M. Thrun, Prof. Dr. Ultsch

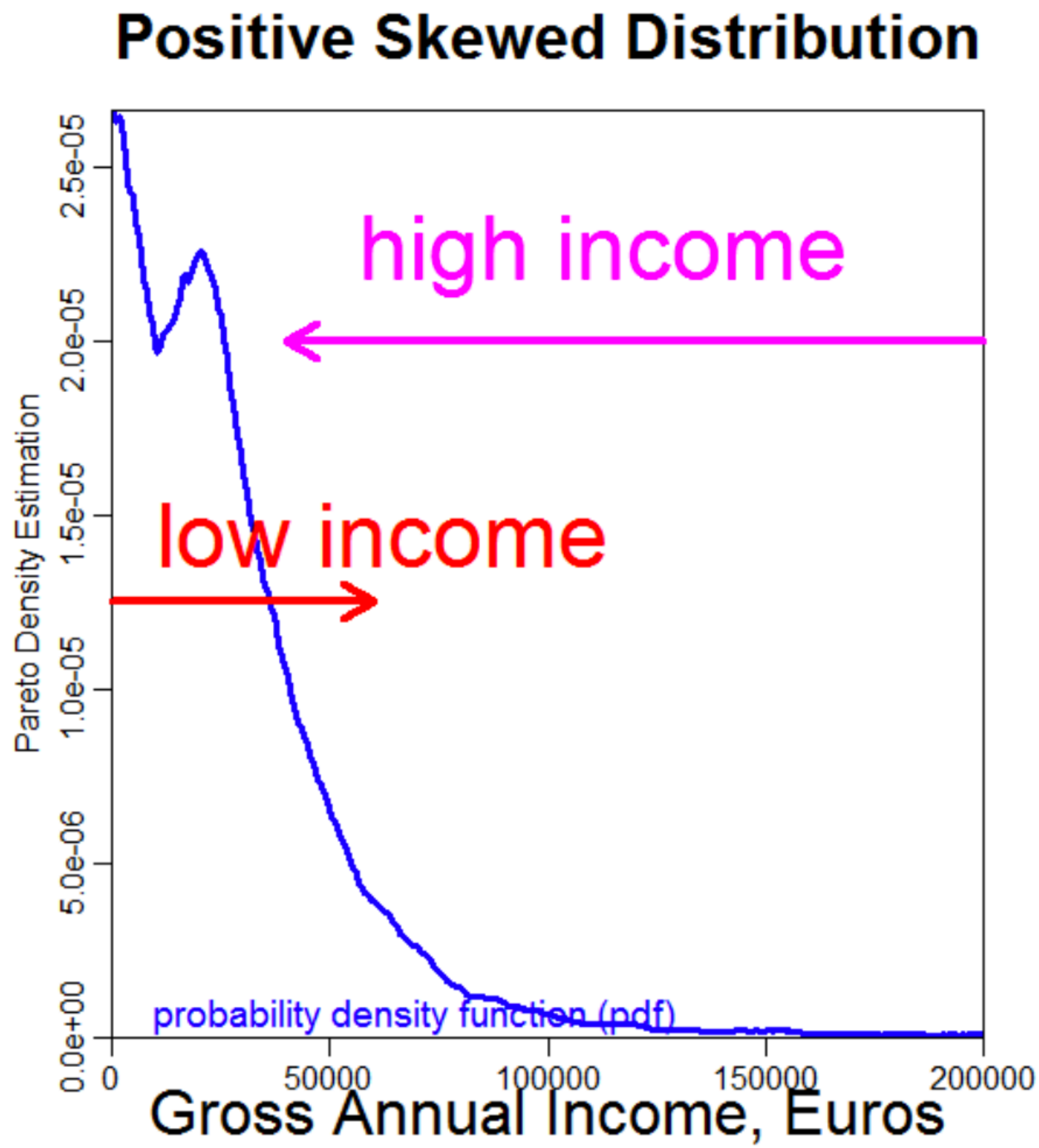
# Models of Income Distributions for Knowledge Discovery

Philipps Universität Marburg

# Income Distributions

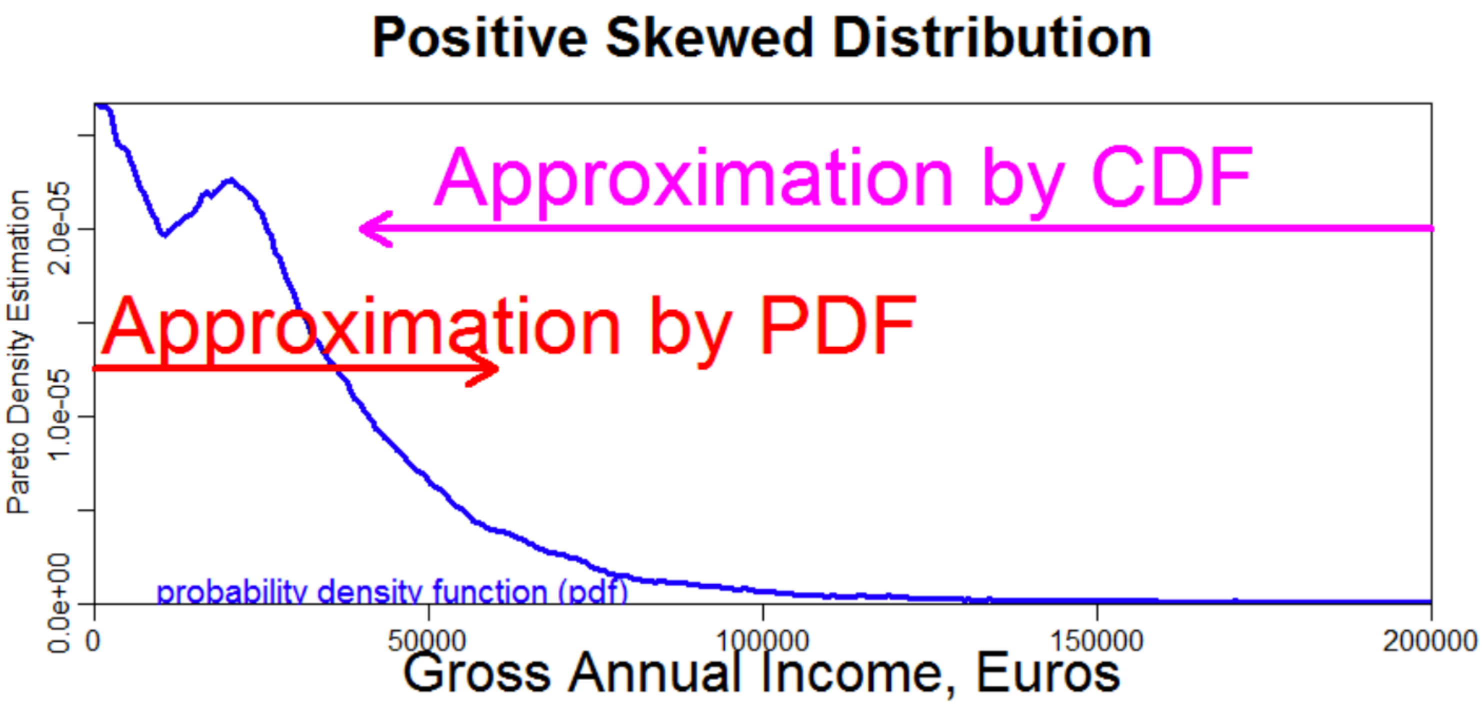
1. Always positively skewed with single mode and long tail [Kakwani 1980, p.14]
2. Properties of income are defined by various distributions

- Models often separate between the upper vs lower parts
  i. No systematic limit between **low** and **high** income
  ii. Different method for **low** and **high** income



