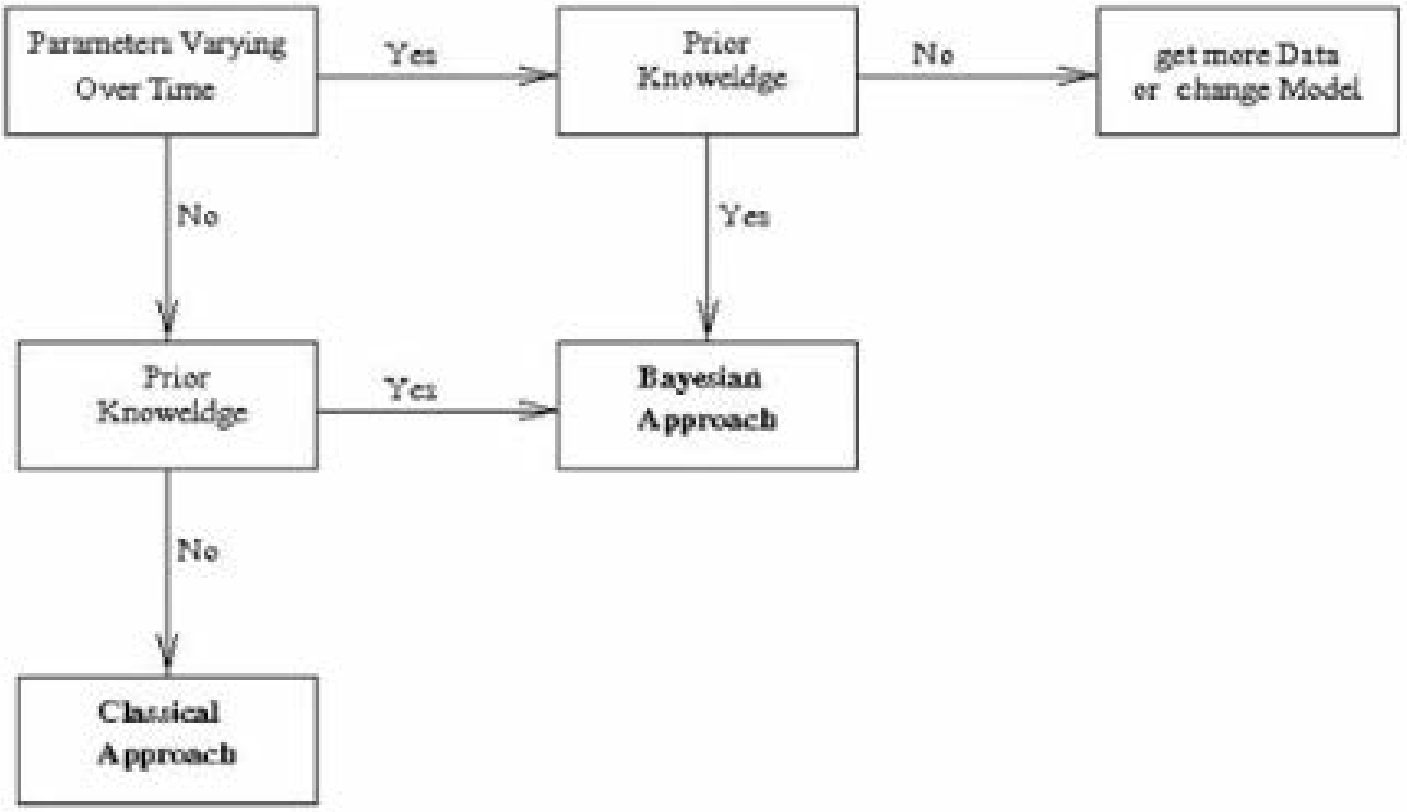
$$\theta = f^{-1}(\mu_k) \qquad (12)$$

$$\hat{\mu}_k = \frac{1}{N}\sum_{n=0}^{N-1} x^k[n] \qquad (13)$$

$$\hat{\theta} = f^{-1}\left(\frac{1}{N}\sum_{n=0}^{N-1} x^k[n]\right) \qquad (14)$$

- We approximate the *kth* moment of data **x**, $\hat{\mu}_k$ by taking average of $\mathbf{x}^{(k)}$ as given by Eq. 13.
- If *f* is an invertible function as given by Eq.12 then substitution of $\hat{\mu}_k$ into Eq.12 results in the estimator $\hat{\theta}$ as given by Eq.14.

