Master's thesis
Astronomy

Your Title Here

Anni Järvenpää

June 2, 2017

Tutor:	Associate Professor Peter Johansson
	Dr. Till Sawala
Censors:	prof. Smith
	doc. Smythe

UNIVERSITY OF HELSINKI
Department of Physics

PL 64 (Gustaf Hällströmin katu 2a)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Physics

Tekijä — Författare — Author

Anni Järvenpää

Työn nimi — Arbetets titel — Title

Your Title Here

Oppiaine — Läroämne — Subject

Astronomy

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	June 2, 2017	0 pages

Tiivistelmä — Referat — Abstract

Abstract goes here.

Avainsanat — Nyckelord — Keywords

Your keywords here

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1 Introduction **1**

1.1 TL;DR version of prerequisite information 1

1.2 History of Local Group Research 1

1.3 Aim of This Thesis 2

2 Theoretical Background **3**

2.1 Local Group 3

2.1.1 Structure 3

2.1.2 Evolution 3

2.2 Expanding universe 4

2.2.1 Discovery 4

2.2.2 ΛCDM Cosmology 4

2.2.3 Hubble flow 4

3 Mathematical and statistical methods **5**

3.1 Statistical Background 5

3.1.1 Hypothesis testing and p-values 6

3.1.2 Distribution functions 7

3.2 Regression Analysis 10

3.3 Error analysis 10

3.4 Goodness of fit testing 10

3.4.1 Comparing two samples drawn from unknown distributions . . 10
3.5 Cluster Analysis 11

4 general simulation thingies 12
4.1 N-body simulations 12
4.1.1 Hierarchical Tree Algorithm 12
4.1.2 Numerical Integrators 12
4.1.3 Halo Finding with Subfind 12
4.2 Description of actual simulations used 12

5 Findings from DMO Halo Catalogue Analysis 13
5.1 Selection of Local Group analogues 13
5.2 Local Anisotropy of the Hubble Flow 13
5.2.1 Hubble Flow Fitting 13
5.3 Statistical Estimate of the Local Group Mass 13

6 Conclusions 14

Bibliography 15

1. Introduction

1.1 TL;DR version of prerequisite information

1. galaxies form
   - Why?
   - When?
   - How?
   - Where?
2. galaxies form in groups
3. our local group is one of these
4. something about large scale distribution of galaxies

1.2 History of Local Group Research

LG objects visible with naked eye -> realization they are something outside our galaxy -> realization they are something very much like our galaxy

First determining distance was difficult, now mass is more interesting question

1.3 Aim of This Thesis

Whatever the main results end up being, presented in somewhat coherent manner and hopefully sugar-coated enough to sound Important and Exciting.

2. Theoretical Background

Think whether LG or LCDM first

2.1 Local Group

Definition of galaxy group, our local group is one of these.

Mass estimate (Li, Yang masses for the LG and MW)

Maybe something about scale of things in our universe, what are galaxy groups made of, what do you get if you go one distance scale up, what's different in galaxy clusters

2.1.1 Structure

Galaxies that are part of LG, distribution of smaller ones around bigger ones

Current mass estimates (at least timing argument, hubble flow and maybe satellites)

2.1.2 Evolution

How have we ended up in a situation described earlier? What will happen in future?

2.2 Expanding universe

2.2.1 Discovery

Make maths, add cosmological constant, make observations, remove cosmological constant

Enough cosmology here or in other sections to make other parts of thesis to make sense and to suffice as master's thesis = basic textbook cosmology and galaxy formation theory

2.2.2 ΛCDM Cosmology

2.2.3 Hubble flow

What is, where seen, what means, how to measure, hotness/coldness

Plot: observations with fitted hubble flow

3. Mathematical and statistical methods

Precision of the used equipment limits accuracy of all data gathered from physical experiments, simulations or observations. Therefore the results are affected by the measurement process and the results have to be presented as estimates with some error, magnitude of which is affected by both number of data points and accuracy of the measurement equipment. [Bohm and Zech, 2010]

Estimating errors for measured quantities offers a way to test hypotheses and compare different experiments[Bohm and Zech, 2010]. This is done using different statistical methods, a few of which are covered here. Methods used in this work are shortly introduced in the following sections together with basic statistical concepts that are necessary to understand the methods.

