THE IMPORTANCE OF NON-PARAMETRIC STATISTICS IN GALAXIES

Diplas Vasileios

2020

ABSTRACT

The dynamical state of galaxy groups at intermediate redshifts can provide information about the growth of structure in the universe. We examine three goodness-of-fit tests, the Anderson–Darling (A–D), Kolmogorov, and χ^2 tests, in order to determine which statistical tool is best able to distinguish between groups that are relaxed and those that are dynamically complex. We perform Monte Carlo simulations of these three tests and show that the χ^2 test is profoundly unreliable for groups with fewer than 30 members. Power studies of the Kolmogorov and A–D tests are conducted to test their robustness for various sample sizes, where we use the Skew Normal Distribution. We then apply these non-parametric tests to a sample from the 2MASS catalog. We applied these well-known normality tests and more tests, such as Lilliefors, D'Agostino, Shapiro Wilk, Cramer-von Misses and Wilcoxon to the velocity and magnitude distributions of these systems to distinguish Gaussian and non-Gaussian clusters. Also we study the Rayleigh distribution. We analyzed galaxy samples from the R package "MASS", the astrostatistics dataset (NASA astrophysics data system (A.M. Garcia 1992), COMBO-17), where we show important histograms for the normality of the group galaxies and uniformity (Rayleigh and Wilcoxon tests) for the morphological type of galaxies (Elliptical, Spiral, Lenticular). Furthermore we use a sample from the Canadian Network for Observational Cosmology (CNOC 2) Field Galaxy Redshift Survey, where we approve some figures from the papers Hou.et al. Finally we compute velocity dispersion profiles(VDPs) for all groups of the CNOC 2 sample with more than 20 members and compare the overall features of the Gaussian and non-Gaussian groups, finding that the VDPs of the non-Gaussian groups are distinct and dynamically more complex from those classified as Gaussian. In the end, we compare the results of the velocities, the masses and the magnitudes with Gaussian and Non-Gaussian galaxies. Our achievement is to apply non-parametric statistical tools for group galaxies we the aim of distinguishing Normal and Non-Normal Groups. We start our project by bringing up some auxiliary tools (physics and statistics).

(This project is only for the purpose to share our results)

1. STATISTICAL TOOLS

1.1. Goodness -of- Fits Tests

1.1.1. Pearson's χ^2 Test

For a random sample $N = (N_1, N_2, \dots, N_k) \sim M_d(n, p)$ follows a polynomial distribution with parameters $n, p = (p_1, p_2, .., p_k)$, where $\sum p_i = 1$, with probability mass function :

$$P(N_1 = n_1, N_2 = n_2, \dots, N_k = n_k,) = \binom{n}{N_1, N_2, \dots, N_k} {p_1}^{N_1} {p_1}^{N_1} \dots {p_k}^{N_k} 1_{\{\sum_{i=1}^{k} N_i = n\}}$$

Then the statistical function:

$$T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \approx \chi^2_{(k-1)}$$

for $n \to \infty$

We have a random sample $X_1, X_2, \dots, X_n$ and we want to test the hypothesis $H_0 : X_i \sim F_o$.

Suppose that *n* observations in a random sample from a population are classified into *k* mutually exclusive classes with respective observed numbers x_i (for $i = 1,2,\dots,k$), and a null hypothesis gives the probability p_i that an observation falls into the *i*th class. So we have the expected numbers $m_i = np_i$ for all *i*.

Suppose $A_1, A_2, \dots, A_k$ are the classes and $p_i = P(X_i \in A_i | H_0 : X_i \sim F_o)$

So, under the H_0 , the statistical function is :

$$T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \approx \chi^2_{(k-1)}$$

Under the $H_1 :$ $H_0 : X_i \sim G \neq F_o$ the statistical function will get higher values and the every N_i will differ from the np_i values.

$$\chi^2 = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{(\mathrm{observed_i} - \mathrm{expected_i})^2}{\mathrm{expected_i}} \ . \qquad (1)$$

1.1.2. The Kolmogorov – Smirnov Test

Definition of Almost Sure Convergence

To say that the sequence $\mathbf{X_n}$ converges almost surely or almost everywhere or with probability 1 or strongly towards X means that

$$\mathrm{P}\left(\lim_{\mathrm{n}\to\infty} \mathrm{X_n} = \mathrm{X}\right) = 1$$

This means that the values of $\mathbf{X_n}$ approach the value of X, in the sense that events for which $\mathbf{X_n}$ does not converge to X have probability 0. Using the probability space and the concept of the random variable as a function from Ω to **R**, this is equivalent to the statement

$$P\left(\omega \in \Omega: \lim_{n\to\infty} X_n(\omega) = X(\omega)\right) = 1$$

Using the notion of the limit superior of a sequence of sets, almost sure convergence can also be defined as follows:

$$P(limsup\{\omega \in \Omega: |X_n(\omega) - X(\omega)| > \varepsilon\}) = 0 \ \ \forall\, \varepsilon > 0$$

Almost sure convergence is often denoted by adding the letters *a.s.* over an arrow indicating convergence:

$$X_n \xrightarrow{a.s.} X$$

For $X_n = F_n$ we suppose :

$$||F_n - F\,||_\infty \xrightarrow{a.s.} 0 \quad (\, n \to \infty\,)$$

$\forall\, \varepsilon > 0, n \geq 1$, F cumulative probability function

$$P(sup\{\omega \in \Omega: |F(\omega) - F(\omega)| > \varepsilon\}) \leq 2e^{-2n\varepsilon^2}$$

$$\sum P(||F_n - F\,||_\infty \to 0) \leq 2\sum_{n=1}^{\infty} e^{-2n\varepsilon^2} < \infty$$

So :

$$P(||F_n - F\,||_\infty \to 0) = 1$$

We assume that $D_n = ||F_n - F\,||_\infty$ is the Kolmogorov-Smirnov statistic. It the distance between the empirical distribution function F_n and the theoretical distribution function F, where we count the distance using the Uniform Distribution $\mathbf{U[0,1].}$

For:

$$\varepsilon = \frac{x}{\sqrt{n}}\ , x > 0$$

$$P\left(\sqrt{n}D_n\right) \leq 2e^{-2nx^2}\ \forall\, x > 0$$

So we have a supremum of the distribution's tail of the random variable $\sqrt{n}D_n\ , \forall\ n \geq 1$.

Suppose F_n will be the e.d.f. of the random sample $X_1, X_2, \dots, X_n$ and U_n the e.d.f. of a random sample from Uniform(0,1) distribution. Then :

$$F_n = U_n\, o\, F \Rightarrow$$

$$F_n\,(x) = \frac{1}{n}\sum_{k=1}^{n} 1_{\{(-\infty, F(x)]\}}(U_k)$$

$$F_n\,(x) = \frac{1}{n}\sum_{k=1}^{n} 1_{\{(-\infty, x]\}}(x_k)\,, x \in \mathbb{R}$$

$$\forall\, k \geq 1\, \exists\, U_k \sim \text{Uniform}(0,1)\ :\ x_k = F^{-1}(U_k)$$

Suppose $H_O : F = F_o$ where F_o is a known distribution function . We set the following equations :

$$D_n{}^{+} = \sup\left(F_n\,(x) - F_O(x)\right), x \in \mathbb{R}$$

$$D_n{}^{-} = \sup\left(F_O(x) - F_n\,(x)\right), x \in \mathbb{R}$$

$$D_n = \max\{\, D_n{}^{+}, D_n{}^{-}\}, \qquad (3)$$

If $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ are ordered observations , where

$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ of the random sample $X_1, X_2, \dots, X_n$ and

$$U_i = F_O(x_i)\ \forall\ \ i = 1,2, \dots, n$$

Then :

$$D_n{}^{+} = supremum\left|\frac{i}{n} - U_{(i)}\right|, \quad (4)$$

$$D_n{}^{-} = supremum\left|U_{(i)} - \frac{i-1}{n}\right|, \quad (5)$$

or

$$D_n{}^{+} = \max_{1\le i\le n}\{\frac{i}{n} - U_{(i)}\}$$

$$D_n{}^{-} = \max_{1\le i\le n}\{U_{(i)} - \frac{i-1}{n}\}$$

The $U_{(i)}$ are the ordered observations of the $U_i \sim \text{Uniform}(0,1)$

The empirical distribution function is :

$$\hat{F}_n(x) = \begin{cases} 0\,, & x \le X_{(1)} \\ \frac{1}{n}\,, X_{(1)} \le x < X_{(2)} \\ \frac{2}{n}\,, X_{(2)} \le x < X_{(3)} \\ \vdots \\ 1\,, & X_{(n)} \le x \end{cases}$$

The goodness-of-fit test or the Kolmogorov–Smirnov test can be constructed by using the critical values of the Kolmogorov distribution. This test is asymptotically valid when $n \to \infty$. It rejects the null hypothesis at level a if

$$D_n > c = D_n(a)$$

The $D_n(a)$ is a critical value of the D_n distribution that does not depend on F_O (distribution free). The exact distribution of D_n is hard to be computed. For that reason we compute the distribution of $\sqrt{n}D_n$. We set $Z_n = \sqrt{n}D_n$

$$P(Z_n \le z) = P\left(\sqrt{n}D_n \le z\right) \to 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2z^2} \quad ,\forall\ z\ \ge 0$$

We compute the p-value asymptotically of a random sample , which gave us $D_n = d$

$$p_{value} = P(D_n > d|H_O) = 1 - P\left(\sqrt{n}D_n < \sqrt{n}d\right) \approx\ 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2nz^2}$$

So $$p_{value} \approx 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2nz^2}$$

with critical value :

$$D_n{}^{critical} > (-\frac{1}{2}\ln(\frac{a}{2}))^{1/2} = D_n(a)$$

Stephens (1974) has simplified the Kolmogorov test with the modification of the D values, called the D^* value(Equation (6)), which allows for comparison with one critical value table, rather than computing critical values for specific sample sizes and significance levels (Massey1951).

$$D^* = D(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}})$$

The use of the Kolmogorov test in place of the χ^2 test for small samples is suggested by Lilliefors (1967). In a comparison of the Kolmogorov and χ2tests, Massey (1951) concludes that the Kolmogorov test is generally more reliable than theχ^2 test, especially for small n, where the effects of binning, required by the χ^2 test but not the Kolmogorov test, can result in a large loss of information.

