Introduction to Probability Distribution Fitting

Jerome Dumortier

22 September 2022

Lecture Overview

Distribution fitting

- Finding the best-fitting theoretical probability distribution for the observed data

Three approaches covered in this lecture:

- MASS: `fitdistr()`
- fitdistrplus: `fitdist()`
- gamlss: `fitDist()`

Notes:

- No need to specify distribution function for the last approach, i.e., `fitDist()`
- Introduction and overview to a very broad field of research

Introduction

Empirical work often requires understanding of the underlying distribution of data:

- Distribution of corn yields in a particular county based on observations to calculate the probability of getting a yield below a certain threshold, e.g., for crop insurance purposes
- Wind speed distribution at a particular location for construction of a wind farm: Electricity production is not possible below and above a certain wind speed

Estimation of one or more parameters characterizing a probability distribution function

Introductory Example

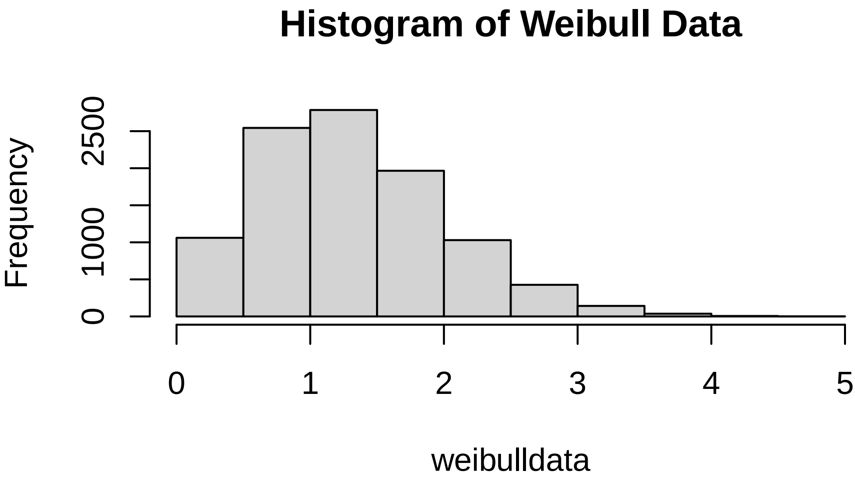
Weibull: Random Data Generation

Random generation of data (N=10000) following a Weibull distribution with two parameters:

- Shape: $k = 2$
- Scale: $\lambda = 1.5$

```
weibulldata = rweibull(10000,2,1.5)
```

Weibull: Histogram



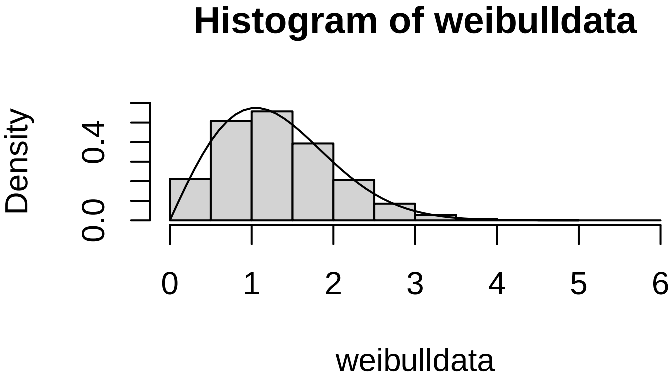
Weibull: Distribution Fitting with fitdistr

```
weibullpara     = fitdistr(weibulldata,densfun="weibull",
                           lower=c(0,0))
shape           = weibullpara$estimate[1]
scale           = weibullpara$estimate[2]
c(shape,scale)
```

```
##    shape    scale
## 1.994347 1.488994
```

Weibull: Observed Data and Estimated Distribution

```
hist(weibulldata,freq=FALSE,ylim=c(0,0.6),xlim=c(0,6))
range              = seq(0,6,0.1)
lines(range,dweibull(range,shape,scale))
```