3.1 Statistical Background

täällä tarvittavat esitiedot ja önnönnöö, listaa mm. mitä aiot kertoa kunhan tiedät itsekään

3.1.1 Hypothesis testing and p-values

A common situation in scientific research is that one has to compare a sample to either a model or another sample in order to derive a conclusion from the dataset. In statistics, this is known as hypothesis testing. For example, this can mean testing hypotheses like ”these two variables are not correlated” or ”this sample is from a population with a mean of 1.0”. [J. V. Wall, 2003] Next paragraphs shortly introduce the basic concept of hypothesis testing and methods that can used to test the hypothesis ”these two samples are drawn from the same distribution”.

Typically the process of hypothesis testing begins with forming of null hypothesis H_0 that is formatted such that the aim for the next steps is to either reject it or deduce that it cannot be rejected with a chosen significance level. Negation of the null hypothesis is often called research hypothesis or alternative hypothesis and denoted as H_1. For example, this can lead to H_0 ”this dataset is sampled from a normal distribution” and H_1 ”this dataset is not sampled from a normal distribution”. Choosing the hypotheses in this manner is done because often the research hypothesis is difficult to define otherwise. [Bohm and Zech, 2010; J. V. Wall, 2003]

After setting the hypotheses one must choose an appropriate test statistic. Ideally this is chosen such that the difference between cases H_0 and H_1 is as large as possible. Then one must choose the significance level α which corresponds to the probability of rejecting H_0 in the case where H_0 actually is true. This fixes the critical region i.e. the values of test statistic that lead to the rejection of the H_0. [Bohm and Zech, 2010; J. V. Wall, 2003]

It is crucial not to look at the test results before choosing α in order to avoid intentional or unintentional fiddling with the data or changing the criterion of acceptance or rejectance to give desired results. Only after these steps should the test statistic be calculated. If the test statistic falls within the critical region, H_0 should be rejected and otherwise stated that H_0 cannot be rejected at this significance level.

[Bohm and Zech, 2010; J. V. Wall, 2003]

This kind of probability based decision making is always prone to error. It is easy to see that α corresponds to the chance of H_0 being rejected when it is true. This is known as error of the first kind. However, this is not the only kind of error possible. It might also occur that H_0 is false but it does not get rejected, which is known as error of the second kind. [Bohm and Zech, 2010]

Despite statistical tests having a binary outcome "H_0 rejected" or "H_0 not rejected", a continuous output is often desired. This is what p-values are used for. The name p-value hints towards probability, but despite it's name p-value is not equal to the probability that the null hypothesis is true. These p-values are functions of test statistic and the p-value for a certain value t_{obs} of test statistic gives the probability that under the condition that H_0 is true, the value of a test statistics for a randomly drawn sample is at least as extreme as t_{obs}. Therefore if p-value is smaller than α, H_0 is to be rejected. [Bohm and Zech, 2010]

3.1.2 Distribution functions

[insert lyhyt aloituskappale here]

As the name suggests, PDF is a function the value of which at some point x represents the likelihood that the value of the random variable would equal x. This is often denoted $f(x)$. Naturally for continuous functions the probability of drawing any single value from the distribution is zero, so these values should be interpreted as depicting relative likelihood of different values. For example if $f(a) = 0.3$ and $f(b) = 0.6$ we can say that drawing value a is twice as likely as drawing value b. [Heino et al., 2012]

Another way to use the PDF is to integrate it over semi closed interval from negative infinity to some value a to obtain CDF, often denoted with $F(x)$:

