Distribution Fitting for the Lazy Scientist

or, All you wanted to know about maximum likelihood estimation but were too afraid to ask.

Christopher Quarles
SFI Complex Systems Summer School 2019

SANTA FE INSTITUTE

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Outline

1. Distribution Fitting: What, Why, and When?

2. Fitting the distribution with MLE

3. Goodness of fit

Regression Example:

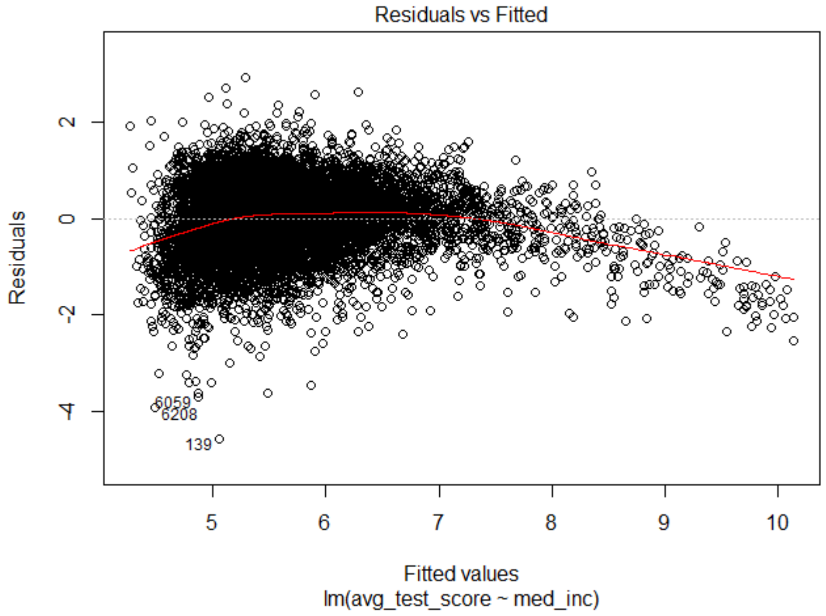
Is income in a school district related to students' test scores in that district?

Variables: average test score, median income (1000's)

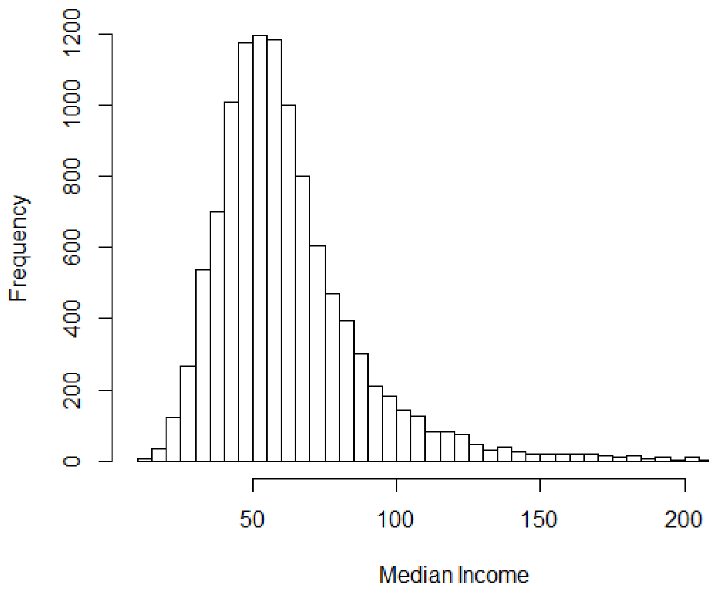
```
lm(formula = avg_test_score ~ med_inc, data = dat)

              Estimate   Std. Error     t value    Pr(>|t|)
(Intercept)    3.9295      0.0168        233.7      <2e-16 ***
med_inc        0.0272      0.0002        109.6      <2e-16 ***

Residual standard error: 0.6962 on 11052 degrees of freedom
Multiple R-squared:  0.5209,  Adjusted R-squared:  0.5209
```