Estimation approaches



- Source: https://slideplayer.com/slide/6240445/

Bayesian Methods

- In classical, frequentist statistics, θ is assumed to be an unknown, fixed quantity.
- In the Bayesian approach, θ is a random variable, and its variation is described by a distribution, the prior distribution, $\pi(\theta)$.
- The prior distribution, $\pi(\theta)$, is subjective, and chosen by the investigator.
- A sample $X_1, X_2, \ldots X_n$ is observed, and in the light of this data the distribution of θ is updated.
- The newly obtained distribution is called the posterior distribution, $\pi(\theta|\mathbf{x})$.
- The posterior distribution is

$$\pi(\theta|\mathbf{x}) = \frac{f(\theta, \mathbf{x})}{m(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}$$

with

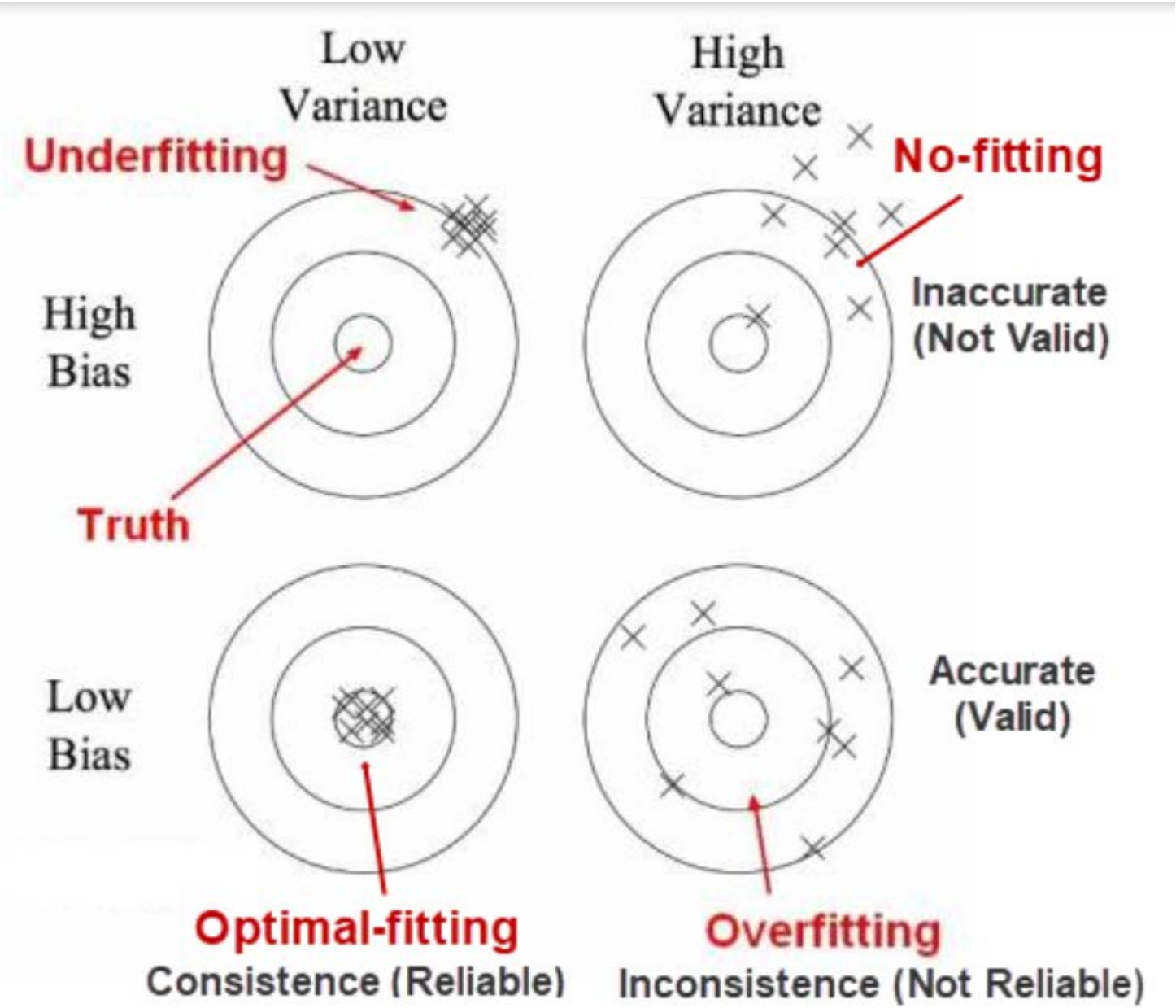
$$m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta$$

- A point estimate for θ is obtained by calculating the expectation (or the median) of the posterior distribution.
- The posterior distribution is proportional to the likelihood function and the prior

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

COMPARING ESTIMATORS

- Bias
- Variance or Precision
- Mean Squared Error [Root mean squared error, Mean Absolute Error (MAE), …]


Low
Variance
High
Variance
Underfitting
No-fitting
High
Bias
Inaccurate
(Not Valid)
Truth
Low
Bias
Accurate
(Valid)
Optimal-fitting
Consistence (Reliable)
Overfitting
Inconsistence (Not Reliable)

Bias $= E\left(\hat{\theta}\right) - \theta$

Variance $V\left(\hat{\theta}\right)$ (or Precision $= \frac{1}{V(\hat{\theta})}$)

Mean squared error $MSE(\hat{\theta}) \equiv E\left((\hat{\theta} - \theta)^2\right) = V\left(\hat{\theta}\right) + (Bias(\hat{\theta}))^2$

Bias

Is the difference between the average value of the estimator and the true value.

$$\text{Bias} = E\left(\hat{\theta}\right) - \theta$$

If Bias ($\hat{\theta}$)=0, then the estimator is said to be unbiased.

- Example1

Determine the Bias of the estimator $\hat{\mu} = \bar{X}$ of the parameter μ from a *poi(μ)* distribution.