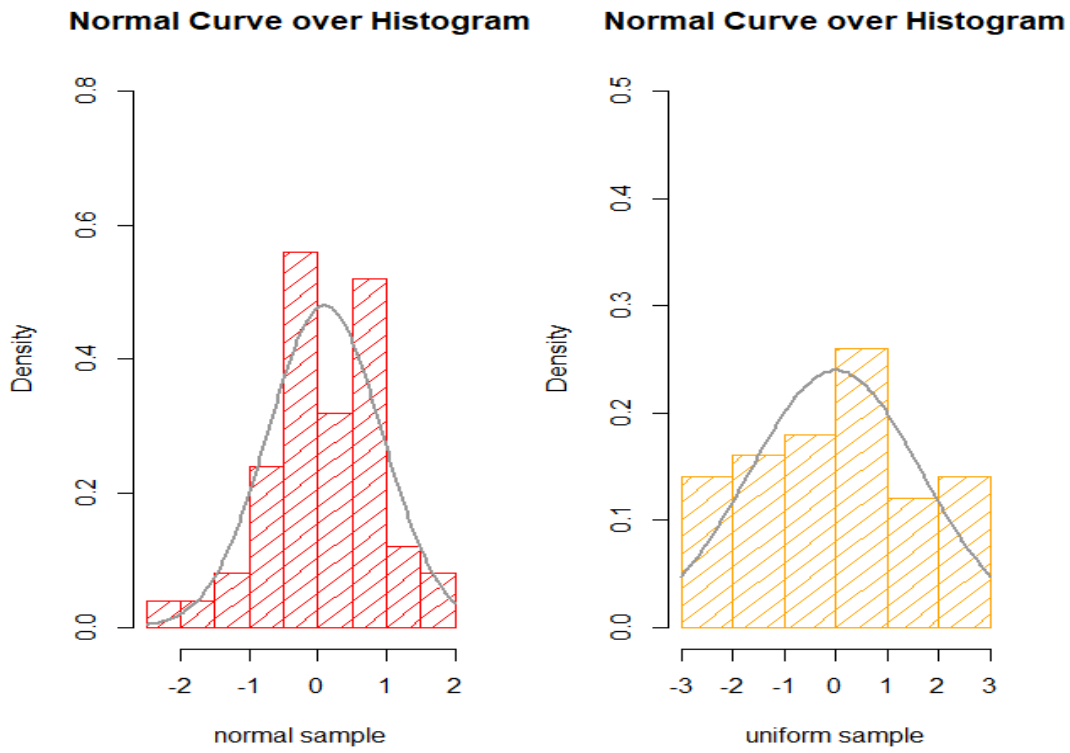


Figure 10

In the above histogram we generate two random samples (normal $N(0,1)$ and uniform $Unif[-3,3]$) and we want to see how the empirical values are close or not to the normal curve.

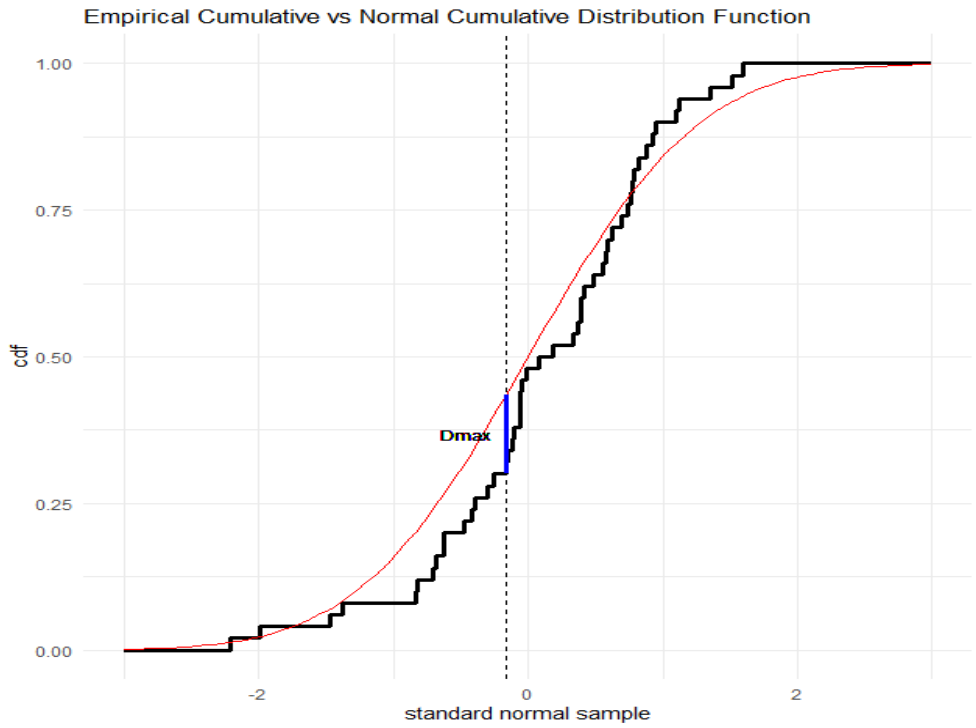


Figure 11

We represent the Kolmogorov-Smirnov test in the above graph for the random sample comes from a normal $N(0,1)$ distribution. The theoretical distribution is a normal $N(0,1)$. We will do the same for the sample comes from a Uniform distribution $Unif[0,1]$

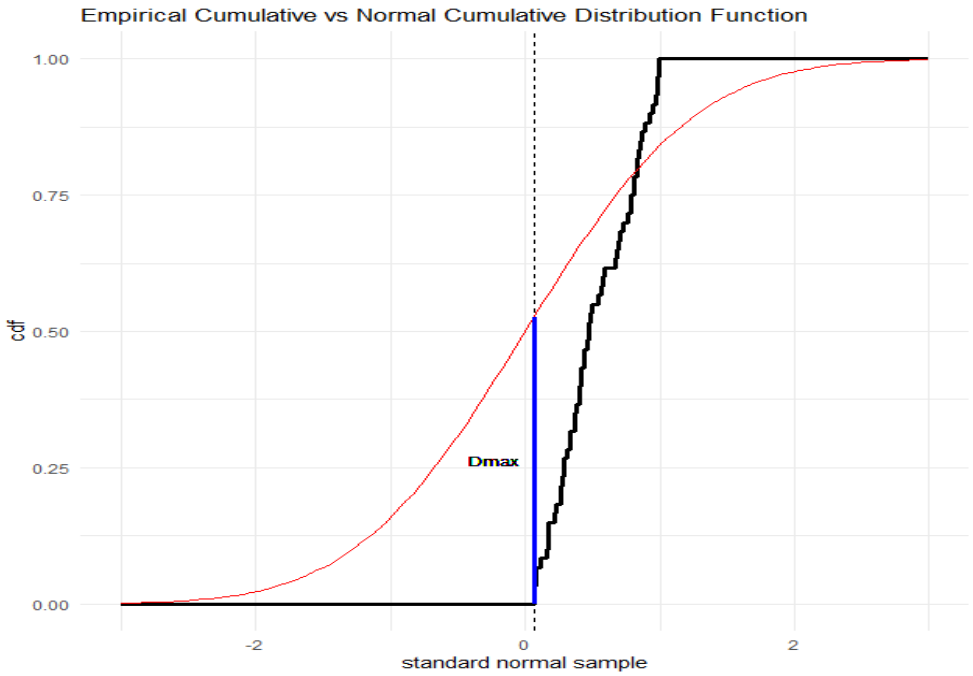


Figure 12

1.1.3. The Anderson – Darling test

Like the Kolmogorov test, the A–D test is also based on EDF statistics and does not require binning or graphical analysis. Despite these advantages, the A–D test is not commonly used in astronomy. The A–D statistic involves the calculation of theA2andA2∗values, starting from the ordered data {xi}:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(ln\Phi(x_i) + \ln\left(1-\Phi(x_{n+1-i})\right)\right), \quad (7)$$

$$A^{2*} = A^2\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right), \quad (8)$$

where xi < x< xi+1 and Φ(xi) are the CDF of the hypothetical underlying distribution. The use of either A^2 or $A^{2*}$for the A–D statistics depends on how well the distribution parameters are known a priori. The A^2 values are modifications for cases where the distribution parameters are not known a priori and must be estimated from the xi values. If the input parameters are known beforehand, then the A^2 values should be used for comparison with critical value tables. From Equation (8), it is clear that A^{2*} approaches A^2 for large n.

In this analysis we take Φ(xi) to be the CDF of a Gaussian distribution , given as :

$$\Phi(x_i) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x_i-\mu}{\sqrt{2}\sigma}\right)\right), \quad (9)$$

where xi is the radial velocity of the galaxy group members, arranged from lowest to highest,μis the computed mean velocity of the group, andσis the computed velocity dispersion. The Φ(xi) values are then used in the A–D computing formulas (Equations (7) and (8)) to obtain the A^{2*} values.

Error Function:

In Mathematics, the error function, often denoted by erf, is defined as:

$$\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{+x} e^{-t^2} dt = \frac{2}{\sqrt{\pi}} \int_{0}^{+x} e^{-t^2} dt$$

In statistics, for non-negative values of x, the error function has the following interpretation: for a random variable Y that is normally distributed with zero (0) mean and variance ½ , $erfx$ is the probability that Y falls in the range $[-x, +x]$. The error function is essentially identical to the standard normal cumulative distribution function, denoted Φ, also named norm(*x*) by software languages, as they differ only by scaling and translation.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+x} e^{\frac{-t^2}{2}} dt = \frac{1}{2}[1 + \operatorname{erf}(\frac{x}{\sqrt{2}})]$$

1.1.4. The Lilliefors test

The Lilliefors tesy is an EDF omnibus test for thr composite hypothesis of normality. The tset statistic is the maximal absolute difference between empirical and hypothetical cumulative distribution function. It may be computed as $D_n = \max\{ {D_n}^+, {D_n}^- \}$ with

$$ {D_n}^+ = \max_{1 \le i \le n} \{ \frac{i}{n} - p_{(i)} \}$$

$$ {D_n}^- = \max_{1 \le i \le n} \{ p_{(i)} - \frac{i-1}{n} \}$$

, where $p_{(i)} = \Phi([x_{(i)} - \bar{x}]/s)$. Here Φ is the cumulative distribution function of the standard normal distribution and $\bar{x}$ and s are the mean and the standard deviation of the data values. The p-value is computed from the Dallal-Wilkinson (1986) formula, which is claimed to be only reliable when the p-value is smaller than 0.1. If the Dallal-Wilkinson p-value turns out to be greater than 0.1 , then the p-value is computed from the distribution of the modified statistic

$$ Z = D(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}})$$

1.1.5. The Shapiro – Wilk test

The **Shapiro–Wilk test** is a test of normality in frequentist statistics. It was published in 1965 by Samuel Sanford Shapiro and Martin Wilk.