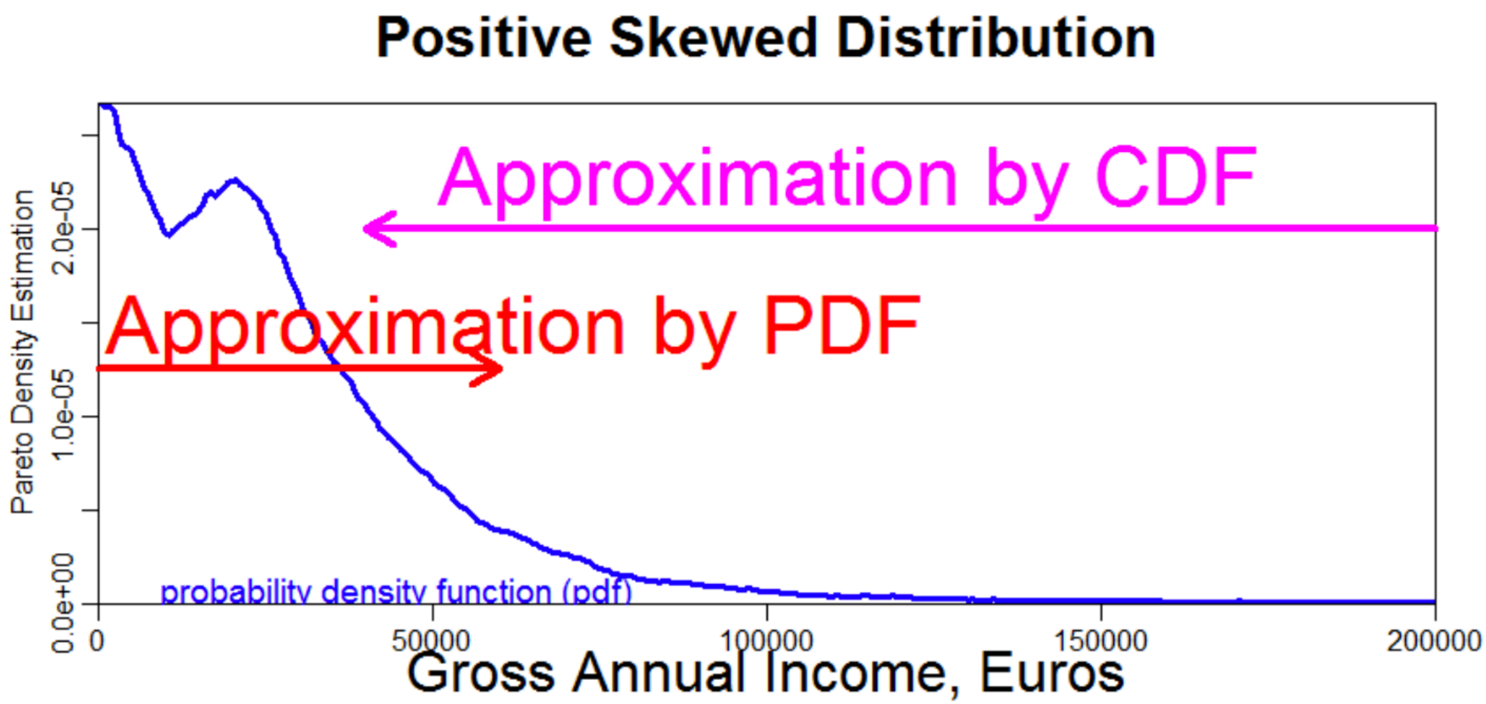
# Examples for Models of Income I

- **Low income**: various distributions, e.g.
  - *Approximation of probability density function (**pdf**)*
  - Log-Normal distribution [Clementi and Gallegati, 2005]
  - Exponential distribution [Chakrabarti, 2006]
  - Gamma distribution [Ferrero 2004, Scafetta 2004]
  - Boltzmann-Gibbs distribution [Drăgulescu/Yakovenko 2001]



# Examples for Models of Income II

- **High income**: pareto distribution
  - *Approximation of cumulative distribution function (**cdf**)*
  - Pareto Power Law distribution [Chaterjee et al. 2005, Levy and Solomon 1997]
  - Covers about 40% of income [Kakwani 1980, p.20]



# Dataset

- gross annual income of German population in 2001
  - a detailed overview Campus-File of income tax statistics 2001
  - for public use [EVAS 73111] discloses a 1% sample
- Dataset of income is preprocessed:
  - Through anonymization process income higher than 500 000 was oversampled

  => we down sampled to 1%
- From now on: **Income:**=gross annual income in Germany

Now two examples are presented...

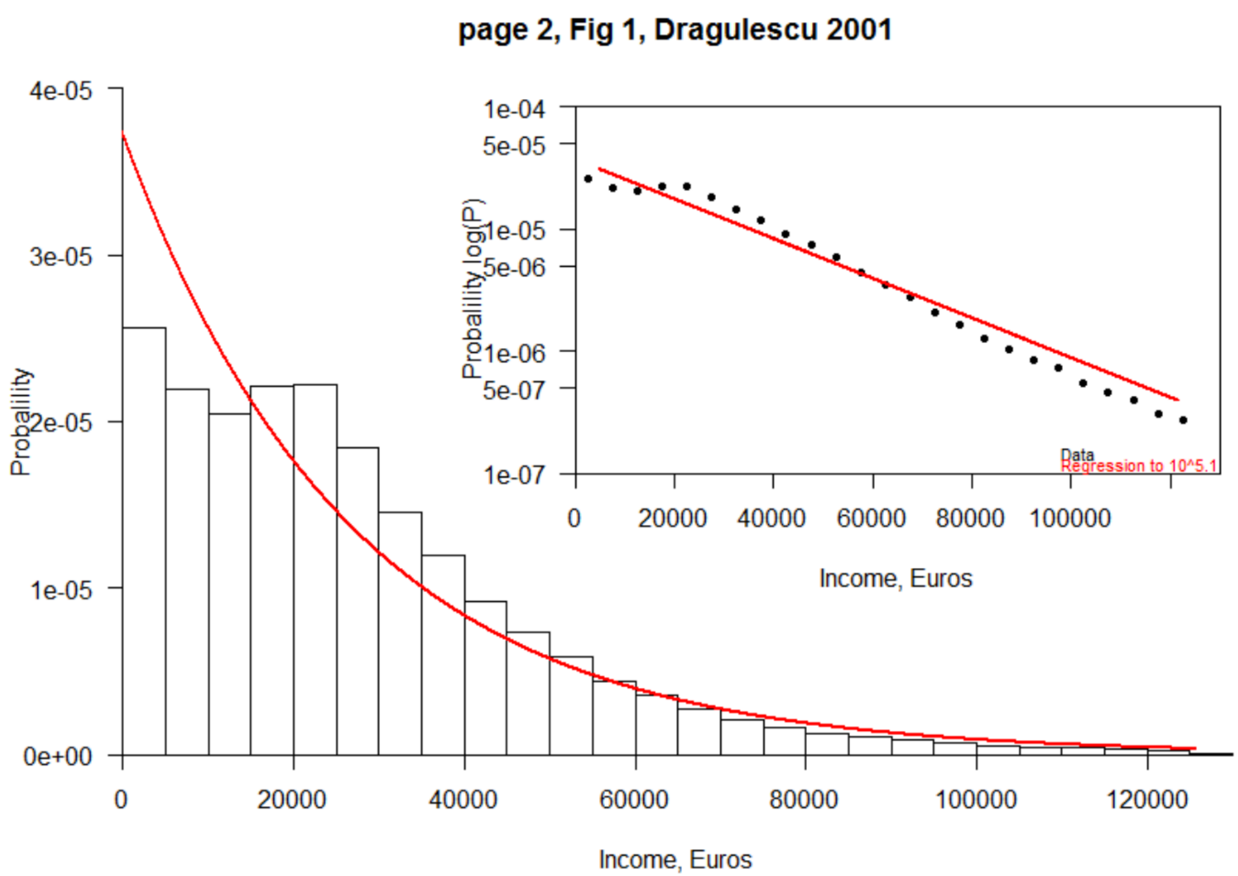
## Low Income with pdf

- **BLACK**: histogram of Income
- **RED**: Boltzmann-Gibbs

$$P(x) = \frac{1}{R} * \mathrm{e}^{-\frac{x}{R}}$$

- Regression and Fit of Range 0-126000 Euro

**pdf estimation of Income**



## High Income modeled with CCDF

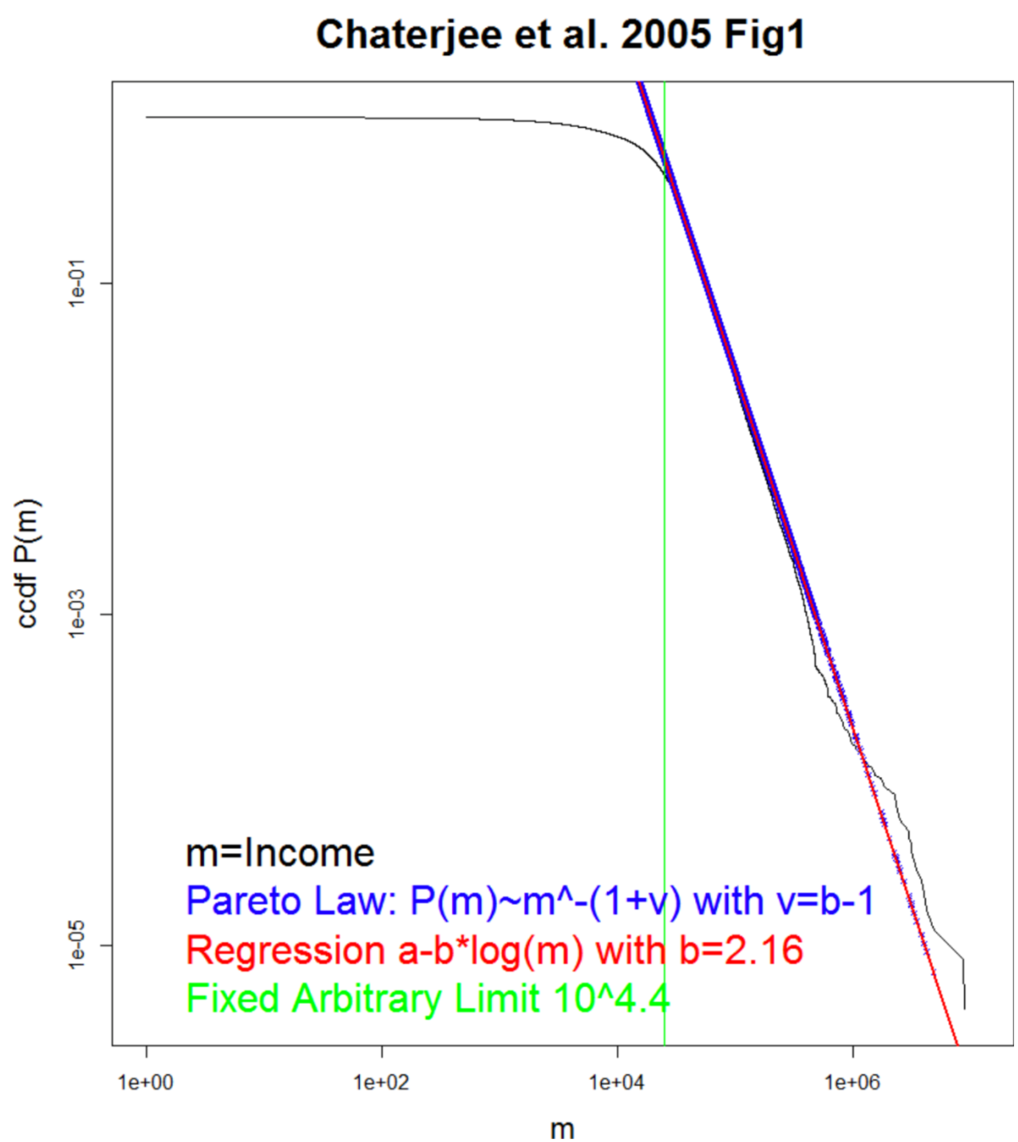
**BLUE**: follows Power Law with $P(m) = a * m^{-b}$