Solution :

$Bais(\hat{\mu}) = E(\hat{\mu}) - \mu$

$= E(\bar{X}) - \mu$ -------- *Since* $\hat{\mu} = \bar{X}$ *is given*

$= E(\frac{1}{n}\sum_{i=1}^{n} x_i) - \mu$

$= \frac{1}{n}\sum_{i=1}^{n} E(x_i) - \mu$

$= \frac{1}{n}\sum_{i=1}^{n} \mu - \mu$ -------- *Since as* $x_i \sim Poi(\mu) \Rightarrow E(x_i) = \mu$

$= \frac{1}{n} n\mu - \mu$ --------- *As* μ *is constant over summation*

$= \mu - \mu$

$= 0$

Which means the estimator $\bar{X}$ *is unbiased . On average the value of the sample mean will give the true mean of the distribution.*

- Example2

Determine the Bias of the estimator $\hat{p} = \frac{X}{n}$ of the parameter p from a *Bin(n,p)* distribution.

Solution :

$Bais(\hat{p}) = E(\hat{p}) - p$

$= E\left(\frac{X}{n}\right) - p$ -------- *Since* $\hat{p} = \frac{X}{n}$ *is given*

$= \frac{1}{n} E(X) - p$

$= \frac{1}{n} np - p$ -------- *Since as* x_i~Bin(n,p) $\Rightarrow E(x_i) = np$

$= p - p$

$= 0$

Which means the estimator $\frac{X}{n}$ *is unbiased for P . On an average will give the true value for p.*

Variance or precision

Variance $V\left(\hat{\theta}\right)$ (or Precision $= \frac{1}{V(\hat{\theta})}$)

Mean Squared Error (MSE)

- Mean square error is The average of the error squared.
- Where error means the difference between the true value of the parameter (θ) and the estimated value ($\hat{\theta}$)of parameter.
- $MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$

  $= V(\hat{\theta}) + \left(Bias\ (\hat{\theta})\right)^2$

This formula gives us the measure of spread of the result around the true value of parameter.

- Thus, MSE has two components, one measures the variability of the estimator (precision) and the other measures its bias (accuracy).
- An estimator that has good MSE properties has small combined variance and bias.
- To find an estimator with good MSE properties, we need to find estimators that control both variance and bias.

- For an unbiased estimator $\hat{\theta}$, we have

$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = V(\hat{\theta})$ and so, if an estimator is unbiased, its MSE is equal to its variance.

- An estimator with smaller MSE is said to be more **efficient.** As if it has a smaller spread around the true value.

- If the MSE of the estimator tends to zero as $n \to \infty$ then it is said to be **consistent.**

- Example1

Determine the MSE of the estimator $\hat{\mu} = \bar{X}$ of the parameter μ from a *poi(μ)* distribution.

Solution :

$MSE(\hat{\mu}) = V(\hat{\mu}) + (Bias\ (\hat{\mu}))^2$

$= V(\bar{X}) + 0^2$ -------- *From example1 of Bias the* $(Bias\ (\hat{\mu}))=0$

-------- *Since* $\hat{\mu} = \bar{X}$ *is given*

$= V\ (\frac{1}{n}\sum_{i=1}^{n} x_i)$

$= \frac{1}{n^2}\sum_{i=1}^{n} V(x_i)$

$= \frac{1}{n^2}\sum_{i=1}^{n} \mu$ -------- *Since as* $x_i \sim Poi(\mu) \Rightarrow V(x_i) = \mu$

$= \frac{1}{n^2} n\mu$ -------- *As* μ *is constant over summation*

$= \frac{\mu}{n}$

For larger values of n our MSE will be more efficient, as you will have smaller MSE and we can see as n gets larger and larger the MSE tends to zero and so this is also a consistent estimator. It becomes more accurate as sample size increases.

- Example2

Determine the MSE of the estimator $\hat{p} = \frac{X}{n}$ of the parameter p from a *Bin(n,p)* distribution.

Solution :

$MSE(\hat{p}) = V(\hat{p}) + (Bias(\hat{p}))^2$

$= V\left(\frac{X}{n}\right) + 0^2$ -------- *From example2 of Bias the* $(Bias(\hat{p}))=0$

-------- *Since* $\hat{p} = \frac{X}{n}$ *is given*

$= \frac{1}{n^2} V(X)$

$= \frac{1}{n^2} npq$ -------- *Since as* $x_i \sim Bin(n,p) \Rightarrow V(x_i) = npq$

$= \frac{1}{n} p\, q$

For larger values of n our MSE will be more efficient, as you will have smaller MSE and we can see as n gets larger and larger the MSE tends to zero and so this is a consistent estimator.

What is better than mean squared error?

MSE values differ based on whether the values of the response variable are scaled or not. A better measure instead of MSE is the **root mean squared error** (RMSE) which takes care of the fact related to whether the values of the response variable are scaled or not.

- The RMSD of an estimator $\hat{\theta}$ with respect to an estimated parameter θ is defined as the square root of the mean squared error:

$$\mathrm{RMSD}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})} = \sqrt{\mathrm{E}((\hat{\theta} - \theta)^2)}$$

Question

Is the mean squared error always equal to the variance for an unbiased estimator?