The Shapiro-Wilk tests the null hypothesis that a sample $X_1, X_2, \dots, X_n$ came from a normally distributed population. The test statistic is :

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where

- $x_{(i)}$ not to be confused with x_i is the ith order statistic, i.e., the ith-smallest number in the sample;
- $\bar{x}$ is the sample mean.

The coefficients are given by :

$$(a_1, a_2, \dots, a_n) = \frac{m^T V^{-1}}{C}$$

, where C is a vector norm:

$$C = \left|\left|V^{-1}m\right|\right| = (m^T V^{-1} V^{-1} m)^2$$

and the vector m,

$$m = (m_1, m_2, \dots, m_n)^T$$

is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution; finally, V is the covariance matrix of those normal order statistics. There is no name for the distribution of W. The cutoff values for the statistics are calculated through Monte-Carlo simulations.

1.2 Monte Carlo Simulations

Firstly, we will represent graphically the Monte Carlo Simulations of the χ^2 test for different sample sizes, using Gaussian random number generator with input values of μ = 0 and σ_{int} = 0 .

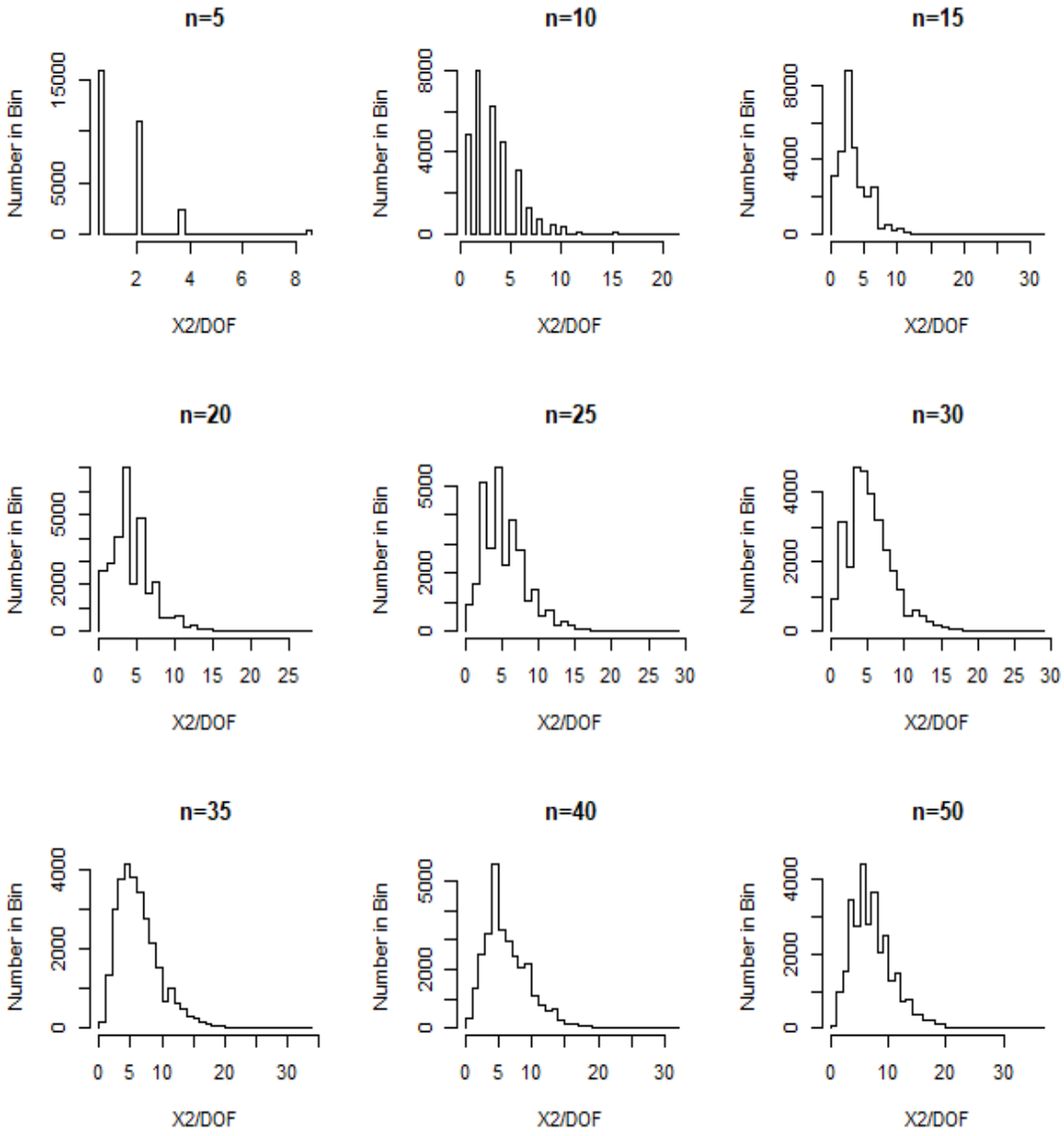


Figure 13 **The above histograms are generated with a bin width of 0.5. For $n < 30$ the χ^2 test completely fails to recover the input distribution. This is the most obvious for the n=5, where the histogram shows three peaks at χ^2/dof=1, 2, 4, instead of the expected peak value of $\chi^2/dof = 1$. Only for $n > 30$ do the simulations recover the expected peak value.**

Secondly, we will represent graphically the Monte Carlo Simulations of the Kolmogorov – Smirnov test for different sample sizes, using Gaussian random number generator with input values of $\mu = 0$ and $\sigma_{int} = 0$.

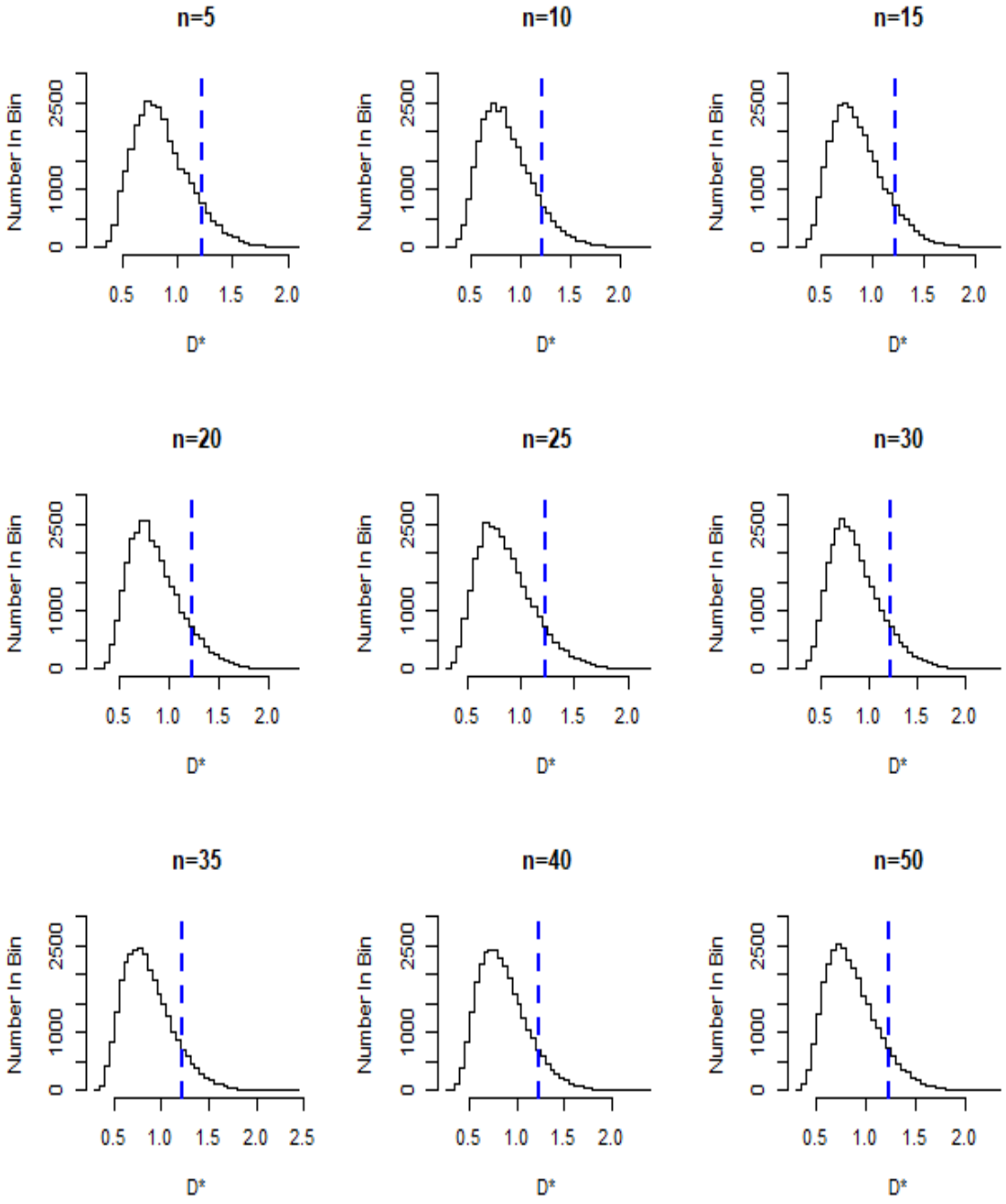


Figure 14 The histogram of D^* values are plotting using a bin width = 0.05. The results of the simulations show that even at n=5 the Kolmogorov-Smirnov test is able to recover the input distribution. The histograms for n=5 to 50, consistently reproduce the expected peak values, indicating that the test is reliable over a wide sample range. The vertical dotted line represents the D^* value above which 10% of the value s lie, our computed values are in complete agreement with the known a = 0.10 critical value of 1.224 (DA86)

Thirdly, we will represent graphically the Monte Carlo Simulations of the Anderson – Darling test for different sample sizes, using Gaussian random number generator with input values of $\mu = 0$ and $\sigma_{int} = 0$.

