Contents

Maximum Likelihood

Introduction

The notes from this lecture are derived from I.J. Myung's paper "Tutorial on maximum likelihood estimation". Journal of Mathematical Psychology 47 (2003) 90-100. The interested reader is directed to review this paper.

In this lecture we will understand an alternative least-squares estimation, Maximum Likelihood Estimation (MLE).

MLE is a preferred method of parameter estimation in statistics and is a general parameter estimation approach, in particular in non-linear modeling with non-normal data.

General Exprimental Approach

In scientific investigatio we seek to "uncover general laws and principles that govern the behavior under investigation".

Because these phenonmenon are not dricely observable, we formulate hyphotheses to test and these hypohtesis are stated in terms of probability using statistical models.

The goal of statistical modeling is to understand underlying processes by testing the viability (e.g. quality, robustness) of the model.

Once the model is specified, and data are collected, we can evaluate how well the model fits the data: 1. Determine parameter values (parameter estimation) 2. Evaluate goodness of fit

For example, think of the process we use in linear modeling

Two general approaches to parameter estimation:

1) LSE (Least-squares estimation)
2) MLE, MLE is a general approach for parameter estimation beyond those we have seen (e.g. Bayesian methods, using missing data, random effects, many other approaches.)

So, how do we do this?

The probability density function

The areas under a fixed distribution The goal of data analysis is to identify the population that is most likely to have generated the sample - i.e. we will estimate the parameters of the candidate model that .

The data vector $y = (y_1, y_2, ..., y_m)$ is a random sample from the unknown population.

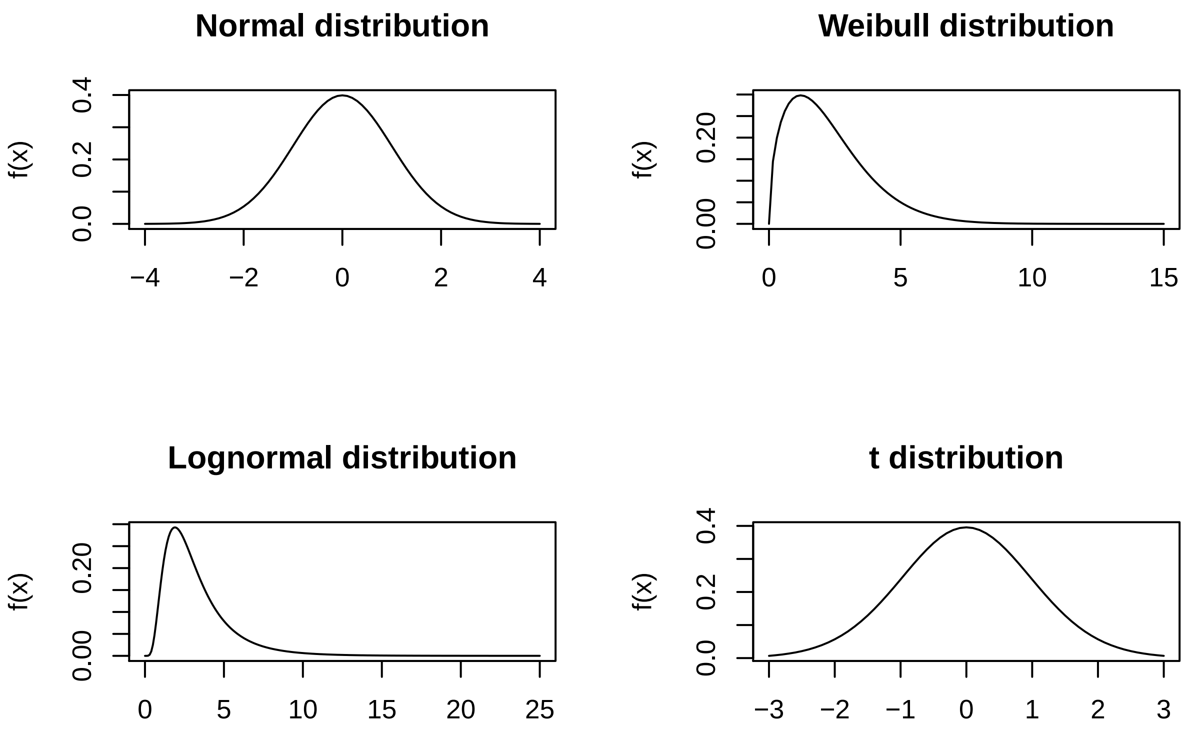
Populations are identified using a probablity distribution and unique values of the parameters - As the parameter changes in value, different probability distributions are generated.

Let $f(y|w)$ denote the *probability density function* (PDF) that specifies the probability of observing data vector y given the parameter w.

The parameter $w = (w_1, ..., w_k)$ is a vector defined on a multi-dimensional parameter space.