Approach

Distribution Fitting Steps

General steps (see Fitting Distributions with R by Vito Ricci for more information)

1. General hypothesis about candidate distributions, e.g., discrete vs. continuous, entire real number line vs. positive numbers only
   - Histogram as a valuable first approach
2. Parameter estimation
   - Example: Calculating shape and scale parameters of the Weibull distribution or mean and variance for a Normal distribution
3. Goodness of fit

Starting point for an overview of various probability distributions: List of probability distributions

Candidate Distributions and Estimation

Meridian Hills: Possible Distributions

Meridian Hills home values:

- Source: https://jrfdumortier.github.io/dataanalysis/
- 101 home values in the Meridian Hills neighborhood in Indianapolis
- Scaling of data to measure home values in $1000

Candidate distributions:

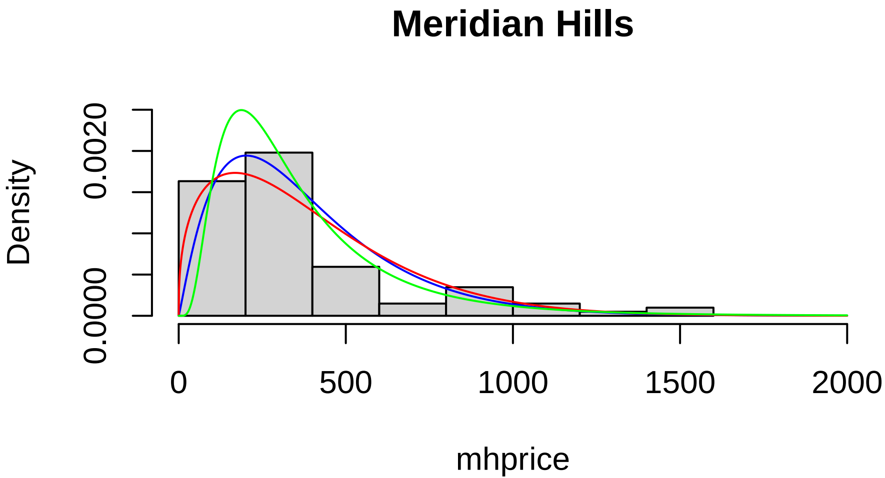
- Gamma distribution: Shape and scale parameter
- Weibull distribution: Shape and scale parameter
- Log-normal distribution, i.e, $Y = \ln(X)$ has a normal distribution: μ and σ

```
mhprice       = mh1$price/1000
mhgamma       = fitdistr(mhprice,"gamma")
mhweibull     = fitdistr(mhprice,"weibull",lower=c(0,0))
mhlognormal   = fitdistr(mhprice,"log-normal")
```

Meridian Hills: Histogram I

```
hist(mhprice,freq=FALSE,ylim=c(0,0.0025),
     xlim=c(0,2000),main="Meridian Hills")
range          = seq(0,2000,1)
lines(range,dgamma(range,mhgamma$estimate[1],
                   mhgamma$estimate[2]),col="blue")
lines(range,dweibull(range,mhweibull$estimate[1],
                     mhweibull$estimate[2]),col="red")
lines(range,dlnorm(range,mhlognormal$estimate[1],
                   mhlognormal$estimate[2]),col="green")
```