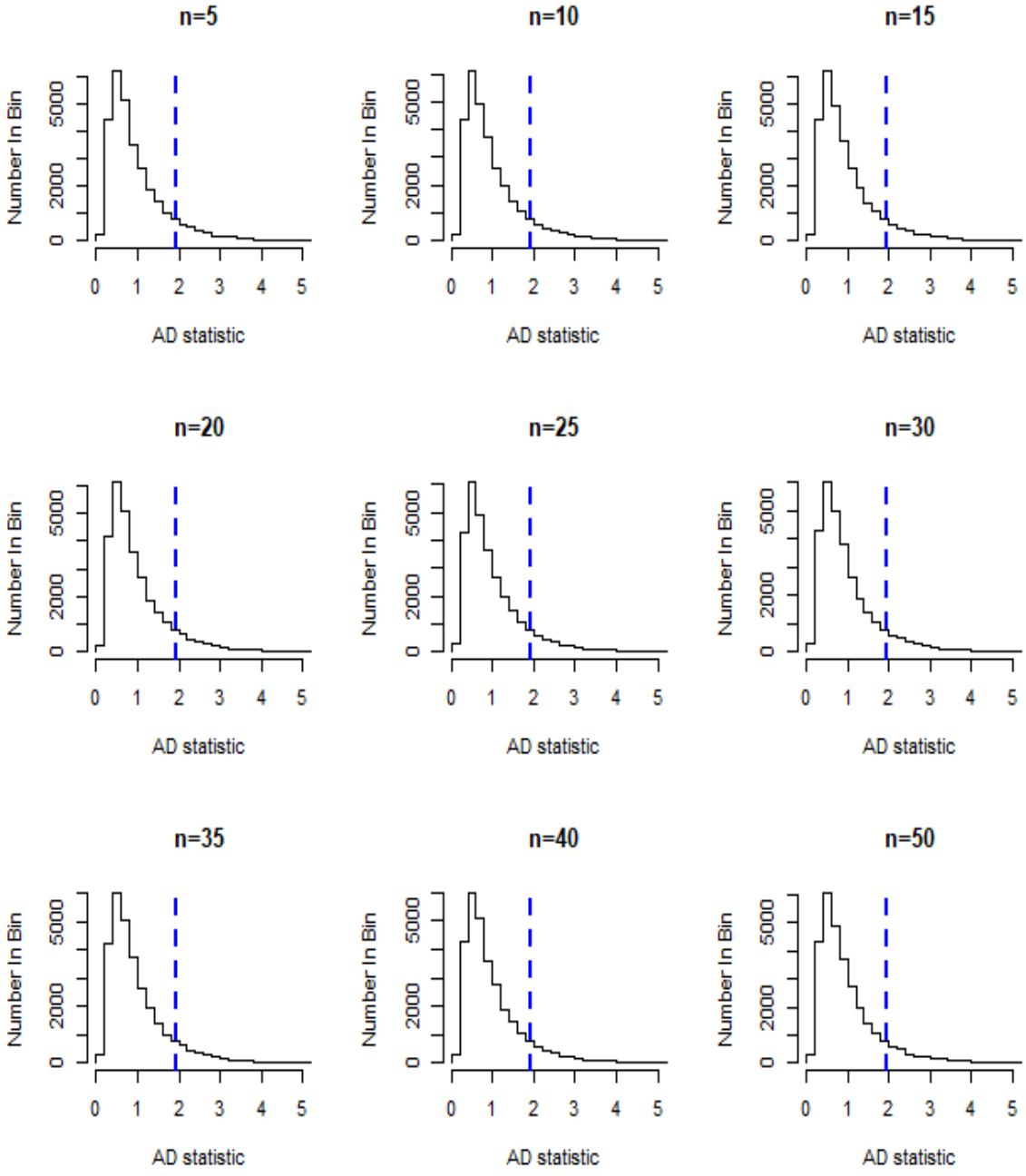


Figure 15 The histogram of A^2 values are plotting using a bin width = 0.2. The histograms for n=5 to 50, consistently reproduce the expected peak values, indicating that the test is reliable over a wide sample range. The vertical dotted line represents the D^* value above which 10% of the values lie, our computed values are in complete agreement with the known a = 0.10 critical value of 1.933 (DA86)

1.3.1.1 Skew Normal Distribution

Consider first a continuous random variable X having pdf of the following form:

$$f(x_i) = 2\varphi(x_i)\Phi(ax_i) \quad , (10)$$

,where α is a fixed arbitrary number and

$$\varphi(x_i) = \frac{e^{\left(-\frac{x_i^2}{2}\right)}}{\sqrt{2\pi}} \quad \rightarrow pdf\ of\ Skew\ Normal \quad (11)$$

$$\Phi(\alpha x) = \int_{-\infty}^{ax} \varphi(t)\,dt$$

The computed α is called the shape parameter because it generates the shape of the density function. Especially:

1) When $\alpha = 0$, the skewness vanishes, and we obtain the Standard Normal Distribution
2) As a increases(in absolute value) , the skewness of the distribution increases
3) When $a \rightarrow \infty$, the density converges to the folded Normal density function
4) If the sign of a changes, the pdf is reflected on the opposite side of the vertical axis

We represent a method to Simulate the Skew Normal Distribution. We denote by $SN(\theta)$ the skew normal distribution of parameter θ and density

$$f(x) = 2\varphi(x_i)\Phi(\theta x_i)$$

, where φ and Φ denote the standard normal $N(0,1)$ probability density function and cumulative distribution function, respectively.

The Skew Normal distribution, due to its mathematical tractability and inclusion of the Standard Normal distribution has attracted a lot of attention in the literature. So if U_1 , U_2 are identically and independently distributed $N(0,1)$ random variables, then :

$$\frac{\theta|U_1| + U_2}{\sqrt{1 + \theta^2}}\ has\ the\ Skew\ Normal\ Distribution$$

For the simulation of the Skew Normal, we propose a combination of maximum and minimum of the independent and identically distributed $N(0,1)$ random variables.

Let U_1 , U_2 two independent and identically distributed $N(0,1)$ random variables and

$U = \max\{U_1, U_2\}$ and $V = \max\{U_1, U_2\}$

For simulation of the random number variable $X \sim SN(\theta)$, we take this combination of U and V

First note that:

- $\theta = 0$,the density becomes $f(x) = \Phi(x) \Rightarrow X \sim N(0,1)$
- $\theta = -1$,the density becomes $f(x) = 2\varphi(x)\Phi(-x) \Rightarrow X = V$

- $\theta = 1$,the density becomes $f(x) = 2\varphi(x)\Phi(x) \Rightarrow X = U$

For $\theta \notin \{0,1,-1\}$, note that:

$$\lambda_1 = \frac{1+\theta}{\sqrt{2(1+\theta^2)}} \ , \lambda_2 = \frac{1-\theta}{\sqrt{2(1+\theta^2)}}$$

$$\lambda_1{}^2 + \lambda_2{}^2 = 1$$

For simulation of the random variable $X \sim SN(\theta)$ we take the combination of $, V$ in the form :

$$X = \lambda_1 U + \lambda_2 V$$

Proposition

The random variable X defined has the skew normal distribution $SN(\theta)$

Proof

The pair (U,V) has density:

$f_{u,v}(u,v) = 2\varphi(u)\varphi(v)11_{\{v \le u\}}(u,v)$, where 11 is an indicator function

Consider the transformation

$$X = \lambda_1 u + \lambda_2 v \ , Y = \lambda_1 u$$

The inverse transform is defined by:

$$u = \frac{y}{\lambda_1} \ , v = \frac{x-y}{\lambda_2}$$

And the corresponding Jacobian is:

$$\Im = \frac{1}{\lambda_1 \lambda_2}$$

X density is defined by:

$$f(x) = \frac{2}{|\lambda_1 \lambda_2|} \int_{\Delta}^{\cdot\cdot} \varphi(\frac{y}{\lambda_1})\varphi(\frac{x-y}{\lambda_2})\, dy$$

, where $\Delta = \{\frac{x-y}{\lambda_2} \le \frac{y}{\lambda_1}\}$ .Taking into account $\lambda_1{}^2 + \lambda_2{}^2 = 1$, we can write:

$$\varphi\left(\frac{y}{\lambda_1}\right)\varphi\left(\frac{x-y}{\lambda_2}\right)=\varphi(x)\varphi(\frac{y-{\lambda_1}^2x}{|\lambda_1\lambda_2|})$$

So,

$$f(x)=\frac{2\varphi(x)}{|\lambda_1\lambda_2|}\int_{\Delta}^{\cdot\cdot}\varphi(\frac{y-{\lambda_1}^2x}{|\lambda_1\lambda_2|})\,dy$$

For the domain, we have the following three cases:

- Case 1 : $\theta\in(-1,0)\cup(0,1)$, we have

$$|\lambda_1\lambda_2|=\frac{1-\theta^2}{2(1+\theta^2)}\ and\ \Delta=\{y\geq\frac{\lambda_1 x}{\lambda_1+\lambda_2}\}$$

- Case 2 : $\theta\in(-\infty,-1)$, we have

$$|\lambda_1\lambda_2|=\frac{\theta^2-1}{2(1+\theta^2)}\ and\ \Delta=\{y\geq\frac{\lambda_1 x}{\lambda_1+\lambda_2}\}$$

- Case 3 : $\theta\in(1,+\infty)$, we have

$$|\lambda_1\lambda_2|=\frac{\theta^2-1}{2(1+\theta^2)}\ and\ \Delta=\{y\geq\frac{\lambda_1 x}{\lambda_1+\lambda_2}\}$$

We get the result

The absolute value of skewness increases as the absolute value of α increases. The distribution is right skewed if $a>0$ and is left skewed if $a<0$. The probability function with location ξ , scale ω and parameter a becomes,

$$f(x)=\frac{2}{\omega}\varphi(\frac{x-\xi}{\omega})\Phi(\alpha(\frac{x-\xi}{\omega}))$$

Note, however, that the skewness (γ_1) of the distribution is limited to the interval $(-1,+1)$.