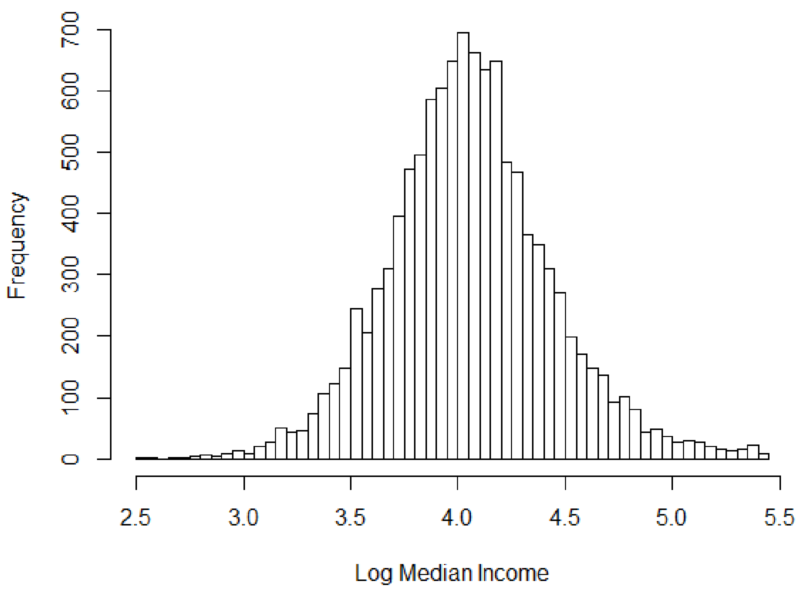
estimated parameters

How confident are we?



Is median income:





Normal?

Additive

Log-Normal?

Multiplicative

1. What are the parameters of each distribution?
2. Which distribution is better?

These represent different processes!

Fit the distribution

"I have a good model for how my data was generated, and want to know the parameters."

- Poisson process
- Power law: Find α in $x^{-\alpha}$.
- Network community detection: SBM

Check whether it fits well enough

"I want to see *which* distribution fits better"

OR

"I want to *test* whether my model fits the data."

It usually doesn't!

How to do MLE:

1. Choose a model

2. Find the log-likelihood function

3. Find the parameters that maximize the log-likelihood
    - Calculate estimators by hand (or look them up)
    - Computationally

Fitting the Distribution

model with parameters $\vec{\theta}$

independent data points $\vec{x} = \{x_1, x_2, x_3, \dots, x_n\}$

Likelihood of a given model

$$\mathcal{L}\left(\vec{\theta} \mid \vec{x}\right) = \mathrm{P}(\vec{\mathrm{x}} \mid \vec{\theta})$$
$$= \prod_i P\left(x_i \middle| \vec{\theta}\right)$$

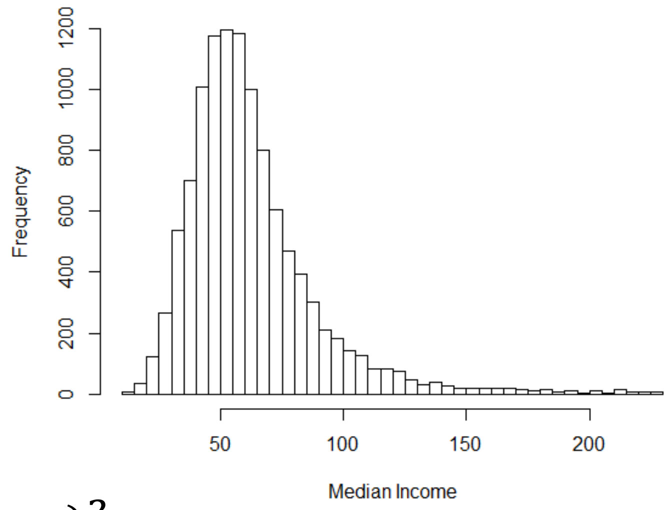
very small

Log-likelihood

$$\log \mathcal{L} = \sum_i \log P\left(x_i \middle| \vec{\theta}\right)$$

maximize this

Example of MLE



model: normal $\mathrm{P}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{\frac{-(\mathrm{x}-\mu)^2}{2\sigma^2}}\ dx$

parameters: μ, σ

$$\mathcal{L} = \prod_i P(x_i|\mu,\sigma) = \prod_i \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