Remember, different distributions are defined using different parameters, so the length of w is distribution specific.

If we have specified a distribution that has a certain set of paramters, for example:



Normal distribution

Weibull distribution

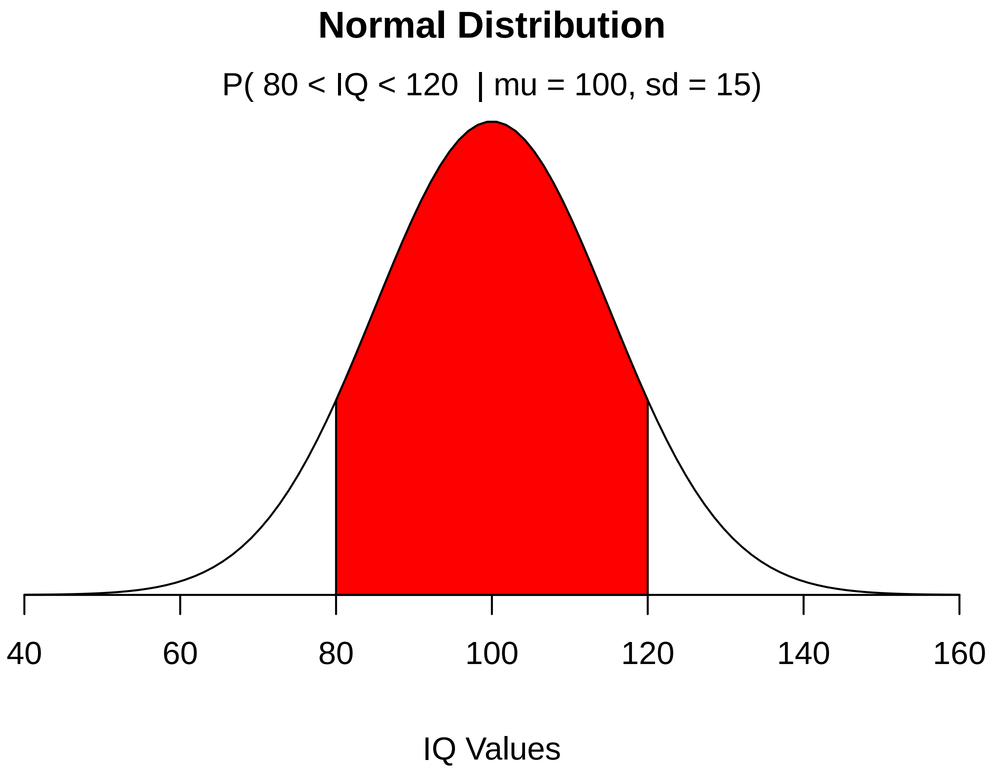
Lognormal distribution

t distribution

We can use a given probability distribution and parameter set to determine the probability of obtaining a value in a population.

For example:

Children's IQ scores are normally distributed with a mean of 100 and a standard deviation of 15. What is the probability of a randomly selected child having an IQ between 80 and 120?



In this case the area under the curve is 0.818 or 81.8% of the integral of the distribution from $-\infty$ to $+\infty$. So, if we take 100 random draws from the population of children's IQs, we will get values of IQ $>$ 80 and $<$ 120, 81 to 82 times…

Let's examine the statement $p(80 < IQ < 120 \mid \mu = 100, \sigma = 15) = 0.0818$.

We are stating that the *probability* of randomly selecting a student the variable characteristics *given* the parameter values *is* 0.818.

If we are interested in finding probabilities of students with different variable characteristics (different IQ values), then we will change the *left* side of the equation.