- Support

$$x\in(-\infty,+\infty)$$

- Parameters

$$\xi\rightarrow location\ (real)$$

$$\omega \rightarrow scale\ (positive, real)$$
$$a \rightarrow shape\ (real)$$

Here we have:

$$Mean = \xi + \omega\delta\sqrt{\frac{2}{\pi}}\ , where\ \delta = \frac{\alpha}{\sqrt{\alpha^2 + 1}}$$

$$Variance = \omega^2\left(1 - \frac{2\delta^2}{\pi}\right)$$

$$Skewness = \gamma_1 = \frac{4-\pi}{2}\frac{\left(\delta\sqrt{\frac{2}{\pi}}\right)^3}{\left(1 - \frac{2\delta^2}{\pi}\right)^{\frac{3}{2}}}$$

$$Kurtosis = \gamma_2 = 2(\pi - 3)\frac{\left(\delta\sqrt{\frac{2}{\pi}}\right)^4}{\left(1 - \frac{2\delta^2}{\pi}\right)^2}$$

We represent some histograms of Skew Normal Distribution

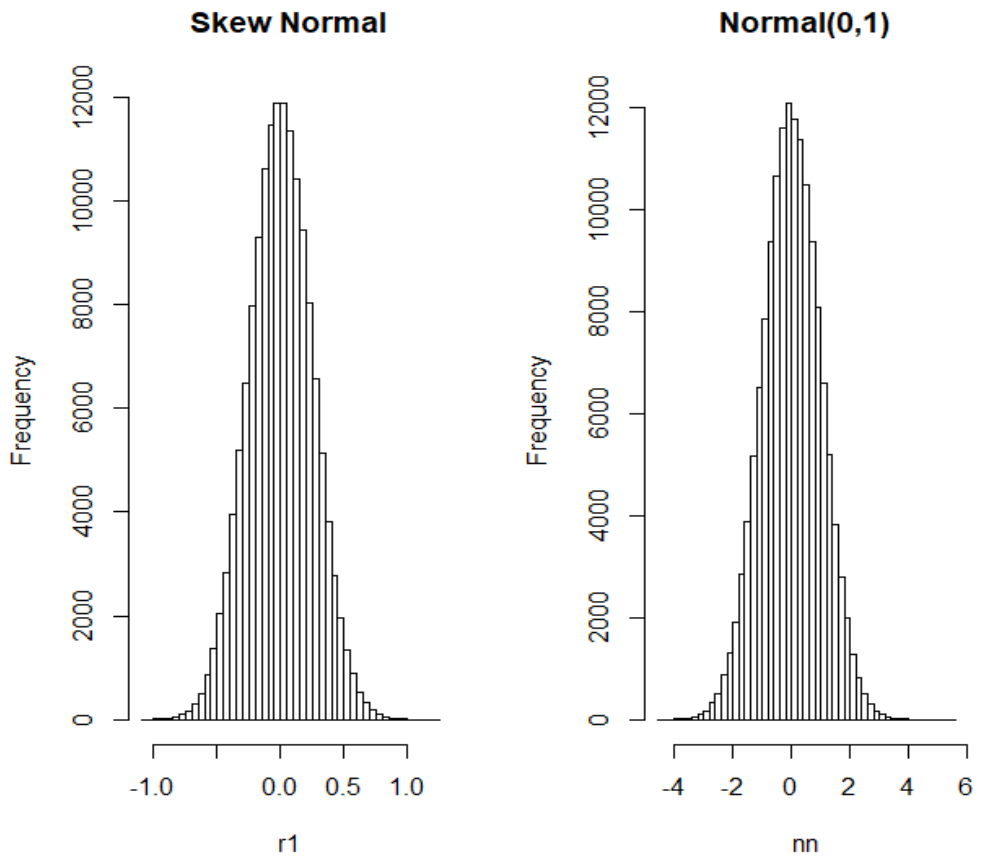


Figure 16 Compare the Skew Normal with ω = ¼ (left) and the Normal(0,1) (right)

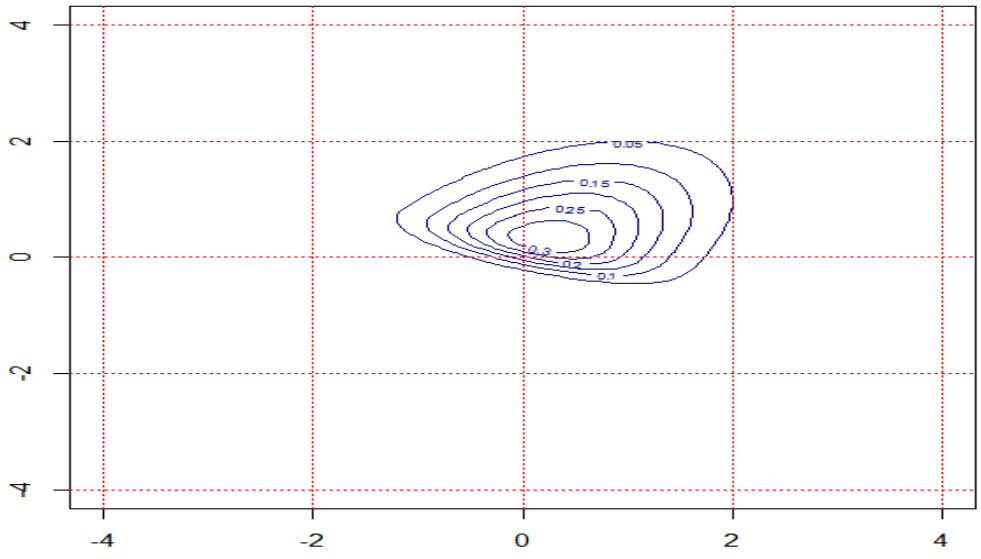


Figure 17 2-D grid pdf of skew normal

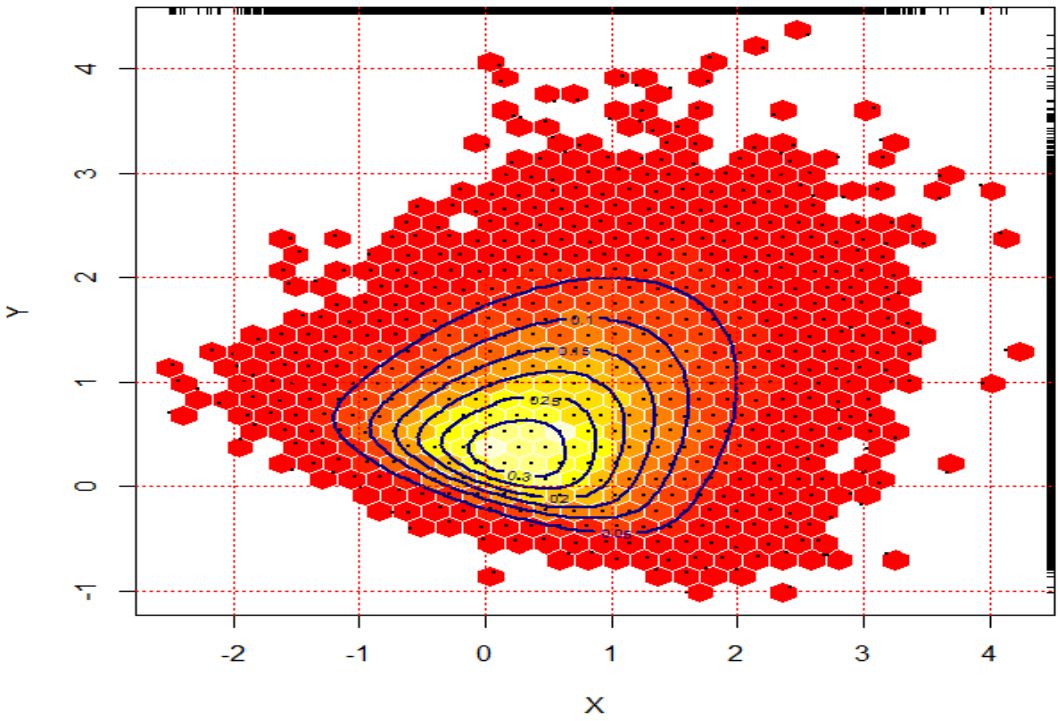


Figure 18 The same contour with colour

We represent the empirical distribution of a skew normal and we compare them with the true distribution function.

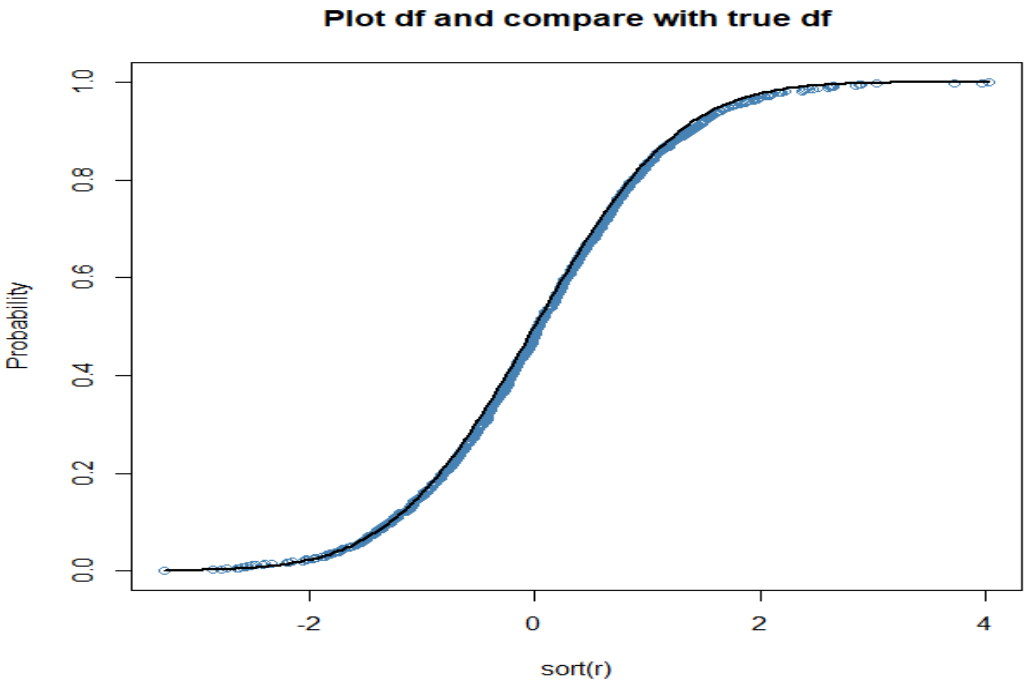


Figure 19 The blue line is the empirical distribution function and the black line is the theoretical distribution function

In the Figure10 we compare the packages "sn" and "fGarch" using the histograms of "ggplot2". In Figure11 we compare the curves of packages "sn" and "fGarch" with the standard normal curve. In Figure12 we represent seven histograms, showing that if the shape parameter increases, then the curve will be moved right.