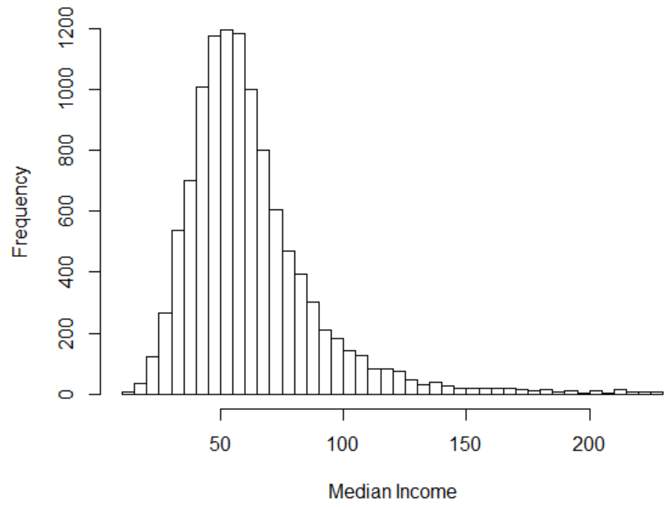
$$\log\mathcal{L} = \sum_i -\log(\sigma) - \log\left(\sqrt{2\pi}\right) + \frac{-(x_i-\mu)^2}{2\sigma^2}$$

$$= -n\log(\sigma) - n\log\left(\sqrt{2\pi}\right) - \sum_i \frac{(x_i-\mu)^2}{2\sigma^2}$$

Example of MLE

Goal: Find μ, σ that maximize

$$\log \mathcal{L}\,(\mu,\sigma) \;= -n\log(\sigma) - n\log\left(\sqrt{2\pi}\right) - \textstyle\sum_i \frac{(x_i-\mu)^2}{2\sigma^2}$$