- To find an estimator with good MSE properties, we need to find estimators that control both variance and bias. For an unbiased estimator $\hat{\theta}$, we have MSE $\hat{\theta}$ = E($\hat{\theta}-\theta)^2$ = Var($\hat{\theta}$) and so, **if an estimator is unbiased, its MSE is equal to its variance**.
- LIKEWISE: For an unbiased estimator, the RMSD is the square root of the variance, known as the standard deviation.

Reading

MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better (in linear regression)?

https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Mean-Squared Error (MSE) and **R-squared** when evaluating the linear regression models:

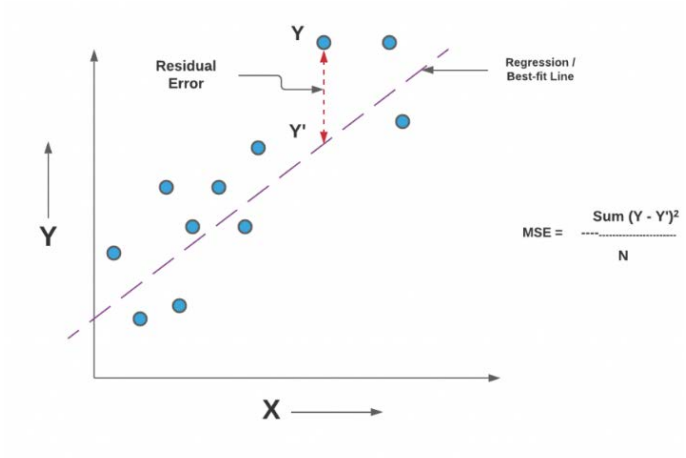


Fig 2. Mean Squared Error Representation

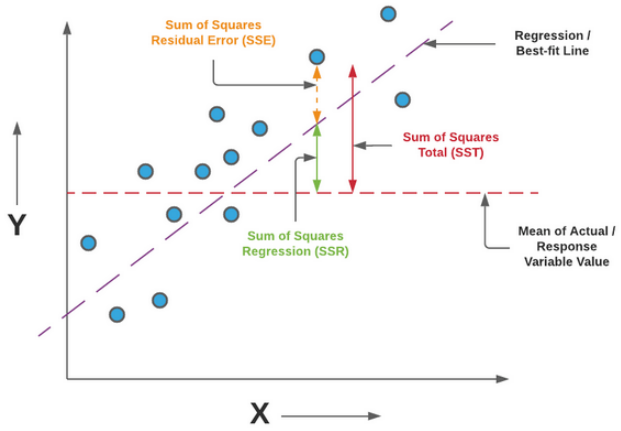


Fig 4. Diagrammatic representation for understanding R-Squared

- MSE represents the residual error which is nothing but sum of squared difference between actual values and the predicted / estimated values divided by total number of records.
- R-Squared represents the fraction of variance captured by the regression model.
- The disadvantage of using MSE is that the value of MSE varies based on whether the values of response variable is scaled or not. If scaled, MSE will be lower than the unscaled values.

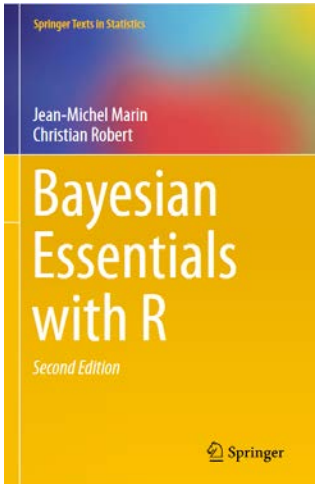
References

- ML Estimation

- Millar, R. B.Maximum Likelihood Estimation and Inference:With Examples in R, SAS and ADMB (Statistics in Practice Book 112) (English Edition) 1o Edición
- Casella, R. & Berger, R. L. (2002) Statistical Inference. Duxbury, Pacific Grove, CA, USA. Second edition, Chapter 7.
- DeGroot, M. H. & Schervish, M.J. (2002) Probability and Statistics. Addison-Wesley. Third edition. Chapter 6.
- Ewens, W. J. & Grant, G. R. (2005) Statistical Methods in Bioinformatics. An Introduction. Springer. Second Edition.

- Bayesian approach

https://michaela-kratofil.com/files/2014_Book_BayesianEssentialsWithR.pdf



Research (One example of)

Maximum Likelihood Estimation of Biological Relatedness from Low Coverage Sequencing Data

https://www.biorxiv.org/content/10.1101/023374v1

EXAMPLES AND EXERCISES

Problem: white and black balls

- A box contains white and black balls mixed up in some unknown proportion. A ball is drawn from the box. Its colour is noted and then it is put back into the box. The experiment is repeated ten times.
- The results are as shown below:

- Which of the following populations do you think the balls were drawn from?
- A: 10% White; 90% Black
- B: 30% White; 70% Black
- C: 50% White; 50% Black
- D: 70% White; 30% Black
- E: 90% White; 10% Black