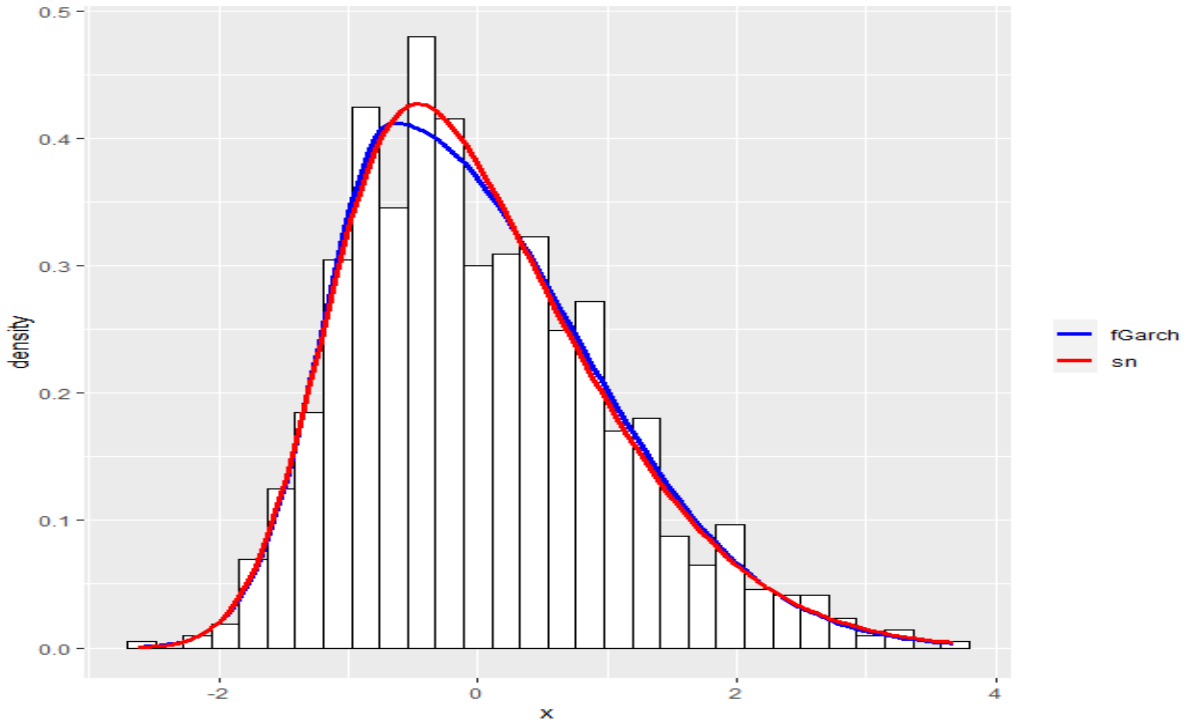


Figure 20

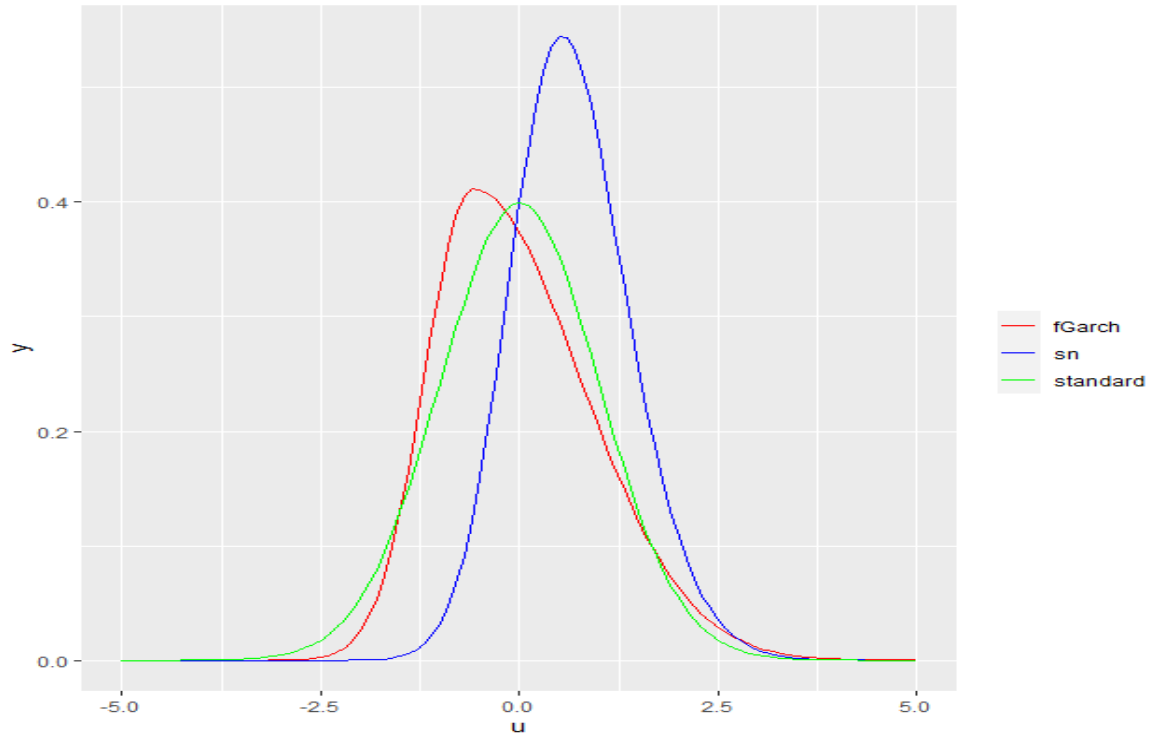


Figure 21

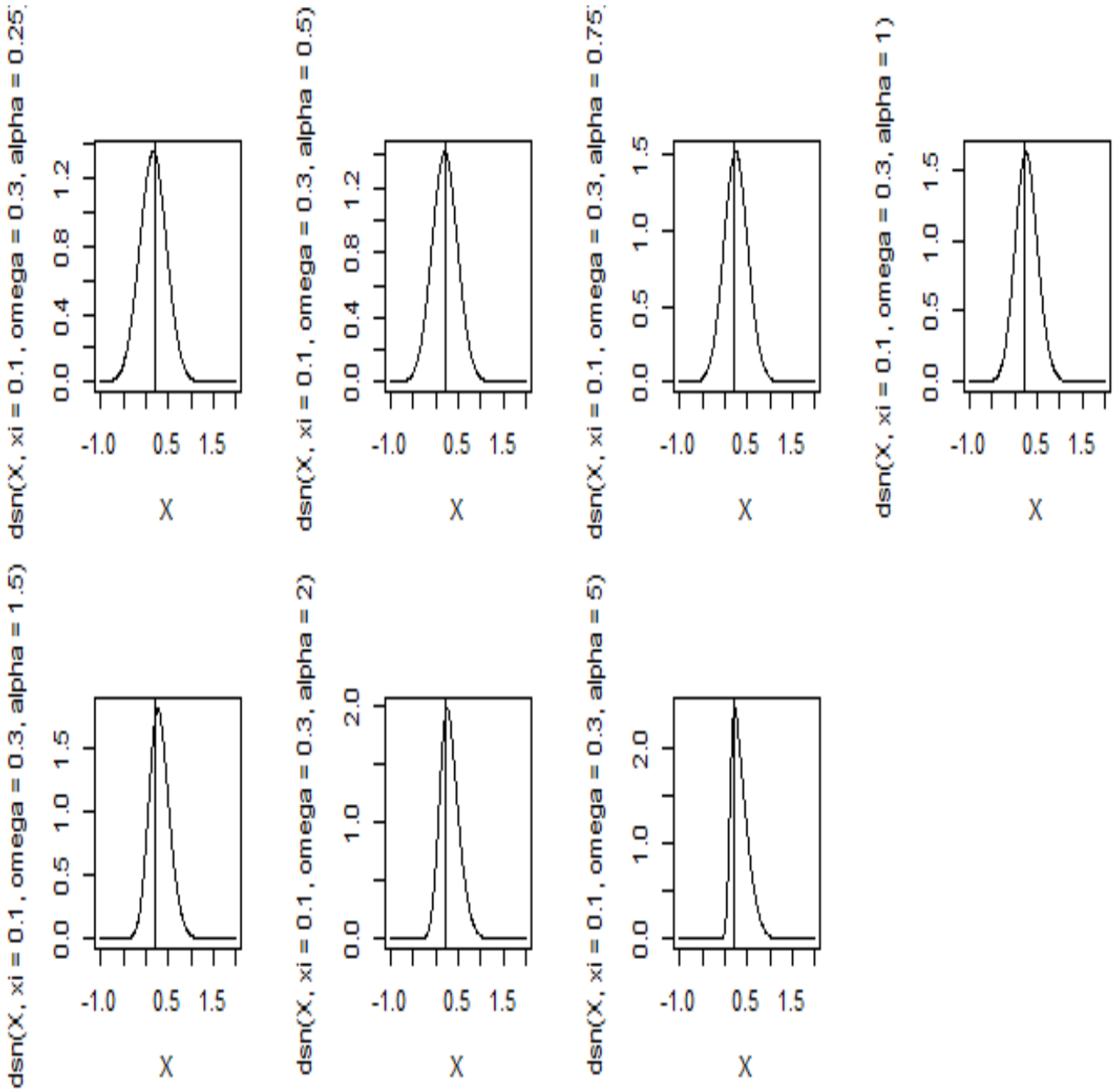


Figure 22 the shape parameter α take values {0.25, 0.5, 1.0, 1.5, 2.0, 5.0}

1.3.2 Power Comparisons of the Kolmogorov - Smirnov and Anderson - Darling Tests

Our Monte Carlo simulations indicate that the χ^2 test is indeed an unreliable statistic for testing Gaussianity (Normality) in small sample sizes, as suggested by DA86. Thus, we can eliminate the use of this test for classification of the galaxy group dynamics. However, the results of our simulation also show that both the Kolmogorov and A–D statistics are reliable down to n=5. In order to determine which test is most suitable for distinguishing between Gaussian and non-Gaussian systems, we perform power studies of the Kolmogorov and A–D tests. Monte Carlo simulations of both tests are performed, using a skewed Gaussian distribution, $F(x_i)$ (Azzalini & Capitanio1999).

$$F(x_i) = 2\varphi(x_i)\Phi(a_s x_i)$$

We draw various sample sizes, $5 \leq n \leq 100$, from a Gaussian random distribution, using 30,000 iterations, and apply varying levels of skewness, $0 \leq \alpha_s \leq 5$, to determine the rejection rate of both the Kolmogorov and A–D tests. The rejection rates are determined using the 10% critical values given in DA86. In order for a test to be considered powerful, the simulations with high α_s should have higher rejection rates, since the underlying distributions are increasingly less Gaussian.