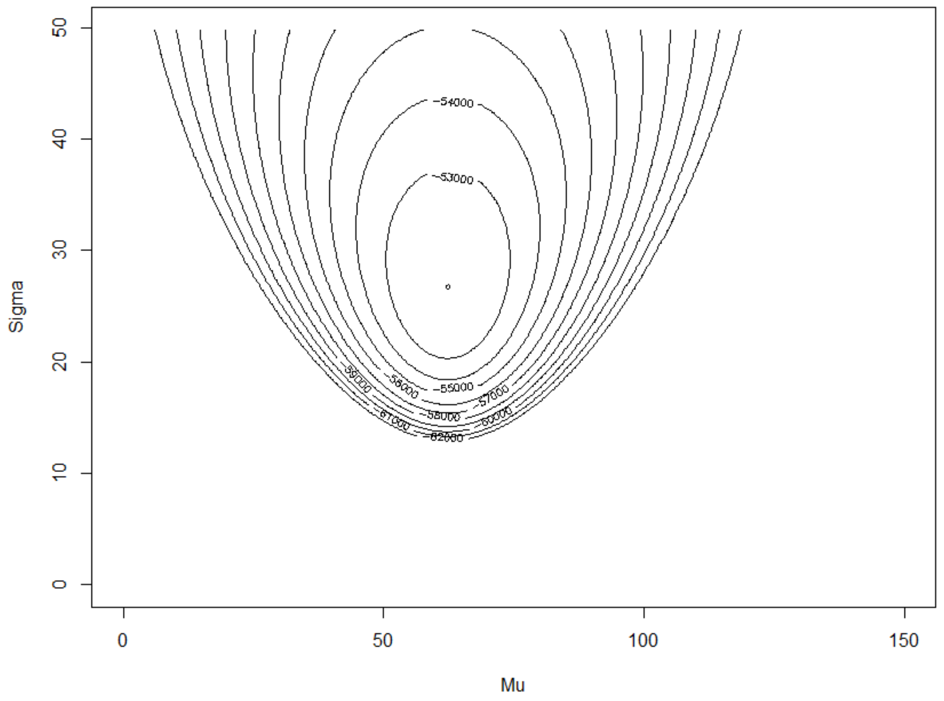
<u>Analytically</u>

$$\frac{d}{d\mu}\log\mathcal{L} = 0$$

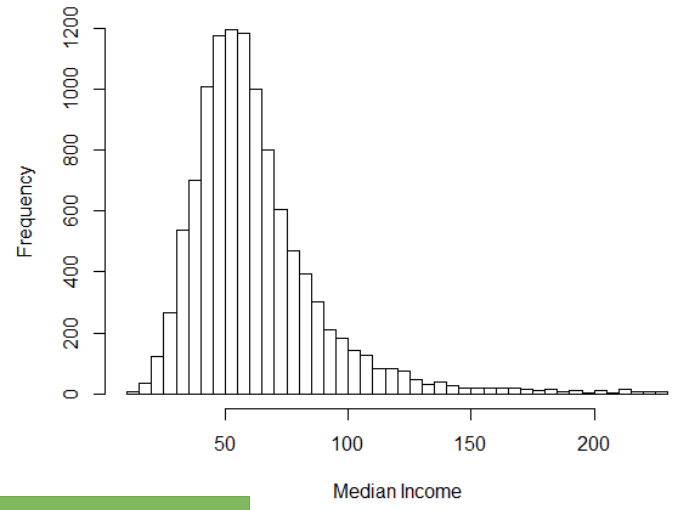
$$\frac{d}{d\sigma}\log\mathcal{L} = 0$$

let's do this!

<u>Computationally</u>



Example of MLE



Analytically

model: normal $\quad P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$

parameters: μ, σ

$$\left.\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_i x_i = \bar{x} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_i [x_i - \hat{\mu}]^2 = \overline{(x^2)} - \bar{x}^2 \end{aligned}\right\}$$

maximize likelihood estimators for the normal distribution

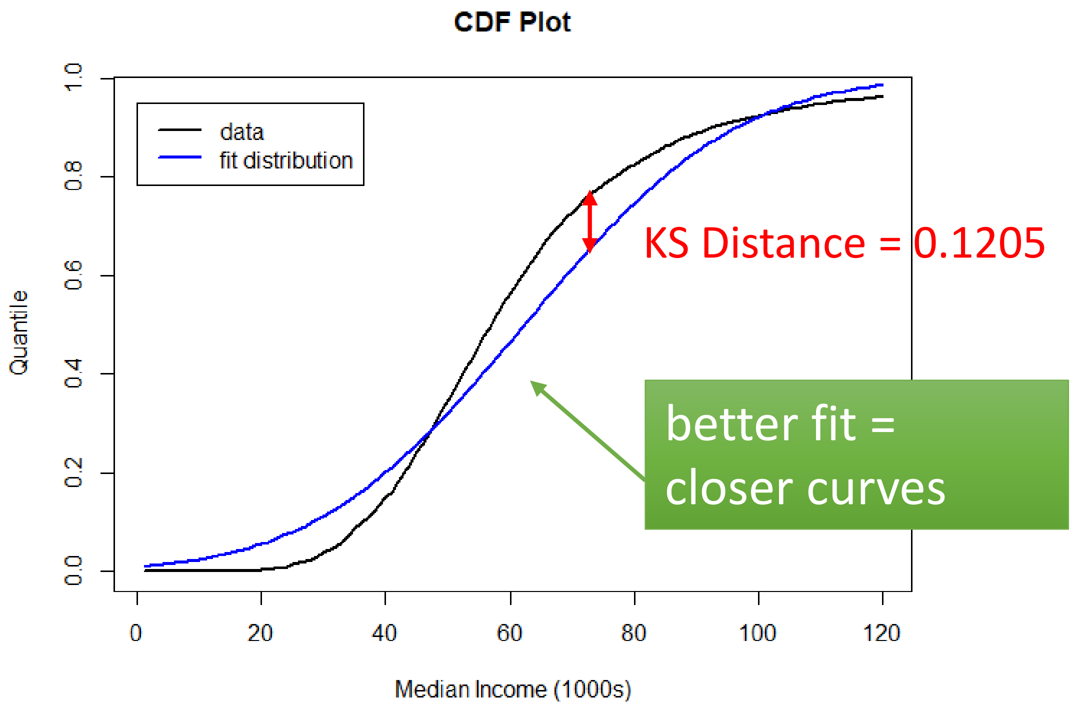
For our data: $\bar{x} = 62.34$, $\overline{(x^2)}$=4600