**RED**: linear Regression of data with $a - b * \log(x) = \log(P)$

**BLACK**: Income

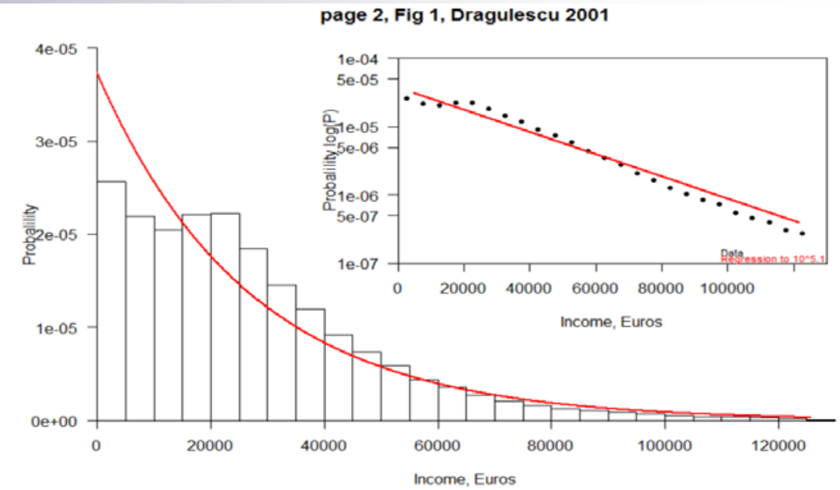
- Log/log plot of $\mathbf{1 - cdf(m) = ccdf(m)}$
- Regression begins somewhere at $10^{4.4} \approx 25000$ Euro
- Imprecise fit for $m > 10^{5.7} \approx 500000$ Euro

## Complementary Cumulative Distribution Function (CCDF)

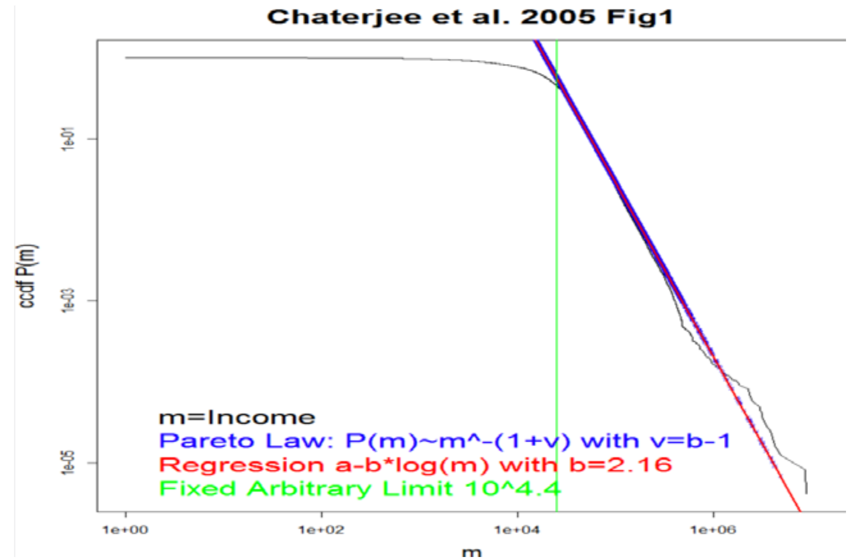


# Problems

- limit for the range of low income unclear
- Right Choice for number and width of bins critical for the right fit of the pdf
  - kernel density estimation with fixed radius




- No clear start point for high income
- Log linear approximation imprecise for vast income




**=> No systematic limit between low and high income**

# New Approach

i. Data logarithmic transformed
   - BoxCox λ$= 0.2$ with $p < 0.01$ [Asar et al. 2015]
ii. Pdf through pareto density estimation (**PDE**) [Ultsch 2005]
iii. Mixture of Gaussians with Toolbox
iv. Visual and statistical verification of model

-> Toolbox „Multimodal“ available in R on CRAN(http://cran.r-project.org/)

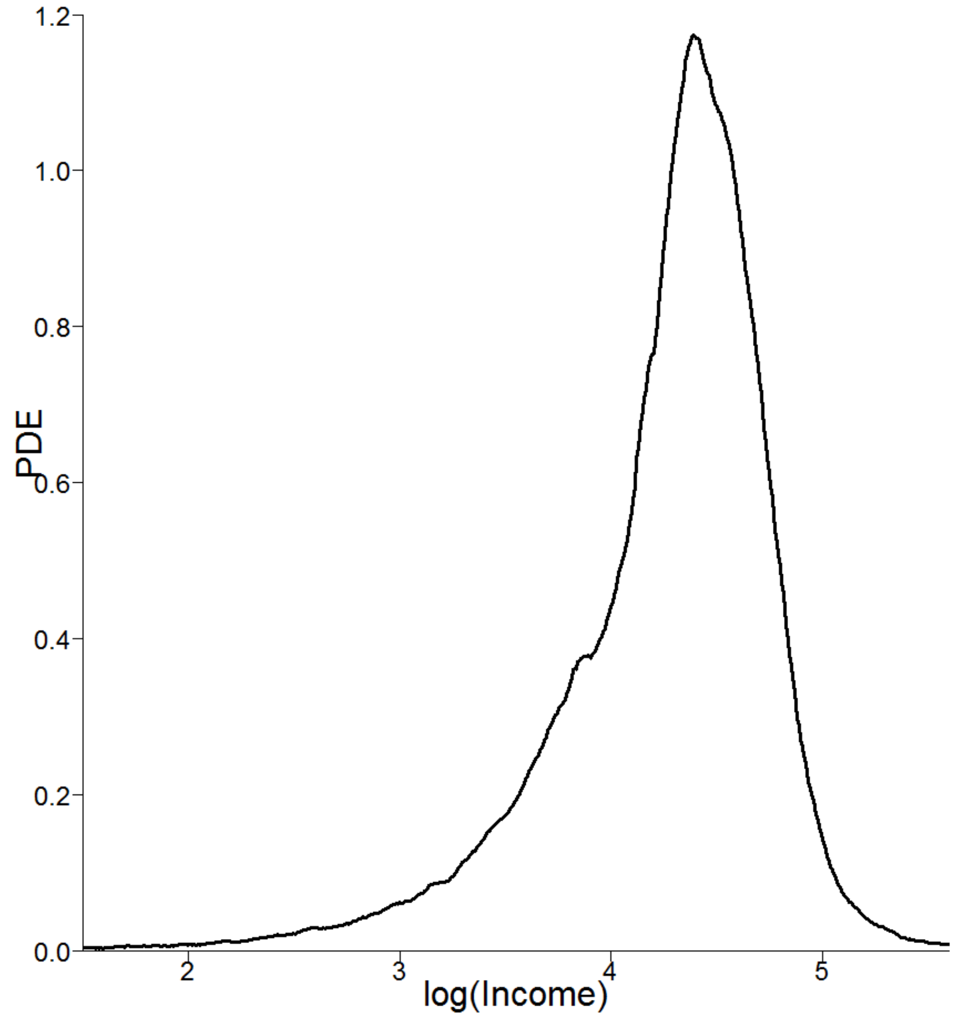
## Estimation of pdf

- Kernel density estimation with variable radius

  -> **PDE** is designed in particular to identify groups in data [Ultsch 2005]

## How to estimate density states within?

**Pareto Density Estimation (PDE)**



# Gaussian Mixture Model (GMM)

- EM-algorithm [Press 2007] estimates a log Gaussian mixture of four density states (Components)