Meridian Hills: Histogram II



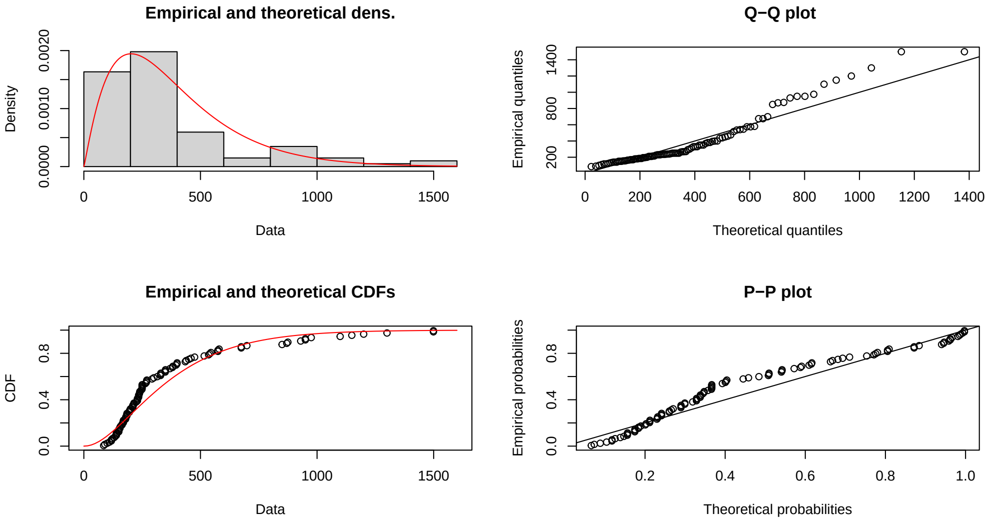
Goodness of Fit

Meridian Hills: Setup for `fitdist()`

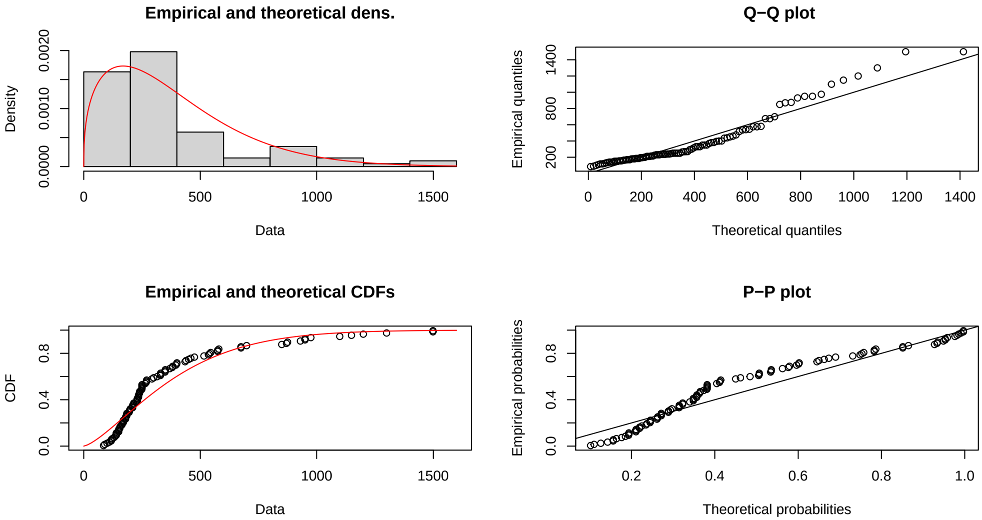
Use of the function `fitdist()` from the package fitdistrplus

```
mhprice       = mh1$price/1000
mhgamma       = fitdist(mhprice,"gamma",lower=c(0,0))
mhweibull     = fitdist(mhprice,"weibull",lower=c(0,0))
mhlognormal   = fitdist(mhprice,"lnorm",lower=c(0,0))
```