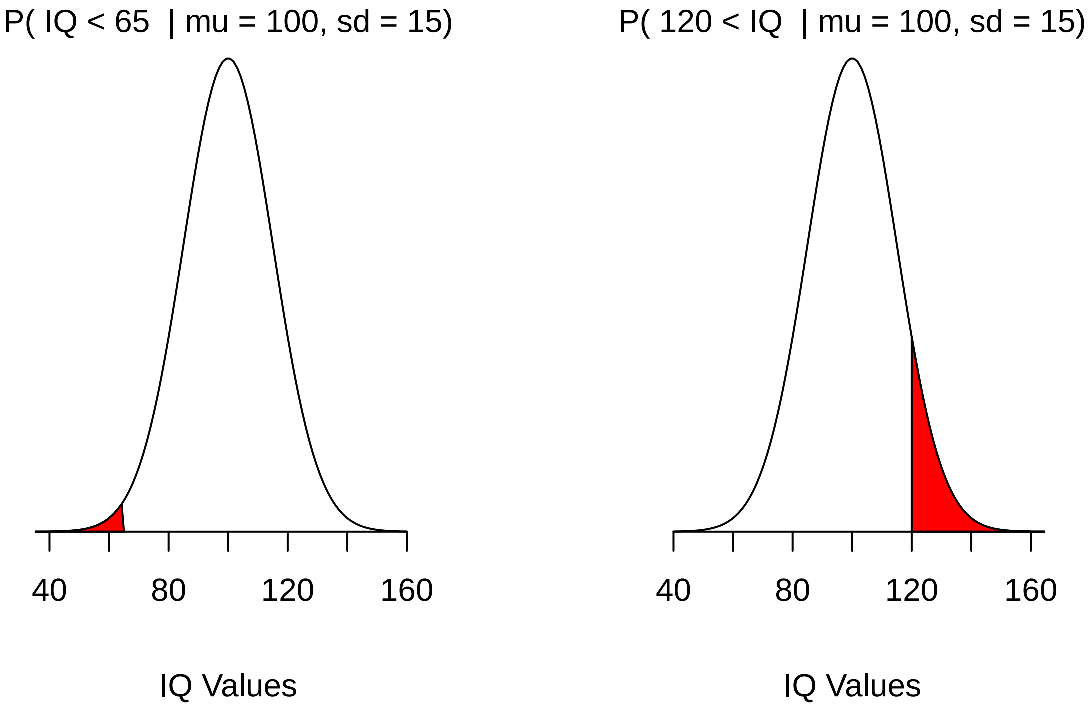
For example:

$p(IQ < 65 \mid \mu = 100, \sigma = 15)$

or

$p(IQ > 120 \mid \mu = 100, \sigma = 15)$

The right side of the equation stays the same - it describes the shape of the distribution. So when we are investigating probability of an event we are quantifying the integral of the curve for the given parameter set bounded by the *left* side of the equation. We change the left side to derive new probability values.



We can determine the probability of obtaining $p(y_1, y_2, ..., y_m) \mid w)$ if the observations are independent.

Think about taking multiple random draws from the population described by the parameter set w.

$p(y_1, y_2, ..., y_m) \mid w) = p(y_1 | \, w) \times p(y_2 | \, w) \times ...p(y_n | \, w)$

$p(y_1, y_2, ..., y_m) \mid w) = \prod_{i=1}^{n} p(y_i | w)$

The likelihood function

Data are fixed, distrbutions change - through alteration of parameter values Given a set of parameter values, w, the corresponding PDF will show that some data are more probable than other data.

In the previous example, the PDF with $90 <$ IQ < 100 is more likely to occur than IQ < 65.

In reality, however, we have already observed the data. Accordingly, we are faced with an inverse problem: Given the observed data and a model of interest, find the one PDF, among all the probability densities that the model prescribes, that is most likely to have produced the data.

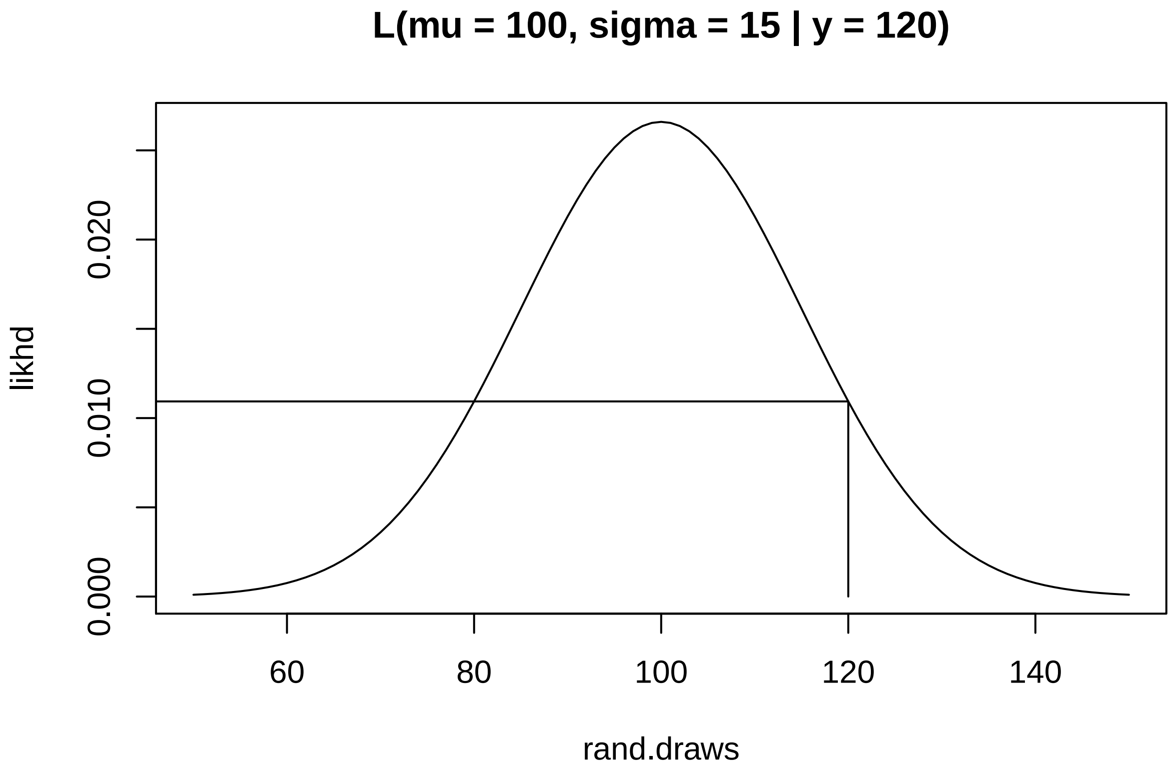
To solve this inverse problem, we define the likelihood function by reversing the roles of the data vector y and the parameter vector w.