Blue: Components $\mathrm{N}(m_i, SD_i)$

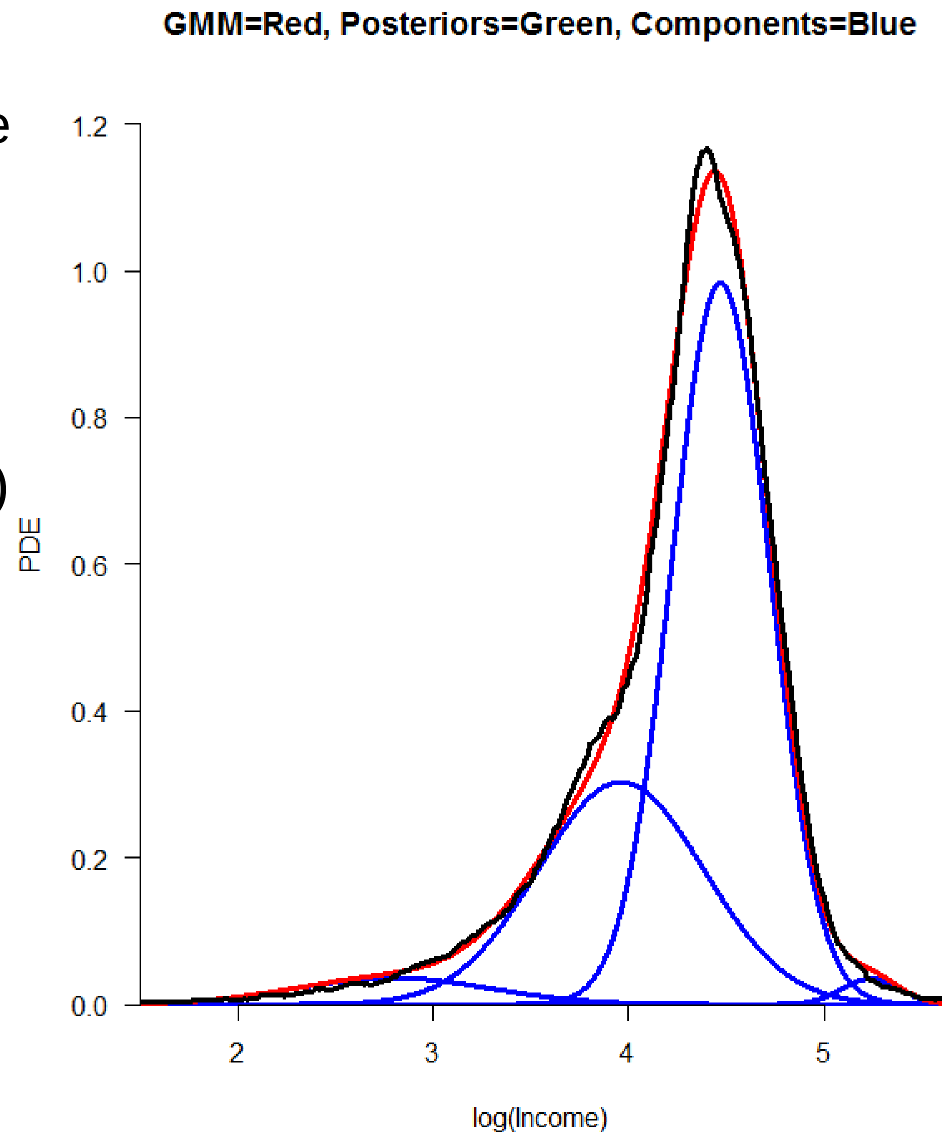
Red:

$$\mathrm{GMM}(x) = \sum_{\mathrm{i}=1}^{4} w_i * \mathrm{N}(m_i, SD_i)$$

$$\sum_{i}^{4} w_i = 1$$

$$\int GMM(x) = 1$$

***How do we calculate limits between components?***



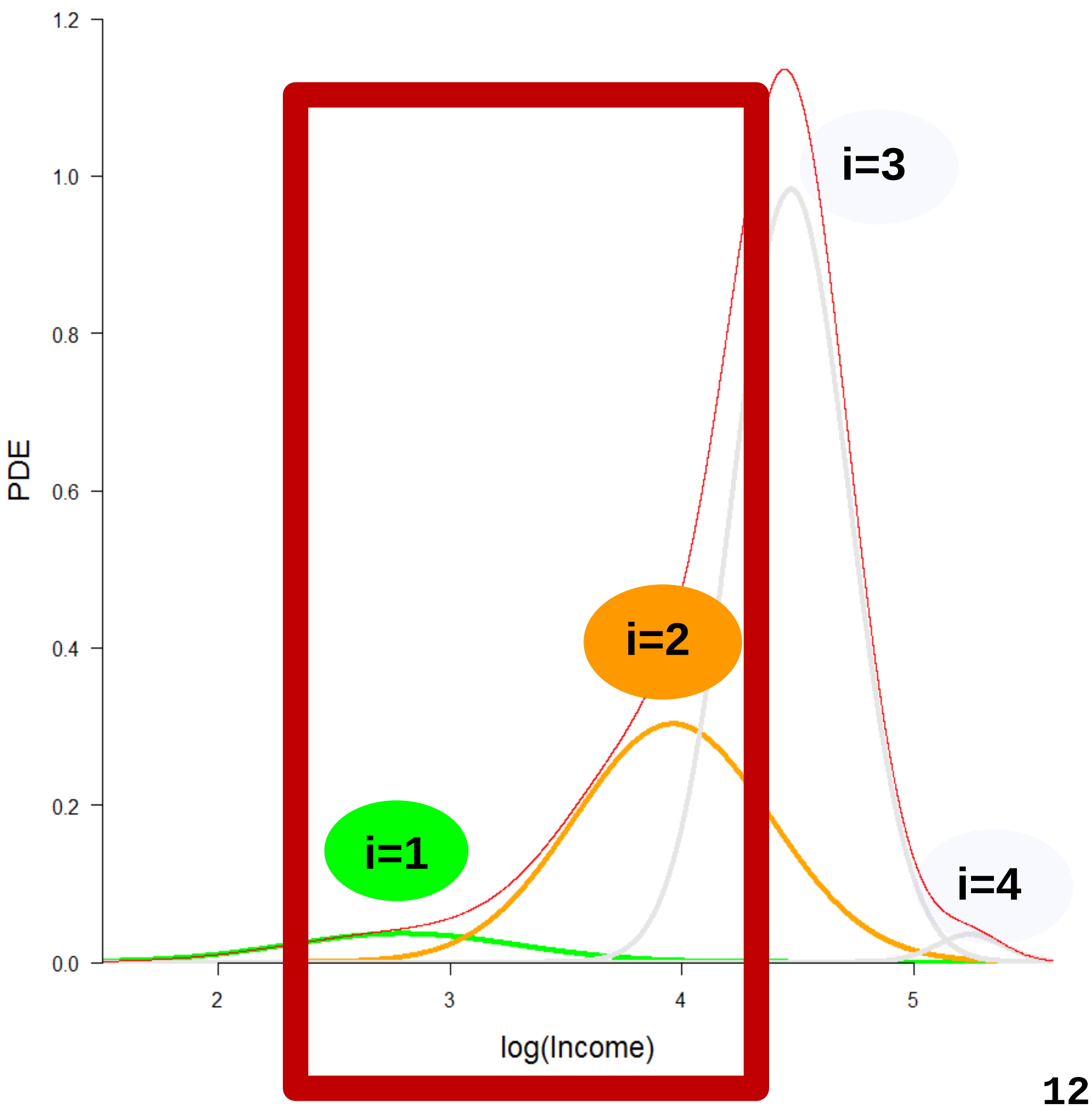
# Application of Bayes Theorem

- Through the likelihood to generate data in a component $c_i$ of the mixture, the conditional $p(x|\, c_i)$ we calculate the posterior $p(c_i|x)$

Blue: Components

Red: GMM(x)

*Example: Lets look at the red window with component $c_1$ and component 2 $c_2$*



# First Boundary in GMM

$$\mathrm{GMM}(x) = \sum_{\mathrm{i}=1}^{4} \mathrm{w_i} * \mathrm{N}(\mathrm{m_i}, \mathrm{SD_i})$$