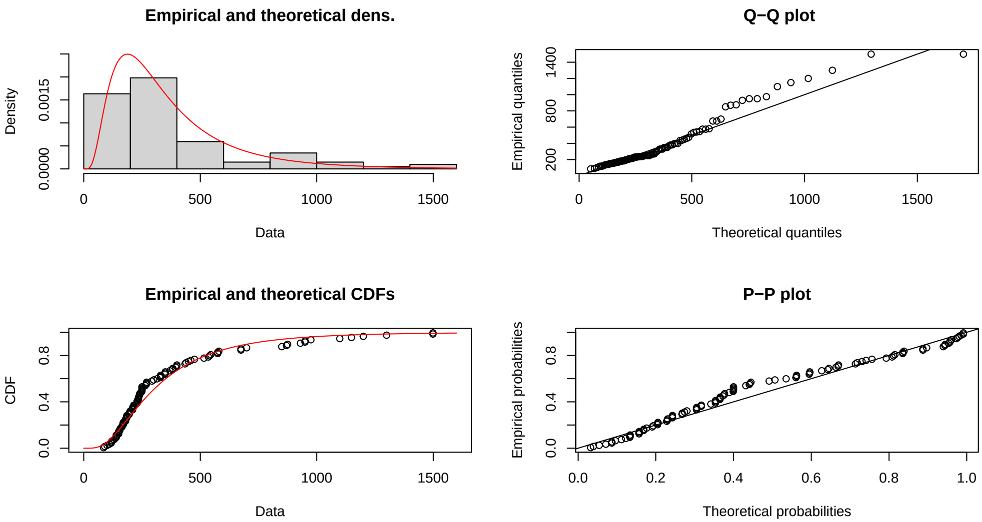
Meridian Hills: Gamma Distribution



Meridian Hills: Weibull Distribution



Meridian Hills: Log-Normal Distribution



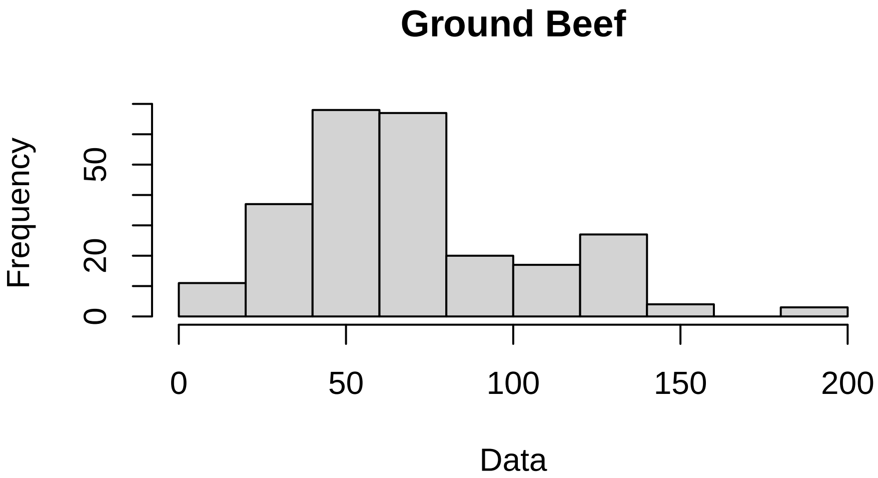
Ground Beef: Possible Distributions

Second example using the function `fitdist()` package:

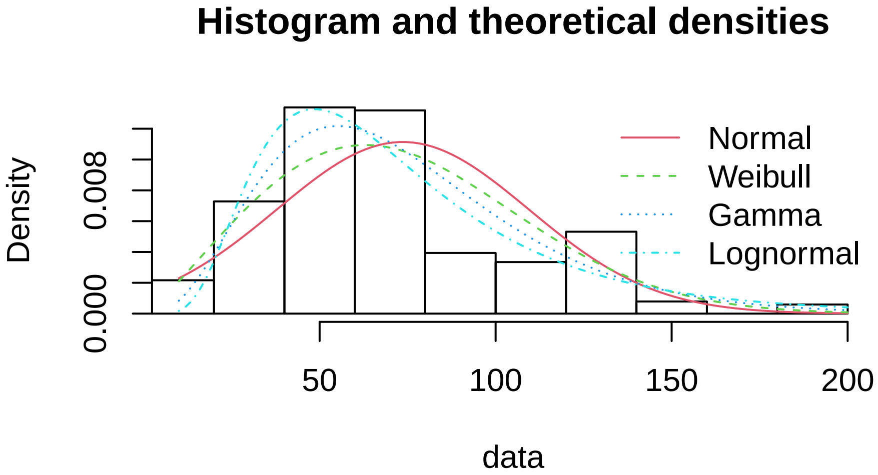
- Use of the data `groundbeef` associated with the package `fitdistrplus`: Serving sizes collected in a French survey, for ground beef patties consumed by children under 5 years old.

```
data("groundbeef")
gbnormal       = fitdist(groundbeef$serving,"norm")
gbweibull      = fitdist(groundbeef$serving,"weibull")
gbgamma        = fitdist(groundbeef$serving,"gamma")
gblognormal    = fitdist(groundbeef$serving,"lnorm")
fitteddist     = list(gbnormal,gbweibull,gbgamma,gblognormal)
plotlegend     = c("Normal","Weibull","Gamma","Lognormal")
```