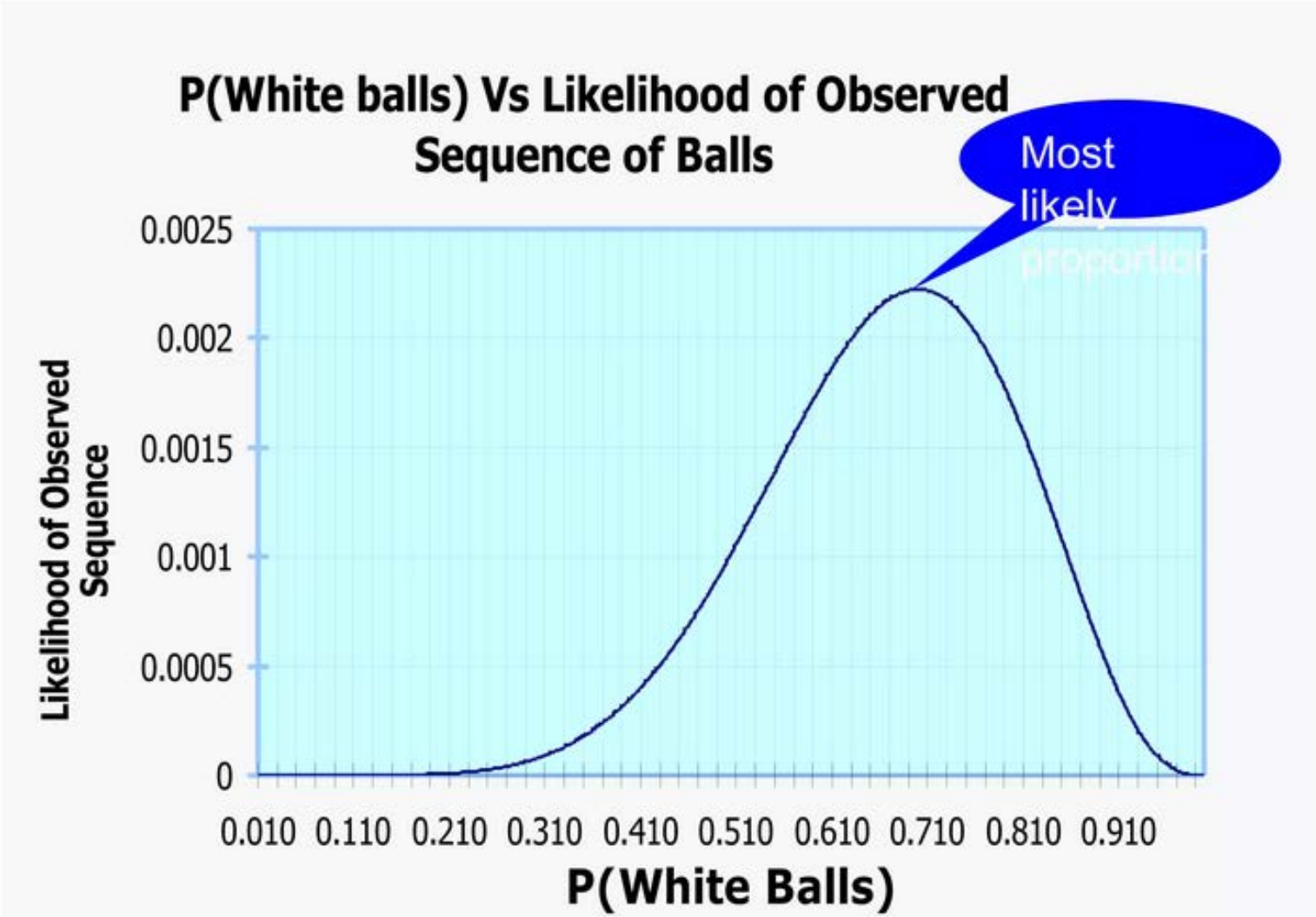
Maximum Likelihood Method

- In terms of colour of balls, the result we have observed is WWWWBWWBBW
- Suppose the proportion of white balls from the box is some unknown quantity, say π, so that the proportion of black is 1- π, then what is the probability of drawing this sequence of balls from the box?
- The balls are selected independently from the box, therefore we can obtain the probability as

  $p = \pi*\pi*\pi*\pi*(1-\pi)*\pi*\pi*(1-\pi)*(1-\pi)*\pi = \pi^7(1-\pi)^3$
- Now let's consider the probabilities we would find for the different values of π in the question above

What is p, for different values of π?

- A: $\pi = 0.1$; $p = 0.0000$
- B: $\pi = 0.3$; $p = 0.0001$
- C: $\pi = 0.5$; $p = 0.0010$
- D: $\pi = 0.7$; $p = 0.0022$
- E: $\pi = 0.9$; $p = 0.0005$
- Therefore, it is most likely that a sequence of balls with colours WWWWBWWBBW will be drawn from a box where the proportion of black balls is 0.7.
- Probability (p) is about FUTURE events: When we looking back into the past, i.e. have already observed a result and we are trying to work out what could have caused it, we talk about likelihood (L).
- So our Maximum Likelihood Estimate of $\pi = 0.7$


P(White balls) Vs Likelihood of Observed Sequence of Balls
Most likely
Likelihood of Observed Sequence
0.0025
0.002
0.0015
0.001
0.0005
0
0.010 0.110 0.210 0.310 0.410 0.510 0.610 0.710 0.810 0.910
P(White Balls)

MLM: How it works

1. A result is observed, e.g. 7 white balls, 3 black
2. Several hypotheses are proposed for what could have caused the observed results, e.g. $\pi = 0.1$, $\pi = 0.3$, etc.
3. For each hypothesis, the likelihood that it could have caused the observed result is calculated.
4. The hypothesis that has the maximum likelihood, e.g. $\pi = 0.7$ is taken as the best estimate for the observed result.

