Parametric Statistics

Week 7 - Fitting the Data

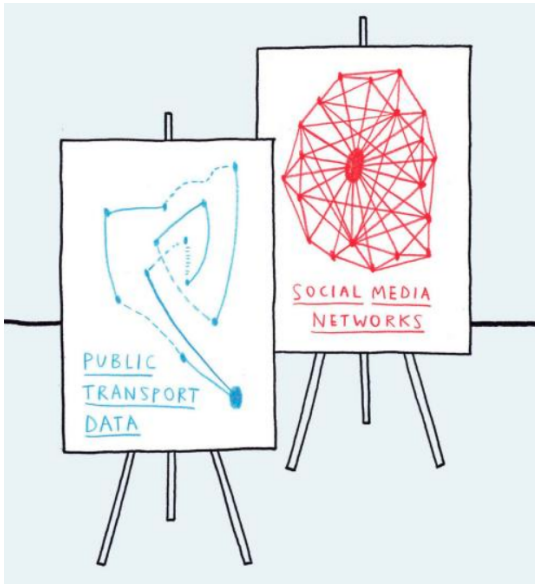
Juan Carlos Medina Serrano

Technische Universität München
Hochschule für Politik
Political Data Science

Munich, 04. December 2019

political data science

https://politicaldatascience.blogspot.de



Fitting the Data

We learned how to find out if our data is normally distributed, but how can we know which distribution goes better with our data = fits our data better.

The fitting process needs two ingredients:

- Finding the best distribution
- Finding the best estimators for the selected distribution's parameters

Normally, we start with the second step. We try to guess which distribution could be better for our data based on our statistics knowledge.

We can do this for several distributions, and then compare which distribution is the best.

Finding the Best Estimators

We need a measure to select the best estimators:

The **Likelihood**!

Remember from our first lecture:

If X and Y are independent :

$$P(X, Y) = P(X) * P(Y)$$

The conditional probability

$$P(X \mid Y)$$

The Likelihood

Likelihood calculates how possible is that we collected this data given a parameter (or parameters).

$$\mathscr{L} = P(X|\theta)$$

Important: The probability depends on the distribution we choose! Since we assume the data to be independent of each other, we can multiply the probability of each data point for the given parameter(s):

$$\mathscr{L} = P(X|\theta) = \prod_{i=1}^{N} P(x_i|\theta)$$

We want to have the parameter estimators that **maximize** the likelihood

The Maximum Likelihood

How to find a maximum?

We could manually try different parameters...would take a long long long time
In mathematics there is a better way to find a maximum: Calculate the derivative and equal to zero.

The Maximum Likelihood

How to find a maximum?

Calculate the derivative and equal to zero.

Example We throw a coin 10 times. We want to estimate the best parameter θ_T that corresponds to the probability of tails (T). This is the data:

$$H, T, H, H, T, H, H, H, T, H$$

$$\mathscr{L} = P(H, T, H, H, T, H, H, H, T, H|\theta_T) = (1-\theta_T)\theta_T(1-\theta_T)(1-\theta_T)\theta_T(1-\theta_T)(1-\theta_T)(1-\theta_T)\theta_T(1-\theta_T)$$

$$\mathscr{L} = \theta_T^3(1-\theta_T)^7$$

$$\frac{d}{d\theta_T}(\theta_T^3(1-\theta_T)^7) = 3\theta_T^2(1-\theta_T)^7 + 7\theta_T^3(1-\theta_T)^6 = 0$$

$$\hat{\theta_T} = 0.3$$

Which we knew already by common sense of seeing 3 tails in 10 throws, $3/10 = 0.3$

Note Most of the time we calculate the logarithm of the likelihood, since it is easier to find the derivative this way.

Distributions

We now know how to find the best parameter estimators for our data, but for this we need to first select a possible distribution.

We need to get to know the most common distributions and their use cases.

Bernoulli Distribution

Checklist

- Discrete Data
- One trial
- Only two trial outcomes: success and failure
- **Parameter:** θ
- Notation: $X \sim Ber(\theta)$

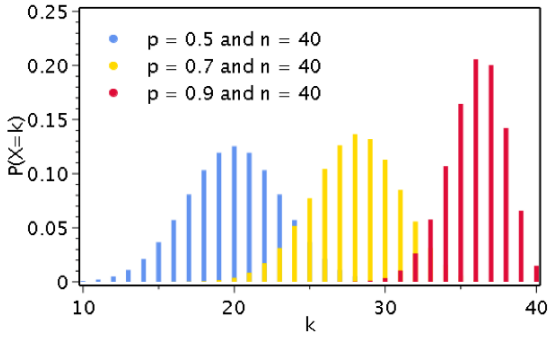
Example: Throwing a coin **ONE** time, with tails being success and heads failure. (we define success/failure)

θ is the probability of success $1 - \theta$ is the probability of failure

Binomial Distribution

Checklist

- Discrete Data
- More than one trial = n trials.
- Only two trial outcomes: success and failure
- Probability of success is the same in each trial (θ)
- Trials are independent
- Outcome: Number of successes (k)
- **Parameters:** θ, n, k
- Notation: $X \sim B(k, n, \theta)$



Binomial Distribution

The previous example with the 10 coin throws and θ_T is from a binomial distribution.