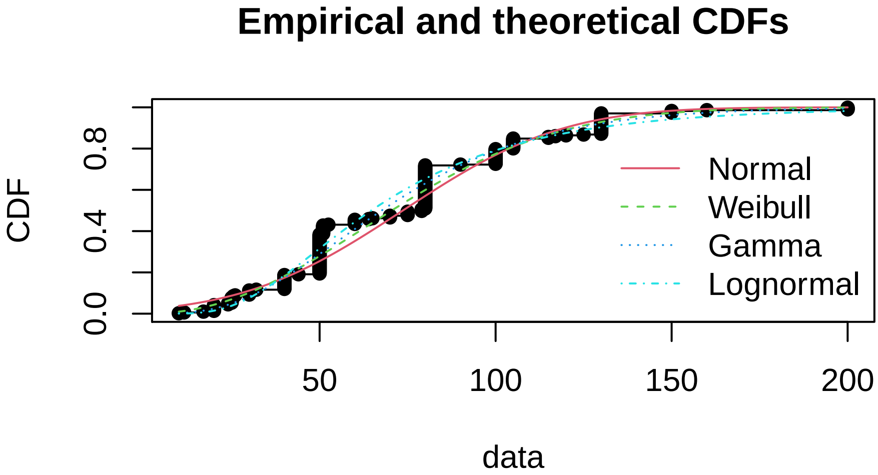
Ground Beef: Histogram



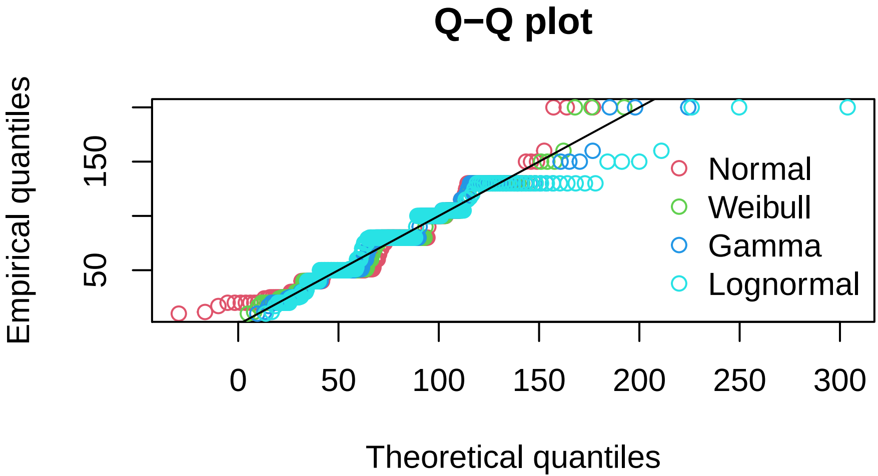
Ground Beef: Results I



Ground Beef: Results II



Results: Q-Q Plot



Unspecified Distribution: `fitDist()`

Use of the function `fitDist()` from package gamlss

```
output = fitDist(mhprice,type="realplus")
```