$$F(x) = \int_{-\infty}^{x} f(x')\, dx'. \tag{3.1}$$

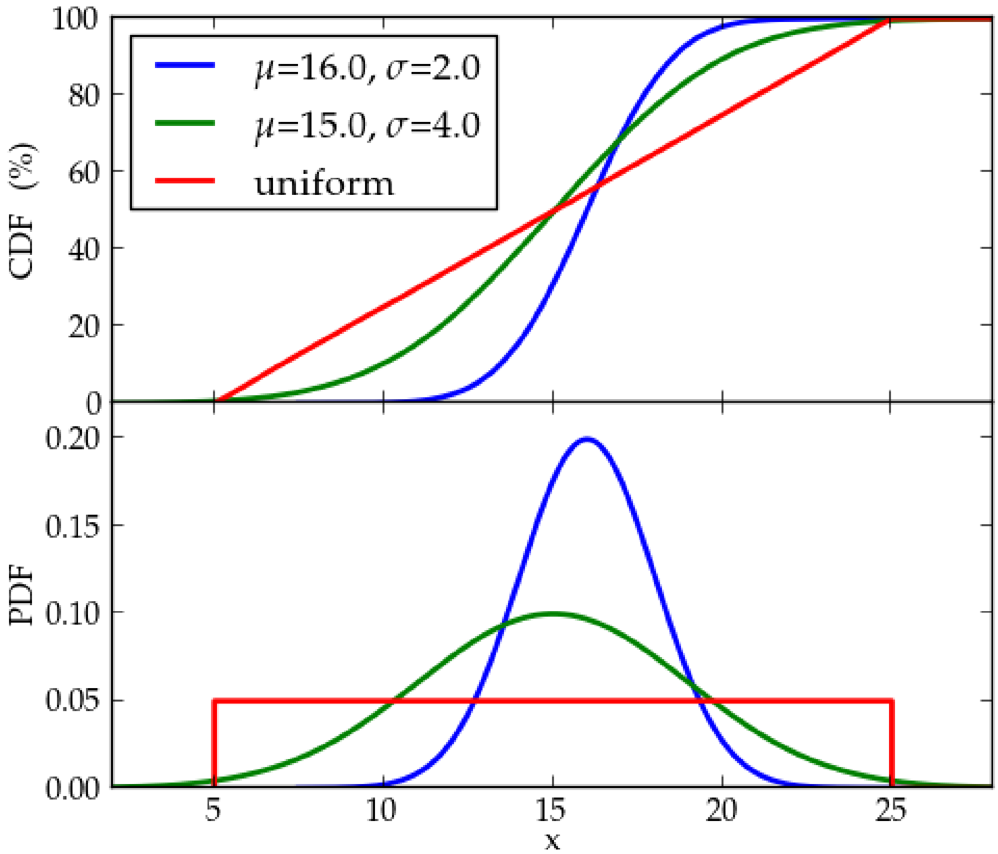

Figure 3.1: Cumulative distribution function (top panel) for three random samples (PDFs shown on bottom panel) drawn from different distributions, two of which are normal and one is uniform.

This gives the probability of a random value drawn from the distribution having value that is smaller than x. Relation between PDF and CDF is illustrated in figure 3.1, where PDFs and CDFs are shown for three different distributions. It is easy to see the integral relation between PDF and CDF and how wider distributions have wider CDFs.[Heino et al., 2012]

Both PDF and CDF both apply to whole population or the set of all possible outcomes of a measurement. In reality the sample is almost always smaller than this. Therefore one cannot measure the actual CDF. Nevertheless, it is possible to calculate a similar measure of how big a fraction of measurements falls under a given

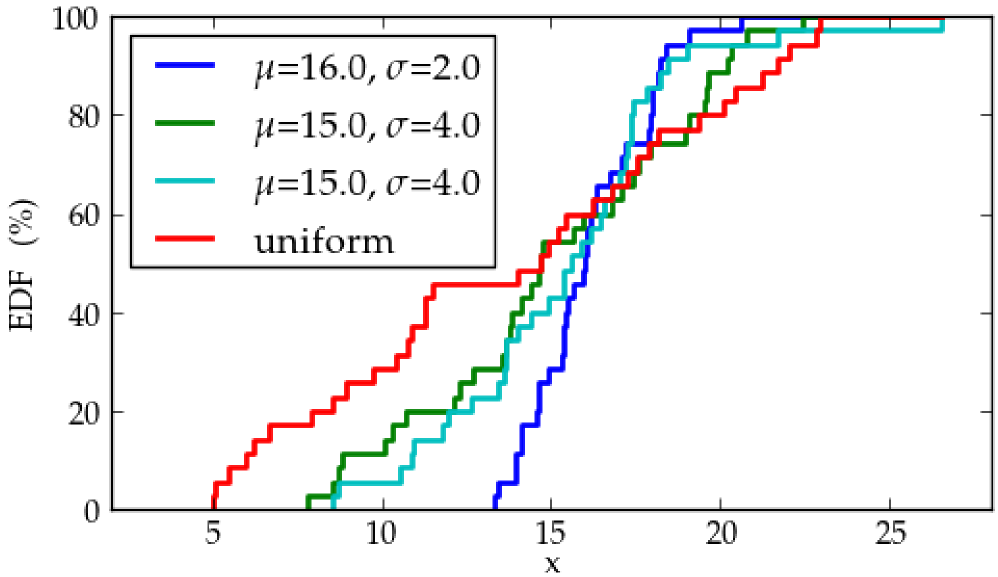

Figure 3.2: Empirical distribution function for four random samples (N=35) drawn from same distributions as in figure 3.1.