$\hat{\mu} = 62.34$

$\hat{\sigma} = 26.71$

$\log \mathcal{L}(\hat{\mu}, \hat{\sigma}) = -51999$

Measuring Goodness of Fit



<u>Some Goodness of Fit Statistics</u>

Kolmogorov-Smirnov (KS) Distance – Maximum distance between CDFs of data and best-fit distribution

Akaike Information Criterion (AIC) = 2(# of parameters) – 2(log-likelihood)

Bayesian Information Criterion (BIC) = $\ln(n)$(# of parameters) – 2(log-likelihood)

YOU CAN NOT COMPARE LOG-LIKELIHOODS OF NON-NESTED MODELS. Use AIC instead

better fit =
SMALLER statistics!

Measuring Goodness of Fit

“I want to see *which* distribution fits better”

1. Fit multiple distributions
2. Compare statistics and look at graphs
3. Choose the best distribution (or none)

“I want to *test* whether my model fits the data.”

"I want to *test* whether my model fits the data."

H_0: Your data is a set of random draws from this probability distribution
H_A: Your data is *not* drawn from this distribution

Strategy:

1. Fit the distribution
2. Generate synthetic distributions of the same size as your data using the probability distribution. Store their KS distance
3. Create a sampling distribution of KS
4. Calculate a p-value

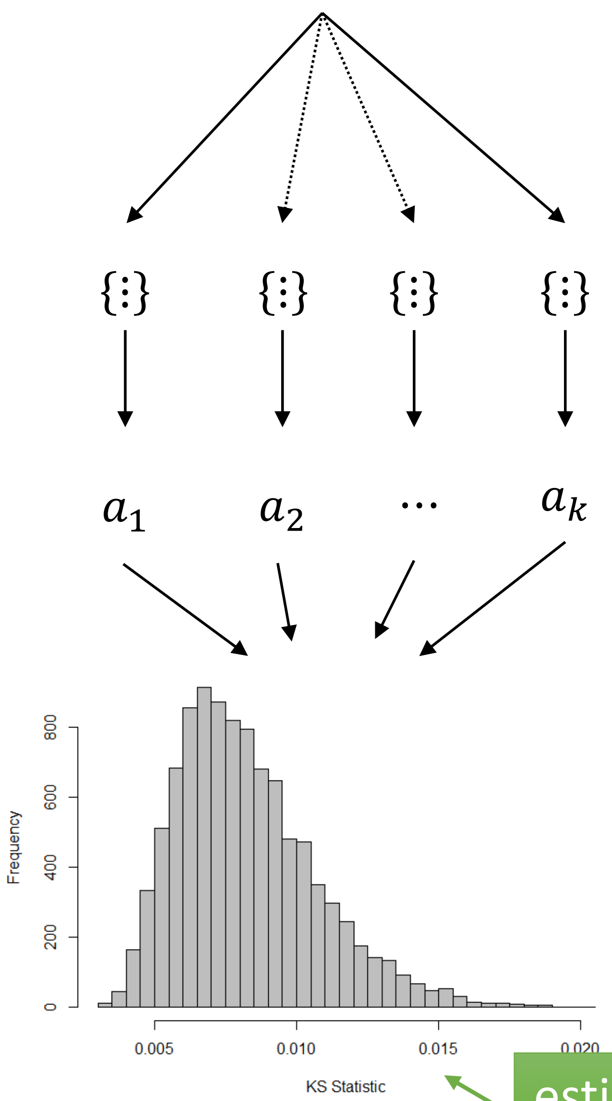
the distribution that we already fit to the data