$$= \sum_{\mathrm{i}=1}^{4} \mathrm{p}(c_i) * \mathrm{p}(\mathrm{x}|c_i)$$

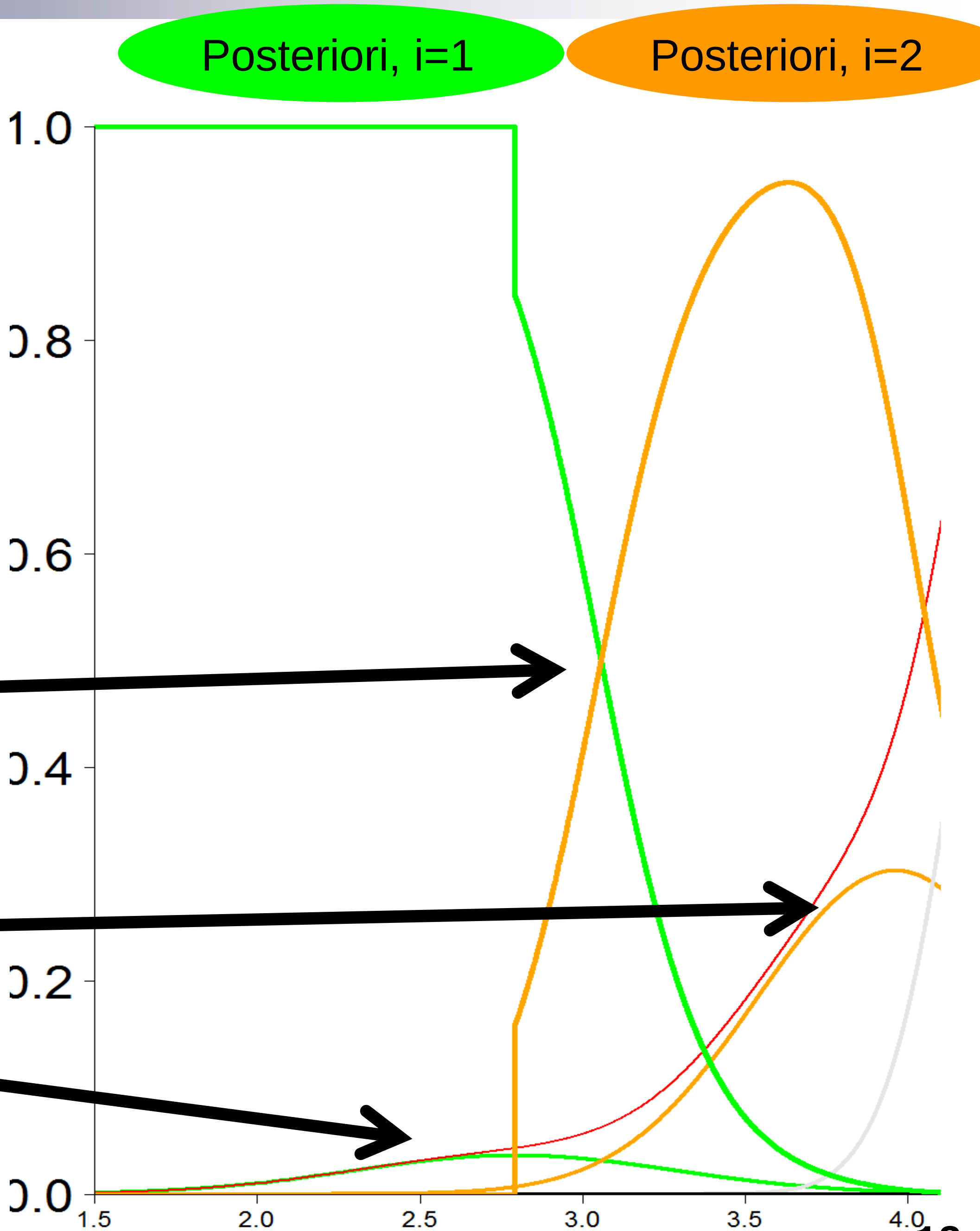
(Details, see Bayes theorem)

*Posteriori = 50%*

Mixture Components:

Orange: $N(m_2, SD_2), c_2$

Green: $N(m_1, SD_1), c_1$



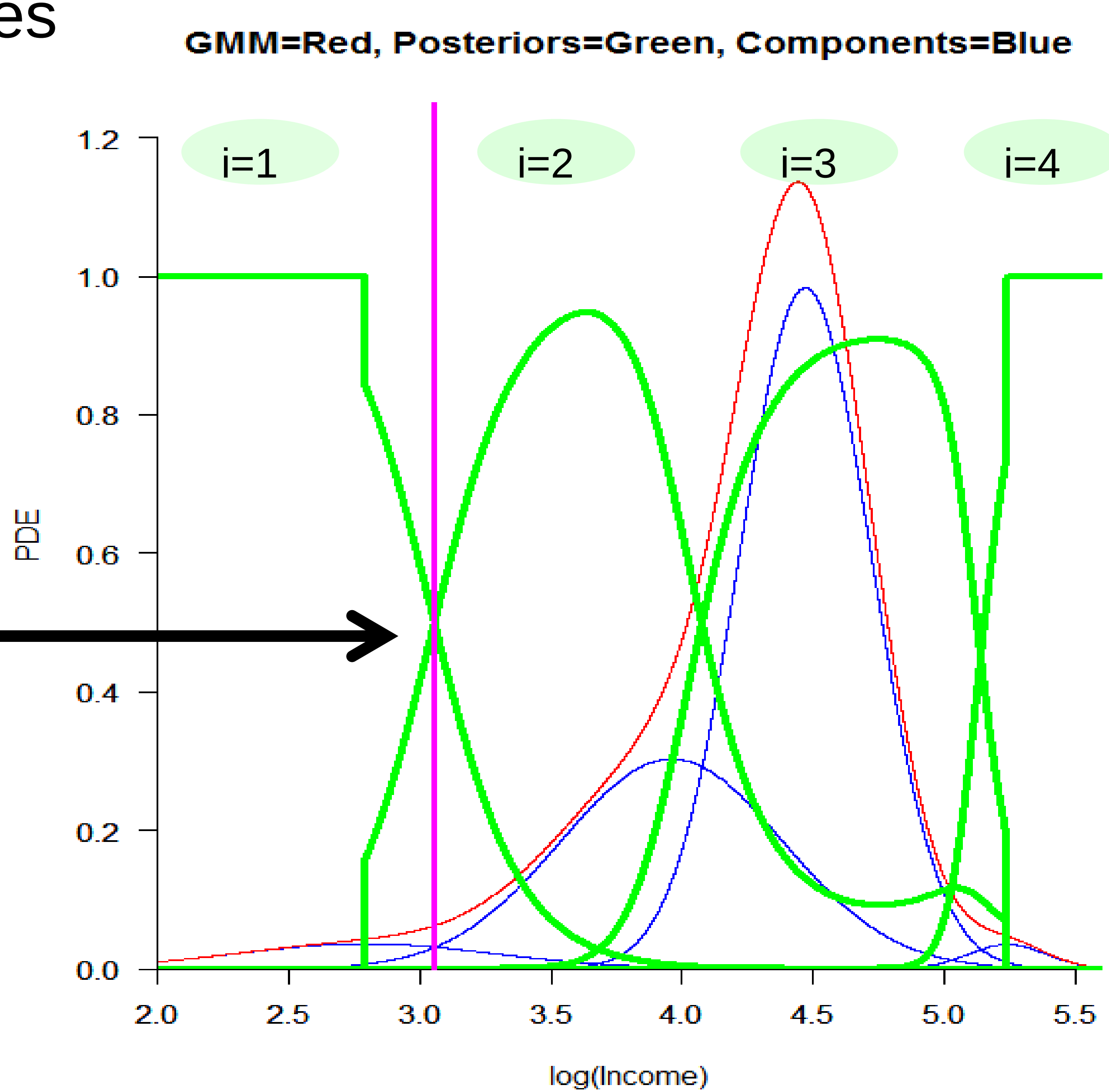
# Exact Boundaries

Green: Calculated posteriori of mixture components

$c_{i,}\ i = 1, \dots, 4$

Posteriori = 50%

$\Rightarrow$ Bayes Boundary between $i = 1$ and $i = 2$ (magenta)



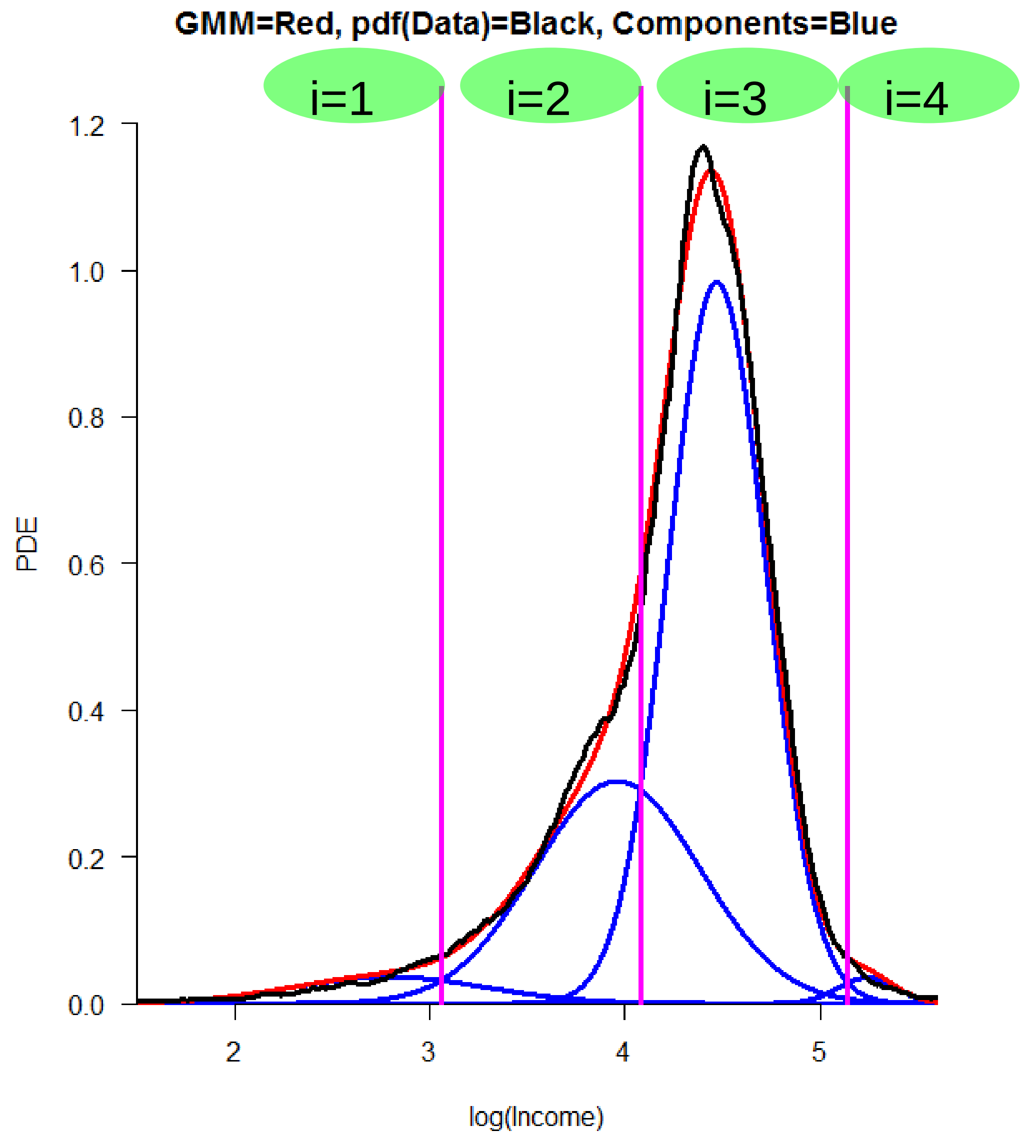
# GMM result for Income

Black = pdf(log(Data))