Example Sarah goes every night of the week (Monday-Friday) to the bar in Olympiapark. Her probability of kissing a guy in one night ist 0.1. (She only kisses maximum one guy per night)

What is the probability that Sarah kisses exactly two guys in a week?

$$B(2,5,.1) = .0729$$

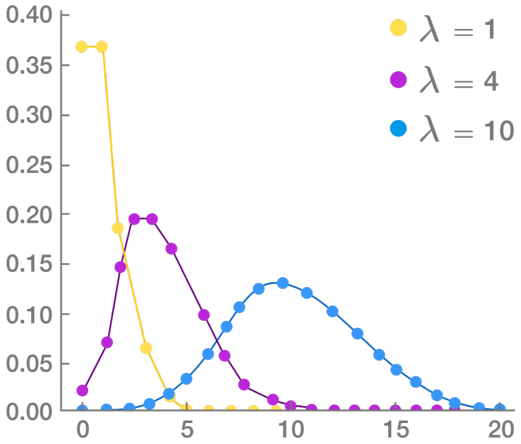
What is the probability that Sarah kisses maximum two guys in a week?

$$P(k <= 2) = B(0,5,.1) + B(1,5,.1) + B(2,5,.1) = .9914$$

Poisson Distribution

Checklist

- Count of discrete events
- Individual events occur at a given **rate** λ and are independent of other events.
- Fixed amount of time or space in which the events can occur
- Outcome: Number of successes (k)
- **Parameters:** λ
- Notation: $X \sim Poisson(\lambda)$



Poisson Distribution

Examples: Number of phone calls received by a call center per hour, number of artificial heart valves failures per year.

Number of Ebola outbreaks in a year? No, since they are dependent.

Example: On a Friday night, the number of people that regularly go to Bahnwärter Thiel is 400.
What is the probability that more than 380 people come?

$$P(X \geq 380 | \lambda = 400) = 0.83$$

Continuous Distributions

The last distributions were all discrete distributions. Now, we look at some continuous distributions.

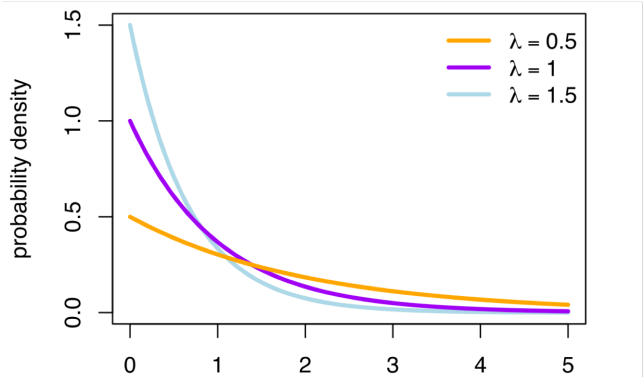
You already know the following:

- Normal Distribution $X \sim \mathcal{N}(\mu, \sigma)$
- t Distribution $X \sim t(df)$
- F Distribution $X \sim \mathcal{F}(df_1, df_2)$
- χ^2 Distribution $X \sim \chi^2(df)$

Exponential Distribution

Checklist

- Continuous, non-negative data
- Used to measure the amount of time or space between events.
- Events occur independently, at a constant rate
- Outcome: Time between events (t)
- **Parameters:** λ
- Notation: $X \sim Exp(\lambda)$



Exponential Distribution

Similar to Poisson distribution, but instead of having the number of events as outcome, here we have time between these events.

Example Your mom calls you very often, you calculate that the rate in which she calls you is every 2,5 hours. (all day all night)

You want to watch Avengers (3 hours), your mom calls you before the movie starts, what is the probability that she calls during the movie?

$$P(t \leq 3 | \lambda = 2.5) = 0.99$$

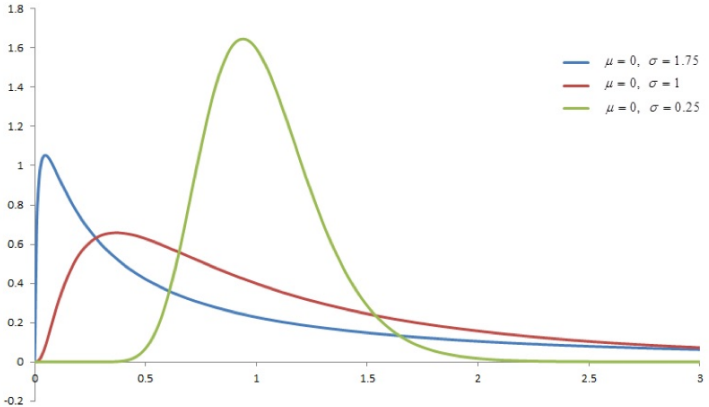
Oh no! She will definitely call you

Log-normal Distribution

Checklist

- Continuous data
- Logarithmic variant of the normal distribution
- Can only represent positive values
- **Parameters:** μ, σ
- Notation: $X \sim Lognormal(\mu, \sigma)$

Important in the description of natural phenomena and human behavior. Example: likes, comments on Facebook.