So, we focuse on $L(w|y)$. This represents the likelihood of the parameter w given the observed data y; and as such is a function of w.

So, the data are fixed, and we modify the parameter vector w (in the case of IQ, μ and σ)

Assume we have some vector of observed data y, these are sampled from some population, for now, assume we take a single value.

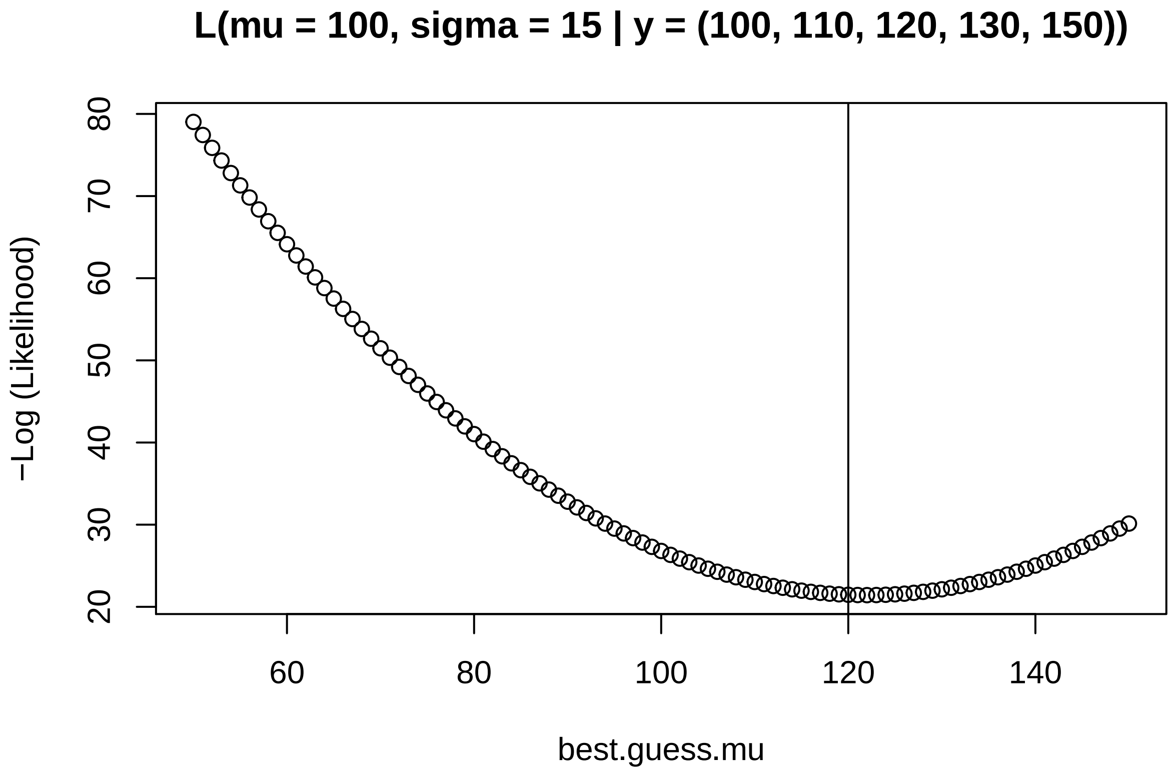
What is the likelihood that $\mu = 100$ and $\sigma = 15$ given our sampled IQ is $y = 120$?



Okay, now what if we have a vector of observations $n = 5$, with mean values centered on 120. So, we want to find the distribution parameters that maximize the likelihood.

We still think that the normal distribution is the most appropriate distribution as a candidate distribution.

$y = (100, 110, 120, 130, 150)$


L(mu = 100, sigma = 15 | y = (100, 110, 120, 130, 150))
−Log (Likelihood)
20
30
40
50
60
70
80
60
80
100
120
140
best.guess.mu