Magenta=Bayes Boundaries

Red=GMM

Blue=Components

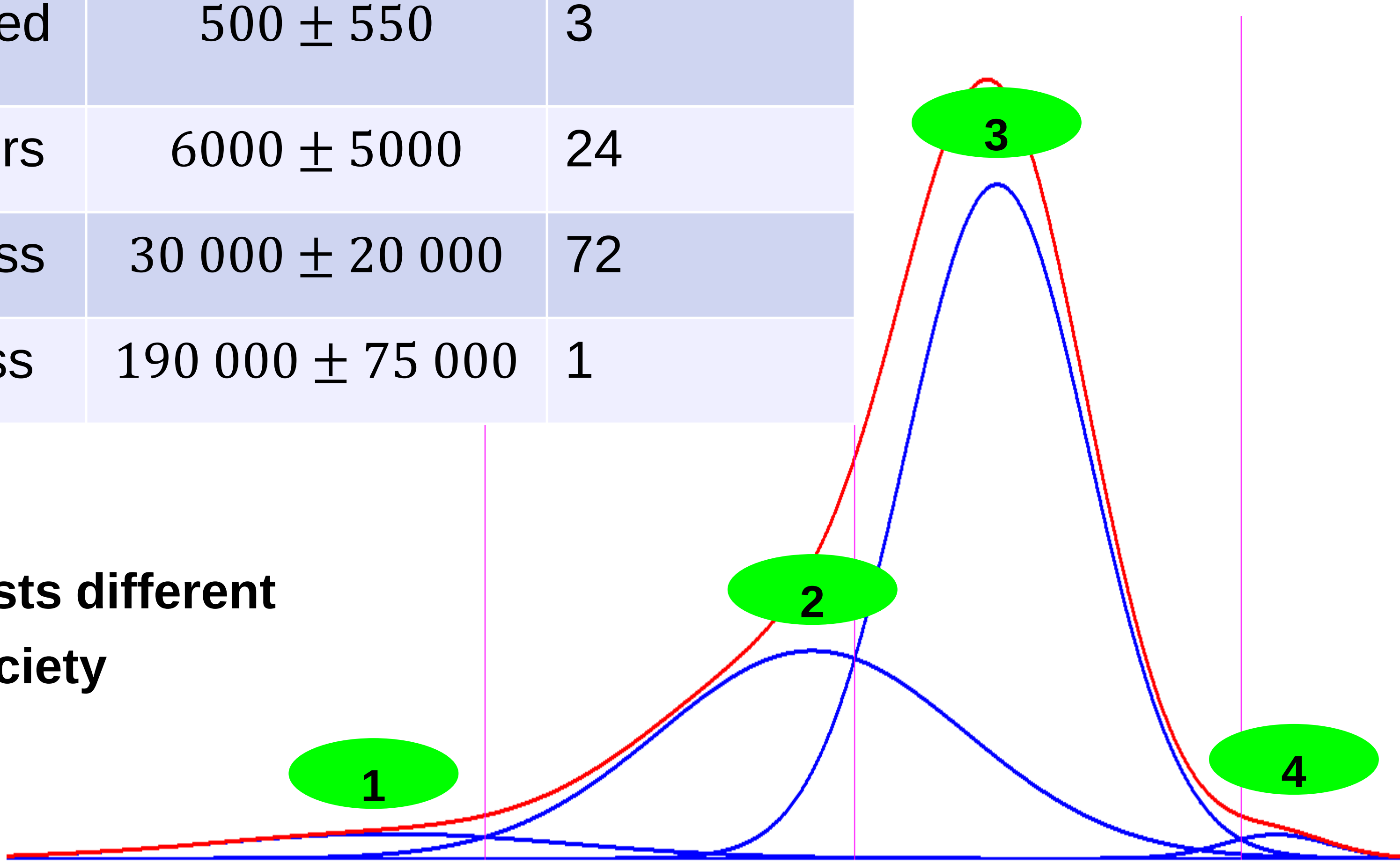



Range:

1. Group: 0-1100 Euro
2. Group: 1100-12000 Euro
3. Group: 12000 -139000 Euro
4. Group: > 139000 Euro

# Knowledge from Income Distribution

| No. | Group | Median $\pm$AMAD in Euro | Population in % |
|---|---|---|---|
| 1 | Unemployed | $500 \pm 550$ | 3 |
| 2 | Low earners | $6000 \pm 5000$ | 24 |
| 3 | Middle class | $30\,000 \pm 20\,000$ | 72 |
| 4 | Upper class | $190\,000 \pm 75\,000$ | 1 |

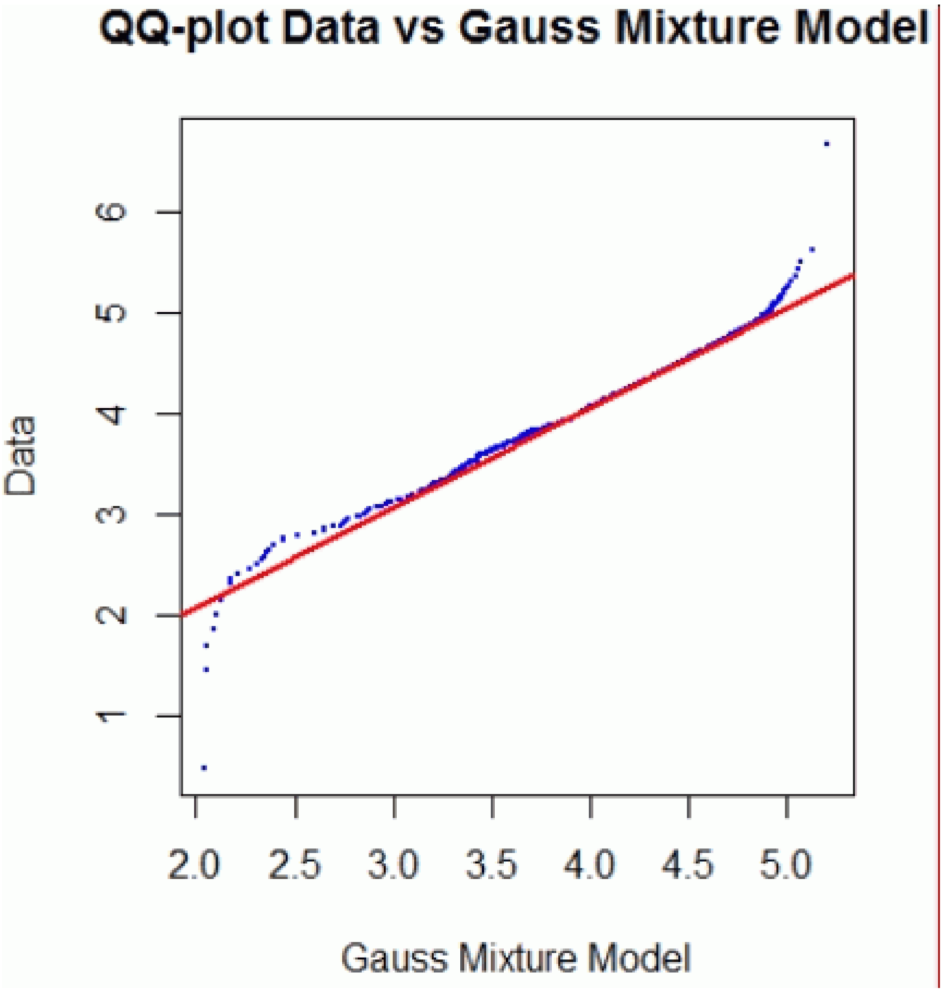



- **Model suggests different classes in society**

# Verification

- Statistical testing: Xi-Quadrat-Test: p<.001

- Visually: QQ plot
  - Compares two distributions by using n quantiles
  - Empirical distribution vs known distribution
  - *If straight line: distributions equal*