1.3.2.1 Rejection rate of the Kolmogorov – Smirnov

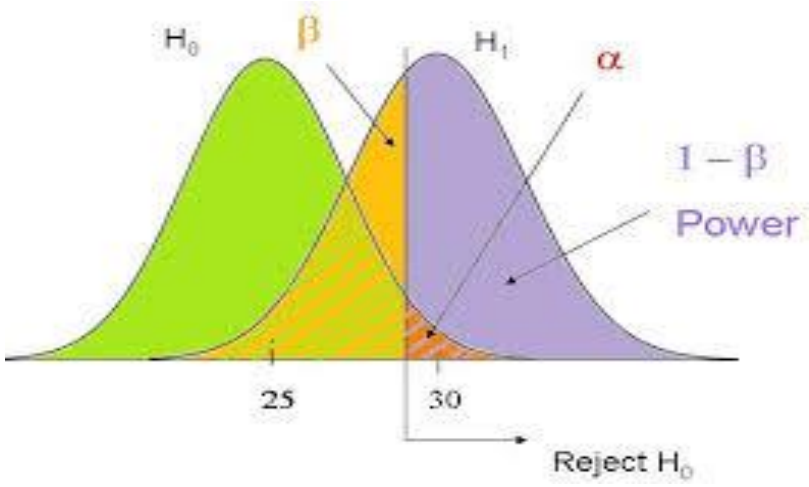
	n=5	n=10	n=15	n=20	n=25	n=30
as=0.25	0.1213333	0.1457333	0.1689667	0.1919667	0.2063000	0.2281000
as=0.5	0.1740333	0.2656667	0.3424000	0.4170667	0.4715000	0.5307000
as=0.75	0.2418333	0.3957667	0.5336333	0.6508667	0.7264000	0.7953667
as=1.0	0.3152667	0.5333667	0.6954333	0.8137000	0.8822667	0.9319667
as=1.5	0.4217333	0.7280333	0.8885000	0.9624667	0.9851333	0.9949333
as=2.0	0.5041333	0.8435000	0.9639667	0.9935000	0.9989000	0.9998333
as=5.0	0.7297000	0.9955667	1.0000000	1.0000000	1.0000000	1.0000000

	n=40	n=50	n=100
as=0.25	0.1793667	0.3125667	0.5052667
as=0.5	0.5241333	0.7241333	0.9460667
as=0.75	0.8195333	0.9443000	0.9991000
as=1.0	0.9550667	0.9937000	1.0000000
as=1.5	0.9988000	0.9999333	1.0000000
as=2.0	1.0000000	1.0000000	1.0000000
as=5.0	1.0000000	1.0000000	1.0000000

Next we have plotted the Kolmogorov-Smirnov p-value and the shape parameter a_s. We want to check the power of the test in function with the shape parameter. The formula of the power of the test is:

$$p_{value} < a = (0.1, 0.05, 0.005)$$

So, as the shape parameter increases then the normality of the hypothesis H_o ceases to exist. In Figure13 we see that as the shape parameter increases then the k-s power test is approximately going to 1.



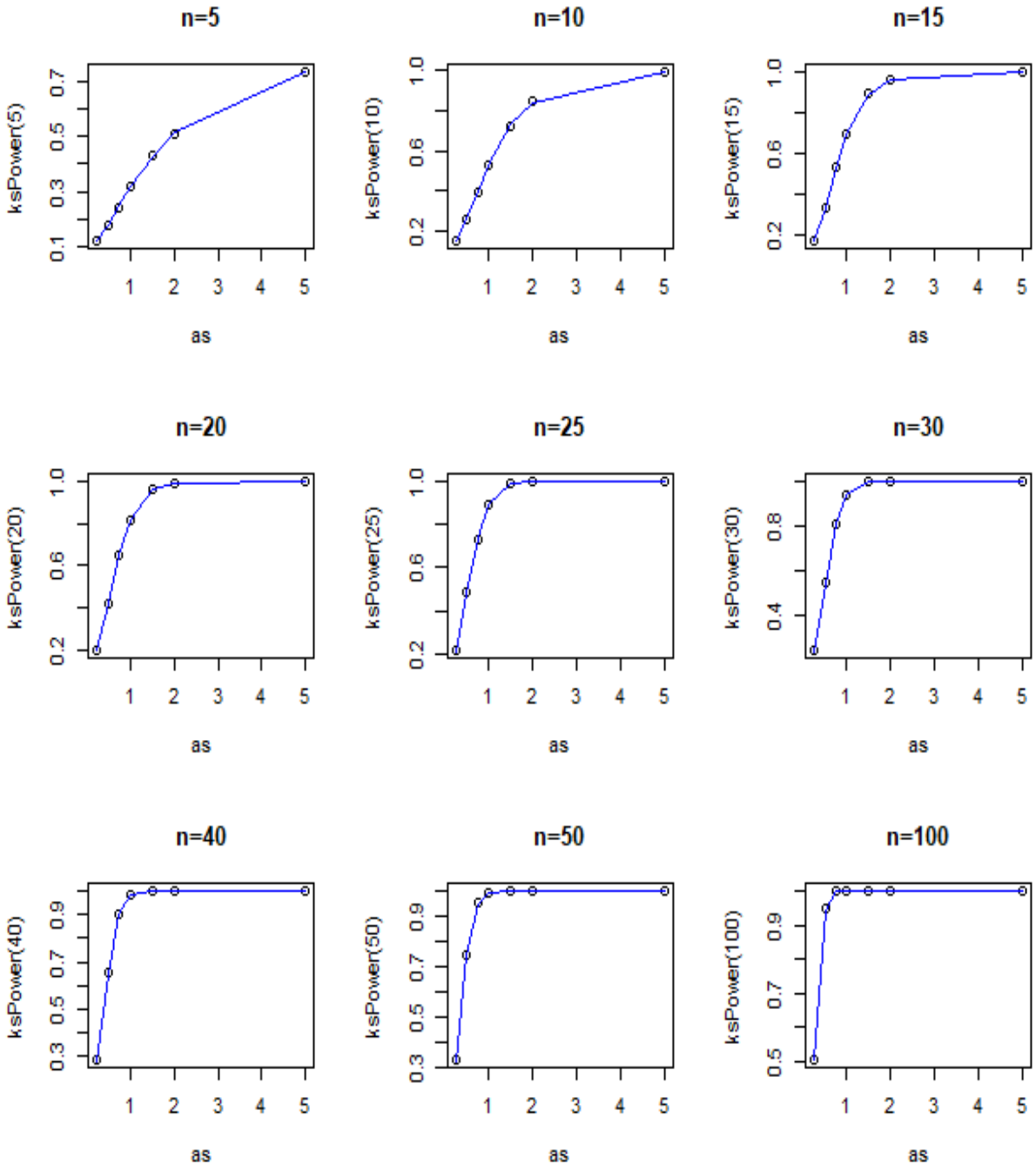


Figure 23 We see that as the shape parameter (as) increases then for large n (e.x. n=40,50,100) the power goes to 1 faster than for smaller n (e.x. n=5,10..)

1.3.2.2 Calculations for K-S Test Rejection

The k-s test function takes into parameters, a data set and a character string naming which distribution we are testing. In our case, we will call ks.test with the parameter "pnorm", as we did in the above calculations and we will examine some of the properties of the power function. This will be done by considering some data with the form $Z(\delta) = X + (\delta)Y$. Where $X \sim N(0,1)$, $Y \sim Unif[-1,1]$ and δ is a value we use to skew the data. The first thing we look at is how the K-S test reacts to changes in δ. Figure14 shows a plot of the observed α level against δ for a few chosen δ values.

We saw at $\delta = 0$, we get a significance level of $a = 0.10$ as to be expected. As δ increases, the observed significance level increases very rapidly. This demonstrates that the K-S test recognizes small alterations to our data and rejects more often. This also indicates that the K-S test is fairly conservative, since it detects changes in the data set.

Now, we fix a small number of δ values and change the sample size n in order to gauge what the power function does when the sample size is changed. In this case we take 30000 samples of $n = 5, 10, 15, \ldots, 100$ and plot the observed rejection rate against n. The test in this case is set up such that:

$$H_o: X \sim N(0,1) \; VS \; H_1 : X \sim N(0,1) + \delta Unif[-1,1]$$

In this case we want this proportion to be greater than or equal to 0.90 and we would like to know what sample size is needed to achieve these requirements. Figure14 shows the percentage of times the test rejected as a function of sample size for various values of δ. This graphic shows that for very contaminated data, we need a very small sample size. This leads to the conclusion that the K-S test has a good power, but we will prove below that the A-D is stricter than the K-S.

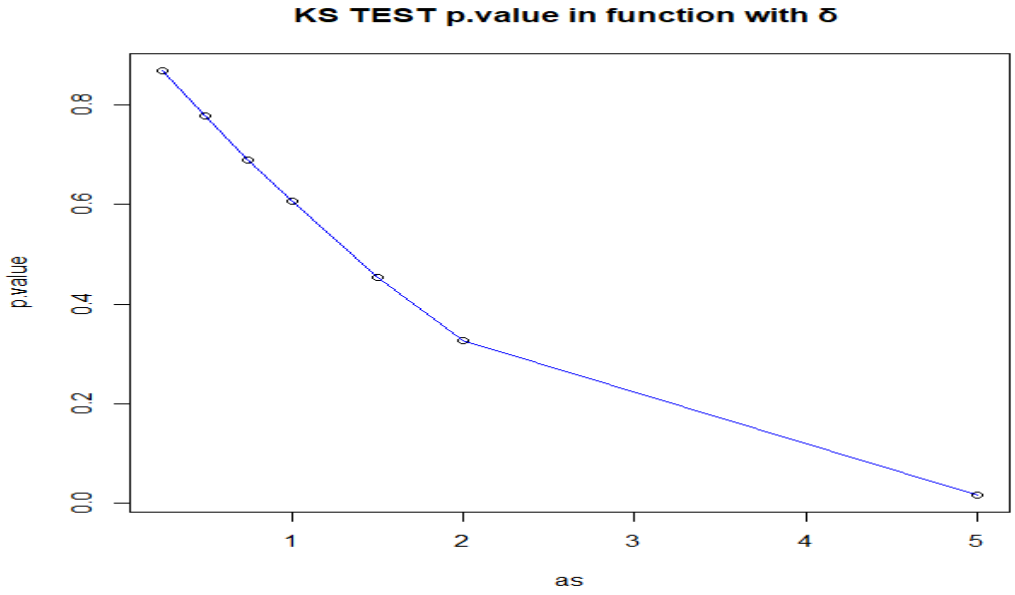


Figure 24

1.3.2.3 Rejection rate of the Anderson – Darling

	n=5	n=10	n=15	n=20	n=25	n=30
as=0.25	0.1324333	0.1564000	0.1783667	0.2008667	0.2294667	0.2589000
as=0.5	0.2035667	0.2781333	0.3531333	0.4227667	0.4994000	0.5652333
as=0.75	0.3262667	0.4771333	0.6060667	0.6956000	0.7798333	0.8353000
as=1.0	0.4707333	0.6594333	0.8007000	0.8794667	0.9335333	0.9612333
as=1.5	0.6758667	0.8753667	0.9542667	0.9849333	0.9952667	0.9985333
as=2.0	0.7813333	0.9408333	0.9850000	0.9964333	0.9995333	0.9997667
as=5.0	0.8837333	0.9828667	0.9984000	0.9998333	0.9999667	1.0000000

	n=40	n=50	n=100
as=0.25	0.3286667	0.4029	0.8266000
as=0.5	0.6854333	0.7797	0.9812333
as=0.75	0.9162667	0.9572	0.9990333
as=1.0	0.9875333	0.9969	1.0000000
as=1.5	0.9999000	1.0000	1.0000000
as=2.0	1.0000000	1.0000	1.0000000
as=5.0	1.0000000	1.0000	1.0000000

Next we compute the expected false positive rates. "Rates" is in the plural because the properties of a test usually vary with the sample size. Since it is well-known that distributional tests have high power against qualitatively small alternatives when sample sizes are large, this study focuses on a range of small sample sizes where such tests of often applied in practice: typically about 5 to 100.