value. This empirical counterpart of the CDF is known as empirical distribution function, often denoted $\hat{F}(x)$, and for a dataset $X_1, X_2, ..., X_n$ containing n samples it is defined to be

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \leq x] \tag{3.2}$$

where I is the indicator function, value of which is 1 if the condition in brackets is true, otherwise 0. [Feigelson and Babu, 2012]

Due to EDF being a result of random sampling, it may deviate from the underlying CDF considerably as can be seen by comparing CDFs in figure 3.1 and corresponding EDFs in figure 3.2. Latter figure also has another EDF corresponding to the green curve in the first figure to illustrate the differences that can arise from random sampling. This randomness also makes determining whether two samples are drawn from same distribution difficult.

3.2 Regression Analysis

line fitting and other trivial things

3.3 Error analysis

3.4 Goodness of fit testing

nyt tarvitsen uuden aloituskappaleen

TODO: oispa parempi otsikko

3.4.1 Comparing two samples drawn from unknown distributions

A common question in multiple fields of science is whether two or more samples are drawn from the same distribution. This can occur for example when comparing effectiveness of two procedures, determining if instrument has changed over time or whether observed data is compatible with simulations. There are multiple two-sample tests that can address this kind of questions, e.g. χ^2, Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling tests. [Bohm and Zech, 2010; Feigelson and Babu, 2012]

In addition to comparing two samples, these tests can be used as one-sample tests to determine whether it is expected that the sample is from a particular distribution. However, some restrictions apply when using the one-sample variants. [Feigelson and Babu, 2012]

Out of these,the χ^2 test uses categorical data, for example ”number of data points between values 1.5 and 1.6” or ”number of galaxies that are active”, and compares numbers of samples in different categories whereas the others compare empirical distribution functions (EDF) of the datasets. Feigelson and Babu [2012]

For astronomers the most well-known of these tests is the Kolmogorov-Smirnov test, which is also known as the KS test. Test statistic is calculated based on empirical distribution functions $\hat{F}_1$ and $\hat{F}_2$ derived from two samples and the test statistic

$$D = \sup_x |\hat{F}_1(x) - \hat{F}_2(x)| \tag{3.3}$$

uses the maximum vertical distance of the e.d.f's as shown in [ehdottomasti kuva tähän]. [Bohm and Zech, 2010; Feigelson and Babu, 2012]

3.5 Cluster Analysis

DBSCAN

4. general simulation thingies

Data used here from EAGLE which uses modified GADGET-2 which is a tree-code that uses leapfrog, other integrators also briefly introduced?

4.1 N-body simulations

4.1.1 Hierarchical Tree Algorithm

4.1.2 Numerical Integrators

4.1.3 Halo Finding with Subfind

4.2 Description of actual simulations used

Volume, number of particles, compare to other simulations, where better and where maybe worse

Resimulation of interesting regions

Simulation has same parameters as EAGLE 800 Mpc volume used schaye 2015 paper DM-only parts: Volker-Springer Gadget and Gadget 2 papers 1999 and 2005 or something, gravity part is more interesting than SPH Zooms can use multiple meshes, only one is used here gravitational softening

5. Findings from DMO Halo Catalogue Analysis

5.1 Selection of Local Group analogues

criteria, how many found, what are like (some plots maybe? distributions of masses, separations, velocities or correlations between two of those?). This might be part of previous chapter too (relevant to resimulation)?

5.2 Local Anisotropy of the Hubble Flow

Hopefully there's something at least mildly interesting to report when I get to look at the new data

5.2.1 Hubble Flow Fitting

5.3 Statistical Estimate of the Local Group Mass

Analysis similar to Fattahi et al 2016 paper

6. Conclusions

Bibliography

G. Bohm and G. Zech. *Introduction to statistics and data analysis for physicists*. DESY, 2010. ISBN 9783935702416. URL `http://www-library.desy.de/preparch/books/vstatmp_engl.pdf`.

Eric D. Feigelson and G. Jogesh Babu. *Modern Statistical Methods for Astronomy: With R Applications*. Cambridge University Press, 2012. doi: 10.1017/CBO9781139015653.

R. Heino, K. Ruosteenoja, and J. Räisänen. *Havaintojen tilastollinen käsittely*. Department of Physics (University of Helsinki), 2012.

C. R. Jenkins J. V. Wall. *Practical Statistics for Astronomers*. Cambridge Observing Handbooks for Research Astronomers. Cambridge University Press, illustrated edition edition, 2003. ISBN 9780521454162,0521454166.