MLM estimation problem: Defective Handbags

- TP Samuel produces leather handbags. Their production line can be set at two different speeds. At the lower speed, the proportion of defective handbags in each batch is 10%. At the higher speed the percentage of defective handbags is 30%.
- The quality sampling plan is to take an unbiased random sample of 7 handbags from the production line and inspect each in turn.
- In a quality check, a sample of 7 handbags is extracted. It is found that two handbags are defective. Which of the two production speed is most likely to have been used in the process?
- Suggested Steps:
- 1. Work out the most appropriate probability distribution for the sampling process
- 2. Consider two hypothesis: (i) The proportion of defectives in the process is 10% or (ii) The proportion is 30%
- 3. Consider the likelihood of each hypothesis is the light of the observed sample results, i.e. for each hypothesis calculate the likelihood that 2 out of 7 handbags are defective, using the probability distribution that you think is most appropriate.

- Solution

...

Bernoulli distribution – R code (I)

```
# likelihood function of a Bernoulli random variable
m <- 100   # number of flips
p <- 0.50
n <- 1     # parameter of the binomial.

set.seed(123)

x <- rbinom(m,size=n,prob=p)
x

p_ml <- mean(x)
p_ml

likelihood <- function(p,x) {
  S <- sum(x)
  n <- length(x)
  L <- (p^S)*((1-p)^(n-S))
  return(L)
}
```

Bernoulli distribution – R code (II)

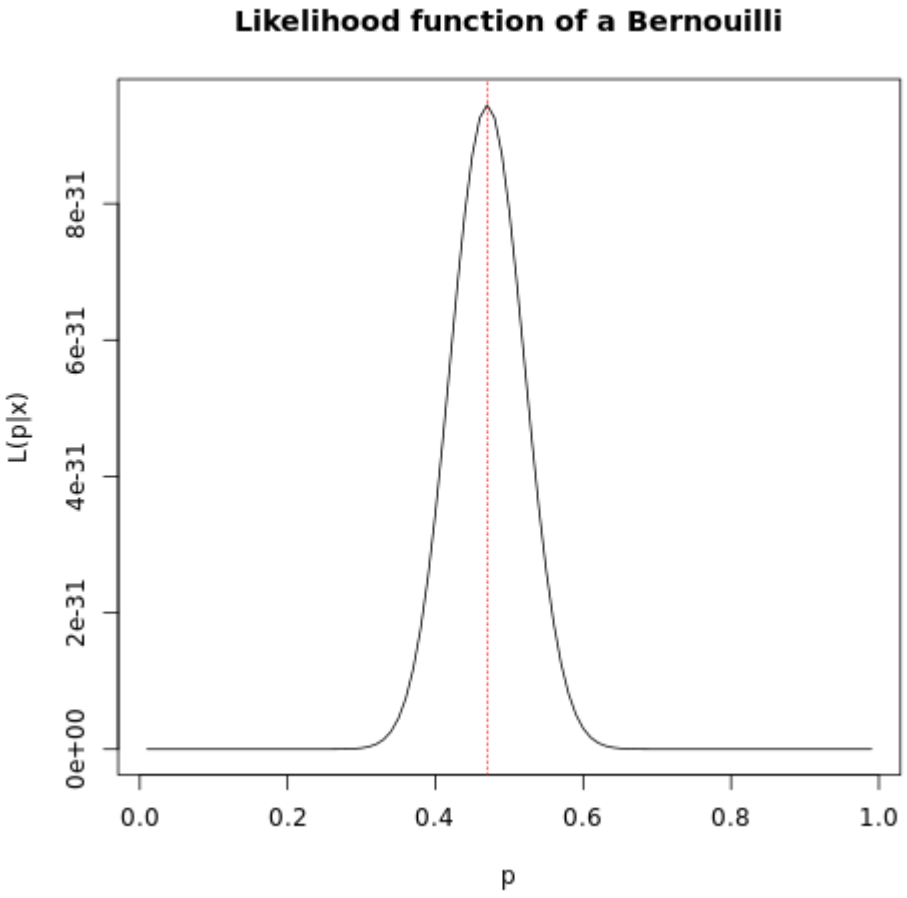
```
loglikelihood <- function(p,x) {
  S <- sum(x)
  n <- length(x)
  L <- (p^S)*((1-p)^(n-S))
  l <- log(L)
  return(l)
}

likelihood(p_ml,x)

vp <- seq(0.01,0.99,by=0.01)
vp

L <- likelihood(vp,x)
L

l <- loglikelihood(vp,x)
l
```