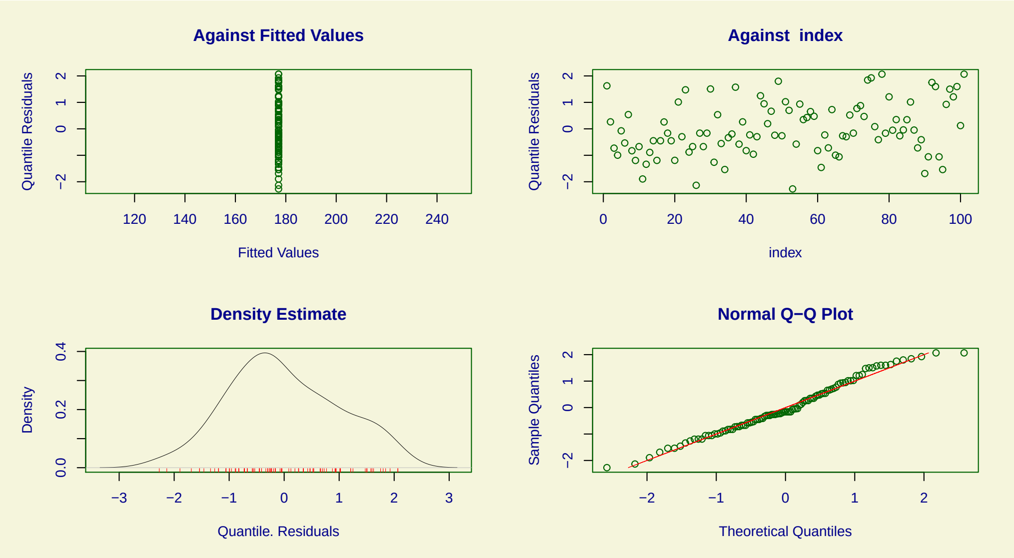
```
output$family
```

```
## [1] "IGAMMA"        "Inverse Gamma"
```

```
output$Allpar
```

```
##     eta.mu  eta.sigma
##  5.1768720 -0.4921408
```

Goodness of Fit with Inverse Gamma

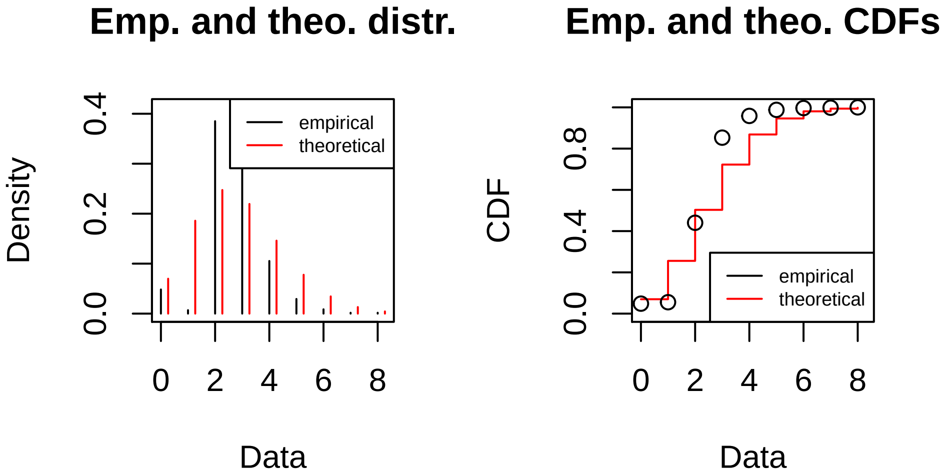


Discrete Data Distribution Fitting

EV Data

```
evpoisson = fitdist(evdata$numcars,discrete=TRUE,distr="pois")
evnbinom  = fitdist(evdata$numcars,discrete=TRUE,distr="nbinom")
```

EV Data: Results Poisson



EV Data: Results Negative Binomial