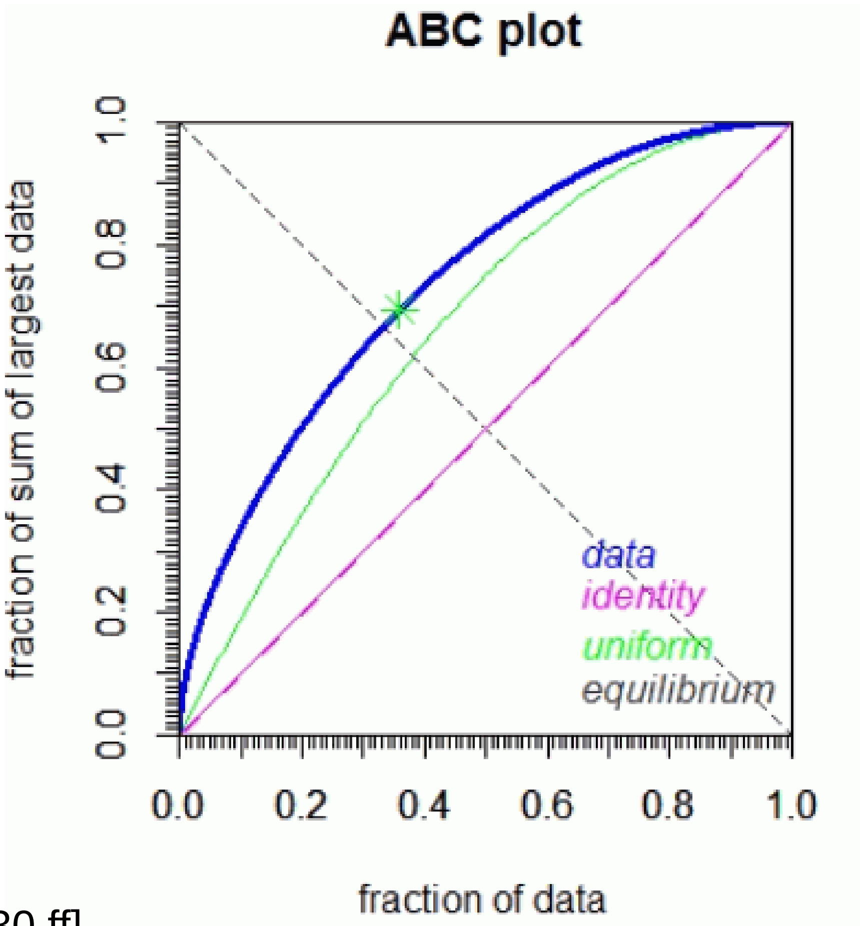
# Inequality – A Property of Distributions

- Instead of comparing income data by pdf or cdf, use *ABCplot* [Ultsch, Lötsch 2015]

- graphical representation of a upturned Lorenz Curve L(P),
- **Equals**: ABC(p)=1-L(1-p)
- **BUT**: Comparing inequality of data to uniform distribution instead of identity distribution

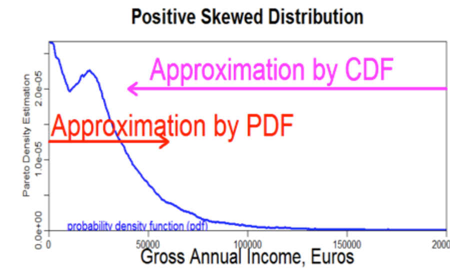
- inequality distribution is more skewed if above uniform distribution

ABCanalysis on CRAN



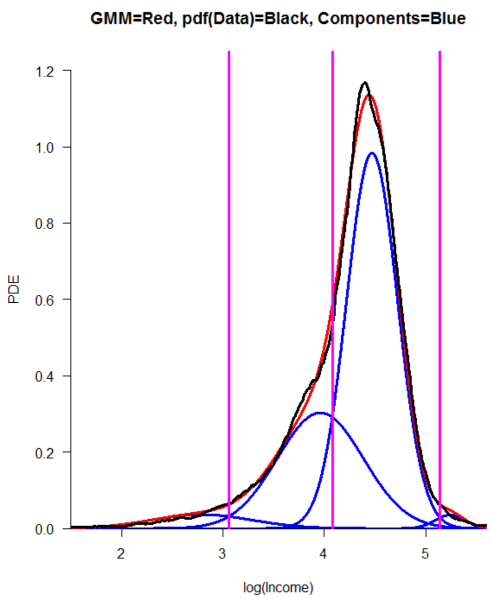

For L(p) see [Kakwani 1980, p.30 ff]

# Summary




**Previous models have the disadvantages:**

- No systematic limit between high and low income
- Inconsistent analysis methods: pdf vs cdf
- Do not explain whole range of Income




**Our model is**

- Simple mathematics founding (Bayes)
- Good fit of the whole range income
- Easily understandable and reproducible

**Open problem:**

- Which parameters of log transformed income do describe the income distribution itself?

# Sources


N. Kakwani, Income Inequality and Poverty, A World Bank Research Publication, Oxford University Press, Oxford 1980

Chatterjee, A., Chakrabarti, B. K., & Stinchcombe, R. B. (2005). Master equation for a kinetic model of a trading market and its analytic solution. *Physical Review E, 72*(2), 026126_026121-026126_026124.

Drăgulescu, A., & Yakovenko, V. M. (2001). Evidence for the exponential distribution of income in the USA. *The European Physical Journal B-Condensed Matter and Complex Systems, 20*(4), 585-589.

Lohn- und Einkommensteuerstatistik (EVAS 73111), Statistische Ämter des Bundes und der Länder, http://www.forschungsdatenzentrum.de/bestand/lest/index.asp, 11.2014 15:15

Campus-File der Einkommensteuerstatistik 2001, Statistisches Bundesamt, Gruppe „Steuern", VID-37313100-04, Wiesbaden, Jan 2008

Asar, O., Ilk, O., Dag, O. (2015). Estimating Box-Cox Power Transformation Parameter via Goodness of Fit Tests. Accepted to be published in *Communications in Statistics - Simulation and Computation*

Press, W.H., *Numerical recipes 3rd edition: The art of scientific computing*. 2007: Cambridge university press.

Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern classification. 2nd.* Edition. New York, 2001, p 512 ff

Ultsch, A., Pareto Density Estimation: A Density Estimation for Knowledge Discovery, in Innovations in classification, data science, and information systems, Springer, New York, pp 91-100, 2005.

Chakrabarti, A.S. and B.K. Chakrabarti, Statistical theories of income and wealth distribution. Economics: The Open-Access, Open-Assessment E-Journal. 4, p. 4, 2010.

Clementi, F. and M. Gallegati, Pareto's law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States, in Econophysics of wealth distributions, Springer, New York, pp 3-14, 2005.

Ferrero, J.C., The statistical Distribution of Money and the Rate of Money Transference. Physica A: Statistical Mechanics and its Applications, 341, p. 575-585, 2004.

Scafetta, N., S. Picozzi, and B.J. West, An out-of-equilibrium Model of the Distributions of Wealth. Quantitative Finance. 4(3): pp 353-364, 2004.