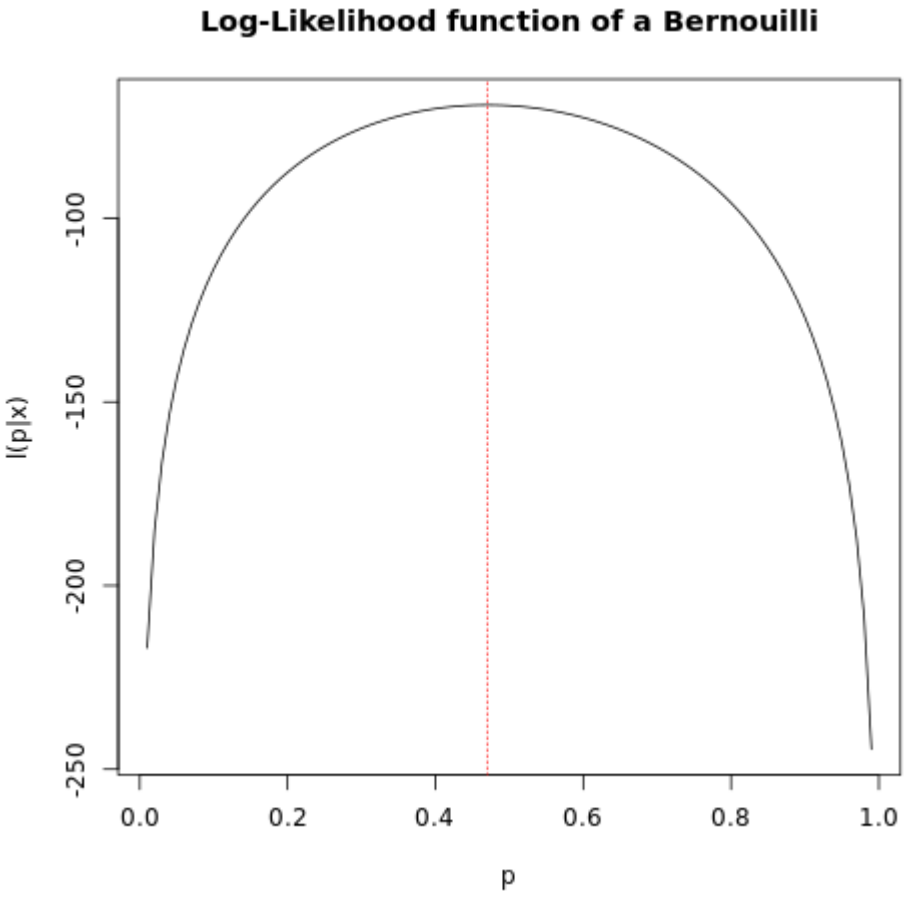
Bernoulli distribution – R code (III)

```
plot(vp,L,type="l",
    main="Likelihood function of a Bernouilli",
xlab="p",ylab="L(p|x)")

abline(v=p_ml,col="red",lty="dotted")


plot(vp,l,type="l",
    main="Log-Likelihood function of a Bernouilli",
xlab="p",ylab="l(p|x)")

abline(v=p_ml,col="red",lty="dotted")
```

Bernoulli distribution – R output

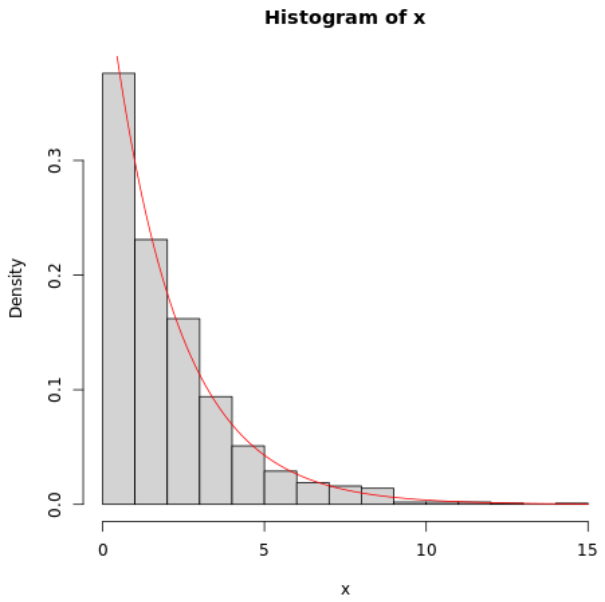
- https://www.mycompiler.io/es/new/r





Exponential distribution

```
# ML estimation of lambda of an exponential distribution
lambda <- 0.5
n <- 1000

set.seed(123)
x <- rexp(n=n,rate=lambda)

hist(x,freq=FALSE)

lambda_ml <- 1/mean(x)
lambda_ml

#Confidence interval calculation
ll <- lambda_ml - qnorm(0.975)*lambda_ml/sqrt(n)
ul <- lambda_ml + qnorm(0.975)*lambda_ml/sqrt(n)
ll
ul

sx <- seq(0,15,by=0.1)
?dexp

fx <- dexp(sx,rate=lambda_ml)
points(sx,fx,type="l",col="red")
```