Fitting the Data

The fitting process needs two ingredients:

- **Finding the best distribution**
- Finding the best estimators for the selected distribution's parameters

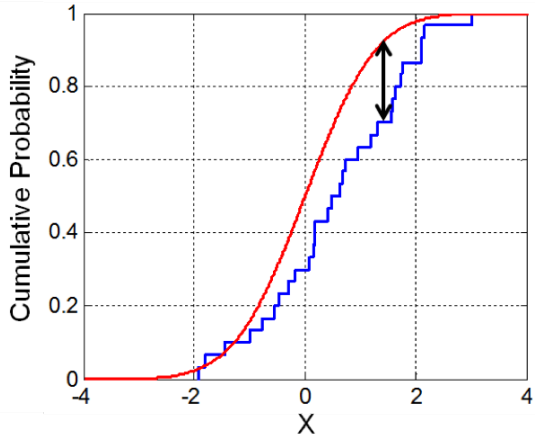
There are two ways to select the best distribution:

Kolmogorov-Smirnov Test

First way to select the best distribution: The **K-S** test

Non-parametric test between the cumulative probability distributions of a reference distribution and the data distribution.

The test statistic is the biggest difference between distributions, D:



The red line is the cumulative distribution with estimated parameters and the blue line the data cumulative distribution divided by the total counts (to make a probability)

Kolmogorov-Smirnov Test

Hypothesis test:

$$H_0 : \textit{The data sample comes from the reference distribution}$$

$$H_1 : \textit{The data sample does not come from the reference distribution}$$

Idea: Compare the data distribution with each of the possible distributions, and find the one with the highest p-value. (We DON'T want to reject the null)

Using the Likelihood

Second way to select the best distribution: Using the likelihood

We selected between estimators to select the best estimator using the maximum likelihood.

Can we use it again to compare distributions?

Yes, but not on its own, because ...

The more parameters a distribution has the easier it is to have a higher likelihood. Unfair for distributions with few parameters!

AIC and BIC

We use a better measure called **AIC**:

$$AIC = -2 * \mathscr{L} + 2 * \#p$$

Which takes the number of parameters ($\#p$) as a penalization

There is a second measure, which also includes the number of datapoints (N) as penalization. The **BIC**:

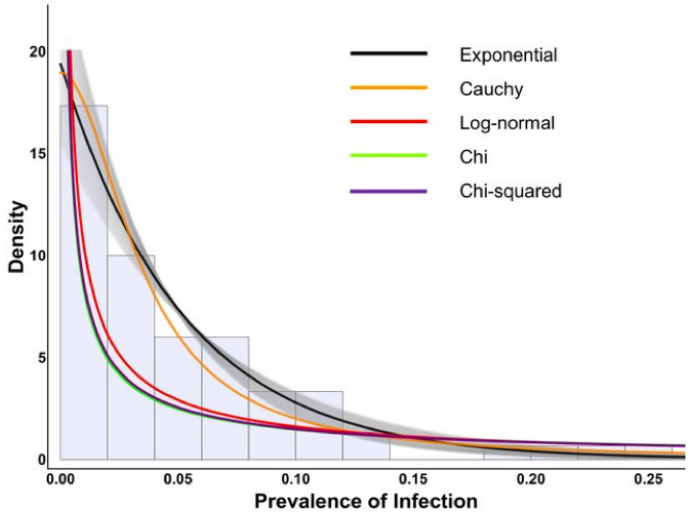
$$BIC = -2 * \mathscr{L} + \#p * log(N)$$

The distribution with lowest AIC/BIC is the best distribution for the data!
Lowest because both AIC and BIC use the negative likelihood

Note The K-S test can only be used for continuous data distributions, whereas AIC and BIC can also be used for discrete data distributions

Comparing Distributions

Real world example: Ocular Chlamydia Prevalence across Tanzanian Communities:



The exponential is the best fit. The gray area correspond to 95% confidence intervals.
Note: Cauchy distribution is a t distribution with one degree of freedom.

Final Words: Statistical Models

When we fit a distribution to our data, we estimate parameters. This is our **statistical model**.

In week 2, we had the simplest model, an estimator of the mean.

Model is something with which you want to represent reality! It can go from something simple (like a point estimate, e.g. the mean) to something more complicated (distributions we learned here) to something very advanced (neural networks).

Literature

- Parametrische Statistik Ch. 3 (Only in German...sorry)