# Thank you for listening, any Questions?

# Boundaries by using Bayes Theorem

**Prior:**

Probability to choose a class

**Conditional Probability:**

Likelihood to generate data in this class

**Posterior:**

Probability, that data x is in class $c_i$

$$p(c_i|x) = \frac{p(c) * p(x|c_i)}{\sum_{i=1}^{L} p(c_i) * p(x|c_i)}$$

$$\sum_{i=1}^{L} p(c\_i) = 1$$

$$\sum_{i=1}^{L} p(c\_i \mid x) = 1$$

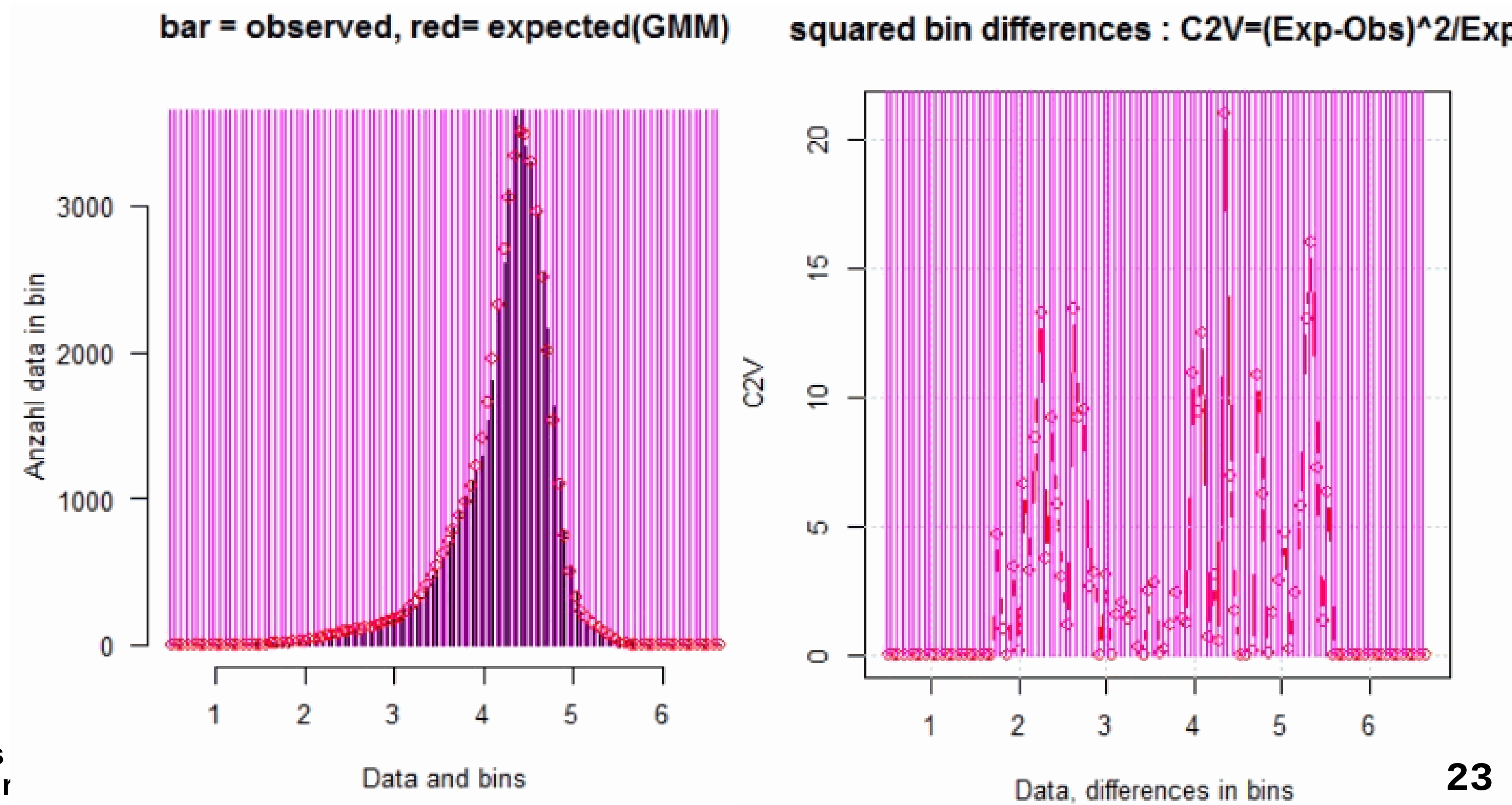
**Normalization, equals**

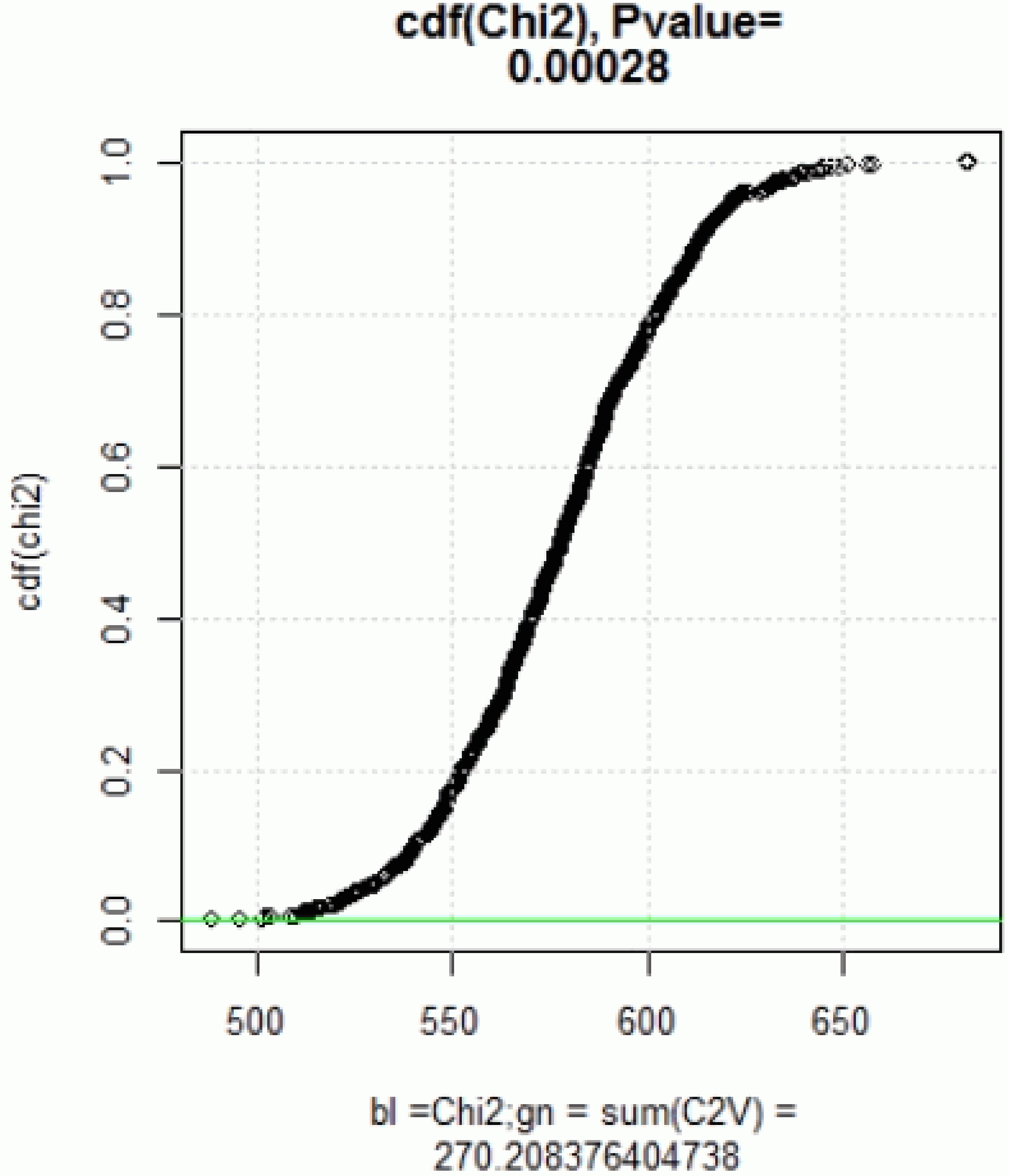
$$\sum_{\mathrm{i}=1}^{\mathrm{L}} w_i * \mathrm{N}(m_i, SD_i)$$

# Xi-Quadrat-Test

- Xi-Quadrat-Test
  - Let m be the number of Bins, $E_i$ one expected and Oi one Observed Bin, then the test statistics (C2V) is
  - degree of freedom is m-2

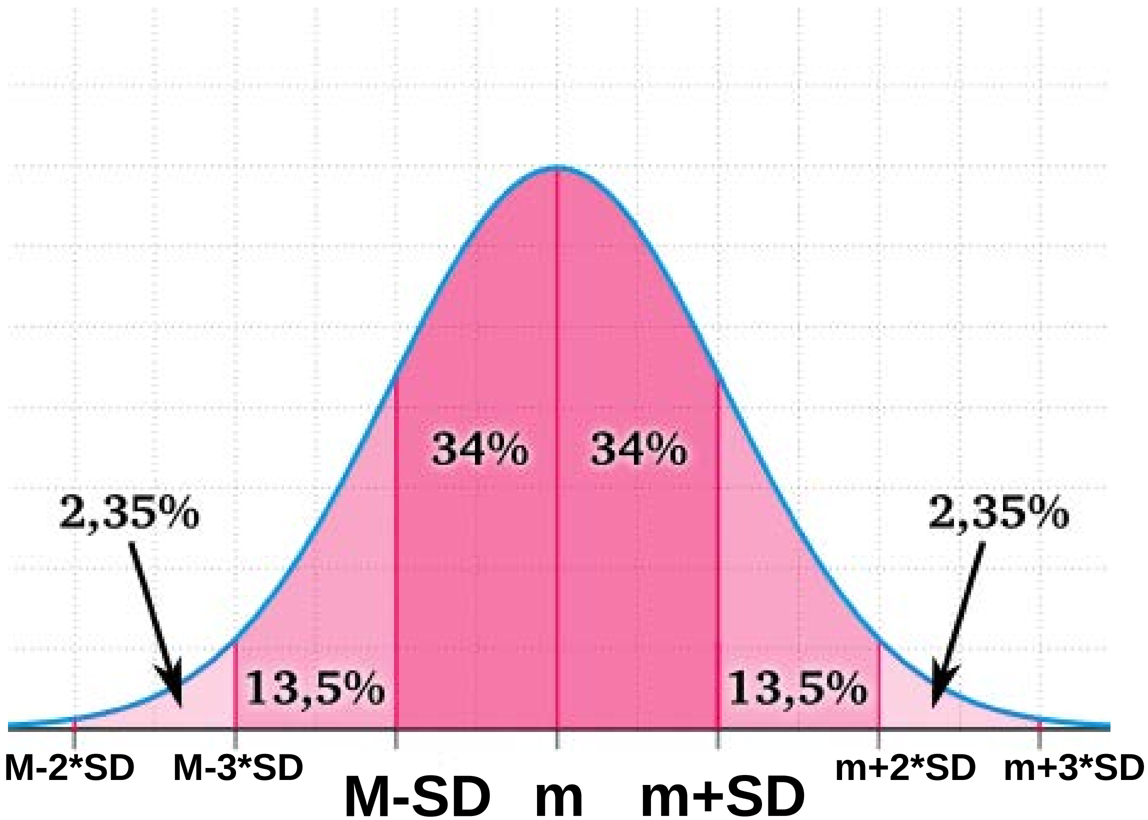
$$\boldsymbol{C2V} = \sum_{i=1}^{m} \frac{(E_i - O_i)^2}{E_i}$$




cdf(Chi2), Pvalue=
0.00028
cdf(chi2)
1.0
0.8
0.6
0.4
0.2
0.0
500
550
600
650
bl =Chi2;gn = sum(C2V) =
270.208376404738

# Definition Gaussian (pdf)

$$N(m_i, SD_i) = \frac{1}{\sqrt{2\pi * SD^2}} \exp\left(-\frac{(x-m)^2}{2 * SD^2}\right)$$




Source of Img: http://matheguru.com/images/normalverteilung_68-95-99.png