$$\mathcal{N}(62.34, 26.71)$$

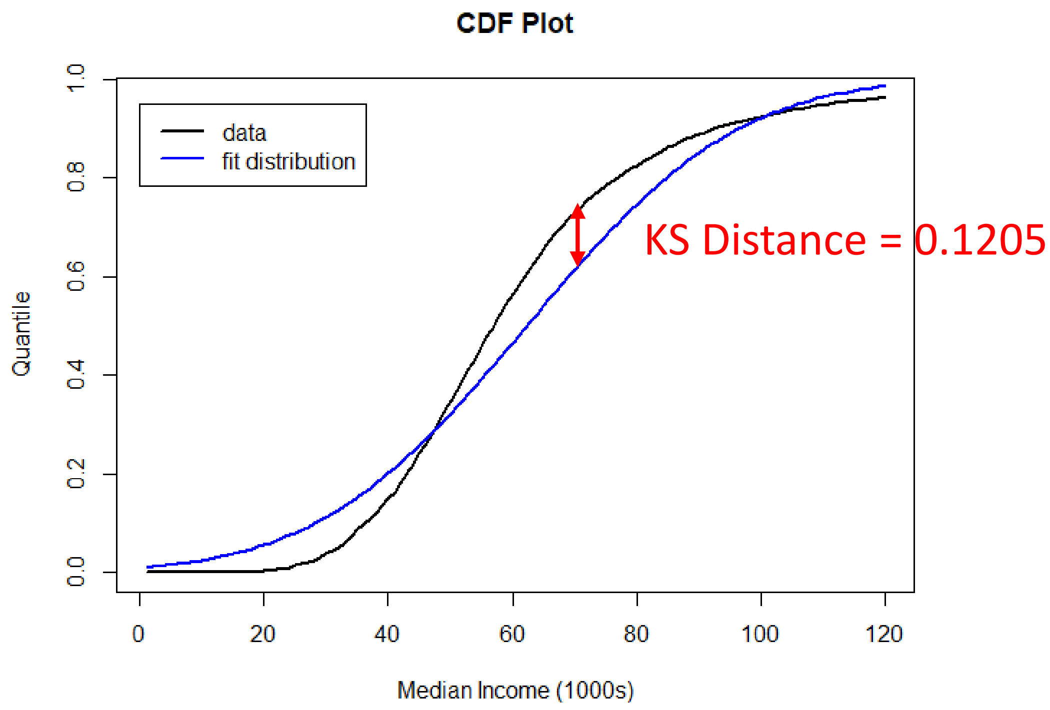
simulated samples n=11054

KS distances

sampling distribution



estimated sampling distribution ASSUMING H_0



Compare our test statistic (KS = 0.1205) to the sampling distribution

p-value = 1

Reject H_0

Our data was not drawn from a normal distribution

Your Turn

1. Download the data
2. Use MLE to fit the following distributions to the data:

log-normal

$$\mathrm{P}(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \mathrm{e}^{\frac{-(\log(\mathrm{x}) - \mu)^2}{2\sigma^2}} \, dx$$

$$\hat{\mu} = \frac{1}{n}\sum_i \log(x_i)$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (\log(x_i) - \hat{\mu})^2$$

power law with $x_{min} = 1$

$$P(x|\alpha) = \frac{\alpha-1}{x_{min}}\left(\frac{x}{x_{min}}\right)^{-\alpha}$$

$$\hat{\alpha} = ???$$

3. Compare the fits of the three distributions (normal, log-normal, power law), and decide which fits the best