ANNEX

Newton-Raphson Method
An overview (I)

We consider a single root, x_r, of the function $f(x)$. The Newton-Raphson method begins with an initial estimate of the root, denoted $x_0 \neq x_r$, and uses the tangent of $f(x)$ at x_0 to improve on the estimate of the root. In particular, the improvement, denoted x_1, is obtained from determining where the line tangent to $f(x)$ at x_0 crosses the x-axis. This represents a single iteration of the Newton-Raphson method and is illustrated in Figure below (a). The next iteration uses the line tangent to $f(x)$ at x_1 to generate x_2 in exactly the same way. This is shown in Figure (b). The direction of the tangent line is determined by the sign of the local gradient at each x_n and, although the illustration has a positive local gradient, the method works analogously for functions with a negative local gradient.

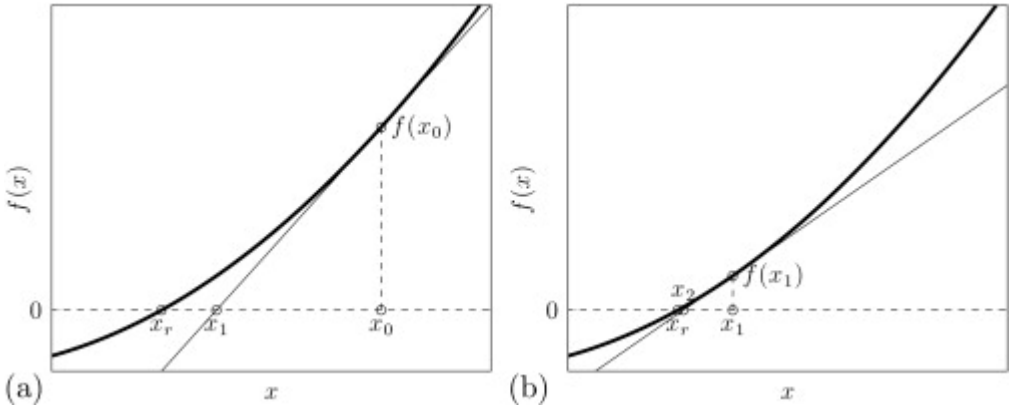


Figure. Illustration of the Newton-Raphson method. (a) First iteration leading to x_1 from the tangent line at $f(x_0)$. (b) Second iteration leading to x_2 from the tangent line at $f(x_1)$.

Newton-Raphson Method An overview (II)

The mathematical formulation of the first iteration of the process. Essentially, we require a expression that gives x_1 from x_0 and the properties of $f(x)$ at x_0. The gradient of the line tangent to $f(x)$ at x_0 is, by definition, $f'(x_0)$ which can be determined from the expression of $f(x)$. We can also form an approximation to this gradient from the fact that the tangent line crosses the points $(x_0,f(x_0))$ and $(x_1,f(x_1)) \approx (x_1,0)$; which of course arises from the definition of x_1. The gradient is then

$$\frac{f(x_0)-0}{x_0-x_1}$$

which is equated to $f'(x_0)$ and rearranged to give $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$

That is, we have an expression for estimate x_1 determined entirely from *known* properties of the function at x_0. Clearly the expression assumes that $f'(x_0) \neq 0$. If this was the case, the tangent line of the function at x_0 would be horizontal and not cross the x-axis; the Newton-Raphson method would then have failed to improve on x_0.

Equation for x_1 can be generalized to give an expression for the $(n + 1)$th estimate under the Newton-Raphson method from properties of the nth iteration,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

$$\text{for } n \in \mathbb{N} \text{ and } f'(x_n) \neq 0$$

The standard error estimate used in an implementation of the Newton-Raphson method is $\epsilon_n = |x_n - x_{n-1}|$. This means that an exit criteria is simply that $\epsilon_n < \epsilon$ for some predetermined tolerance, ϵ. That is, we terminate the iterative process when successive approximations become only marginally different.

Newton-Raphson Method
An overview (III)

Although the description of the Newton-Raphson method has been given for functions with a single root, the method can be applied perfectly well to functions with multiple roots. The root on which the method converges is of course determined by the starting value, x_0. As with the interval methods, it is sensible to have a rough idea of the number of roots that a function may possess and the approximation location of each. There are no definitive rules on how an initial estimate should be determined such that the Newton-Raphson method approaches the particular root of interest and common sense should be applied in practice.

Source:
https://www.sciencedirect.com/topics/mathematics/newton-raphson-method

Webliography

Some examples and exercicises taken from:
Max Chipulu
Senior Teaching Fellow (Management Science) at University of Southampton
Source: https://www.slideshare.net/maxchipulu/the-method-of-maximum-likelihood

_____.

Examples of MM and ML estimator calculation:
https://www.slideshare.net/ssuser23a30e/estimation-methodspptx

_____.

Further detail of (generalized) method of moments can be looked up at:
https://www.eco.uc3m.es/~ricmora/mei/materials/Session_12_GMM.pdf

_____.

Good explanation (with very few formula) on how the MLE and LRT work:
https://www.slideshare.net/meerooahlawy/maximum-likelihood-estimator-intro