We plotted the p-values and we take the power of the Anderson – Darling Test as we proved in the above tables.

For shape parameter $a_s = 0.25$

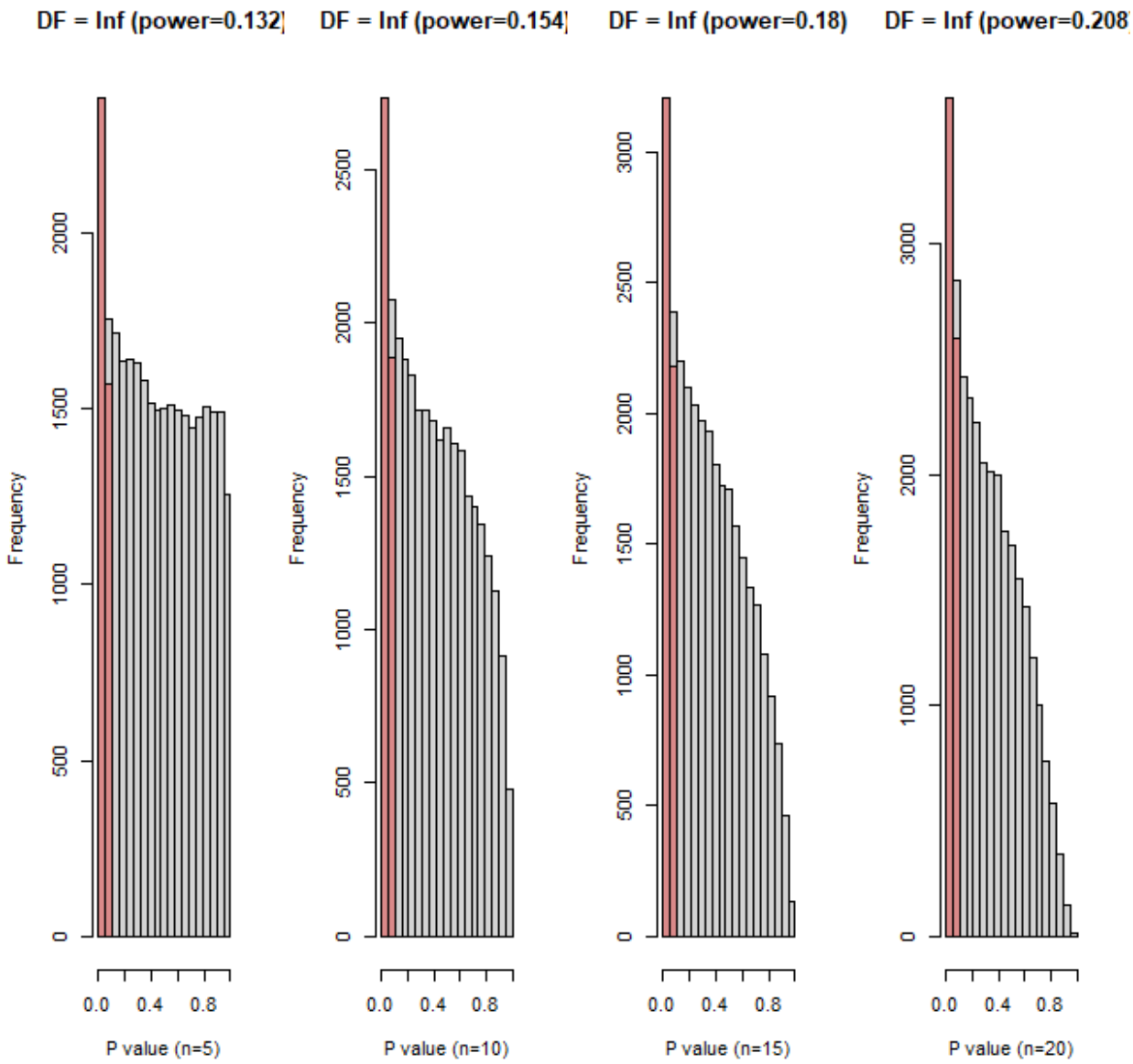


Figure 25

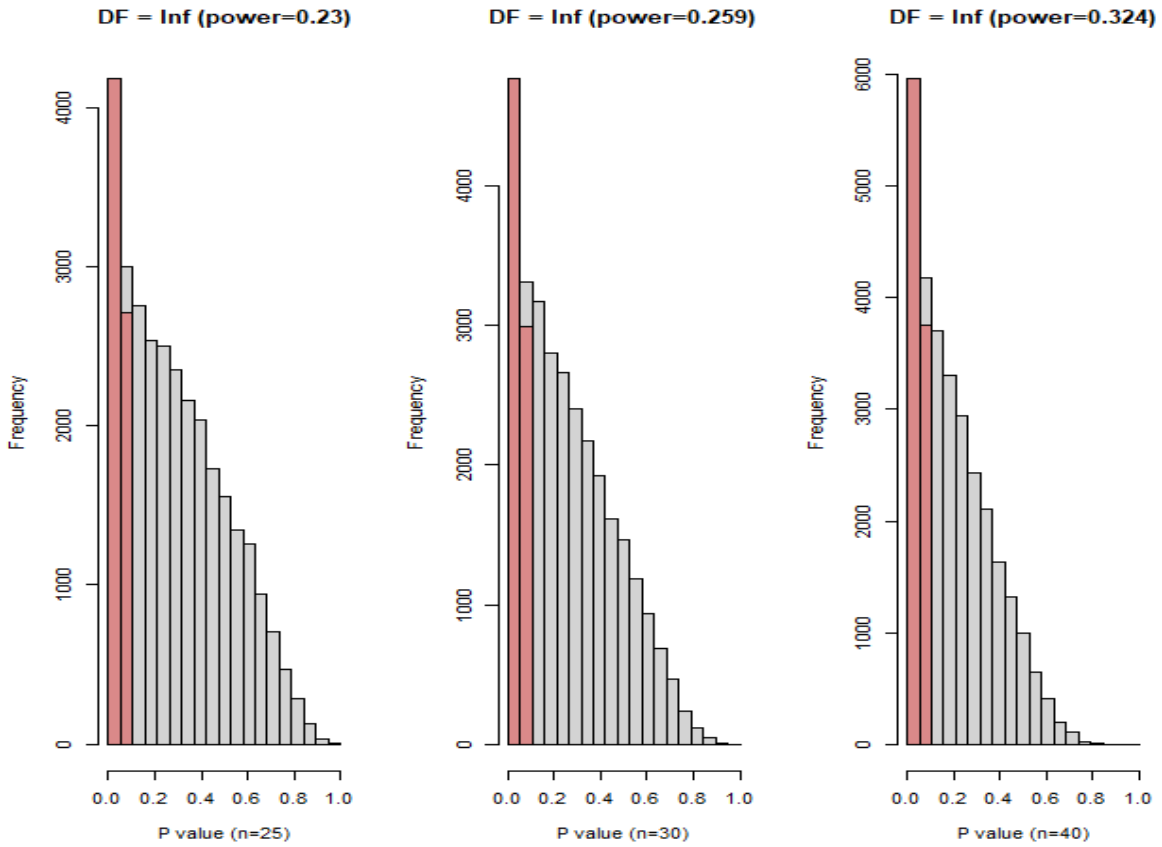


Figure 26

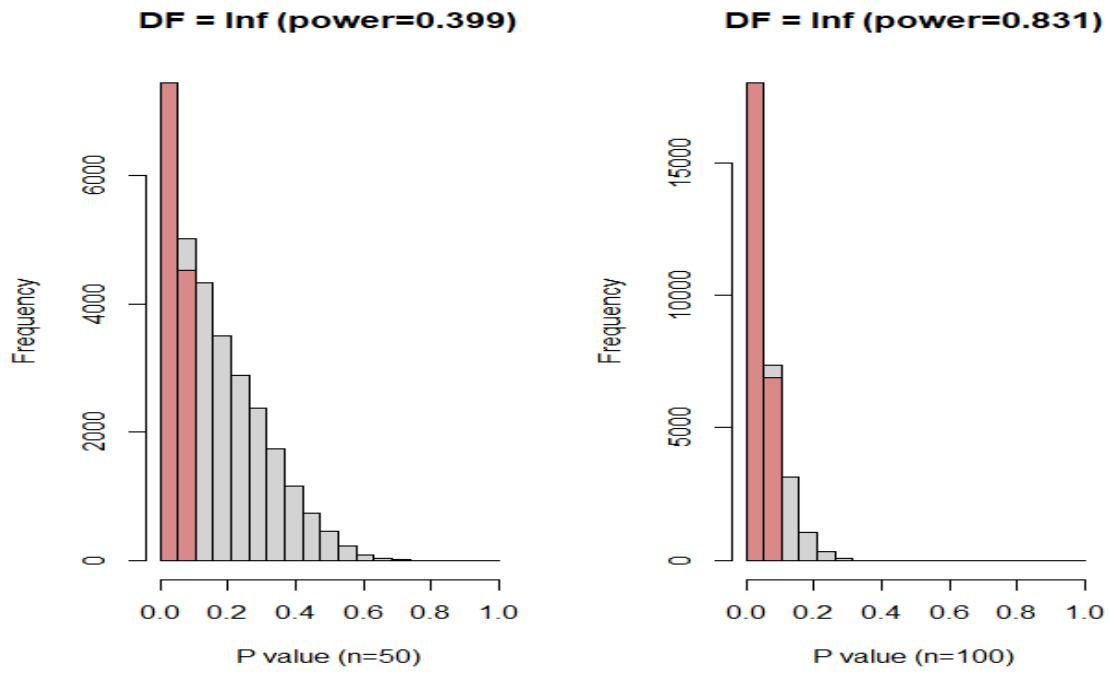


Figure 27

For $a_s = 0.50$

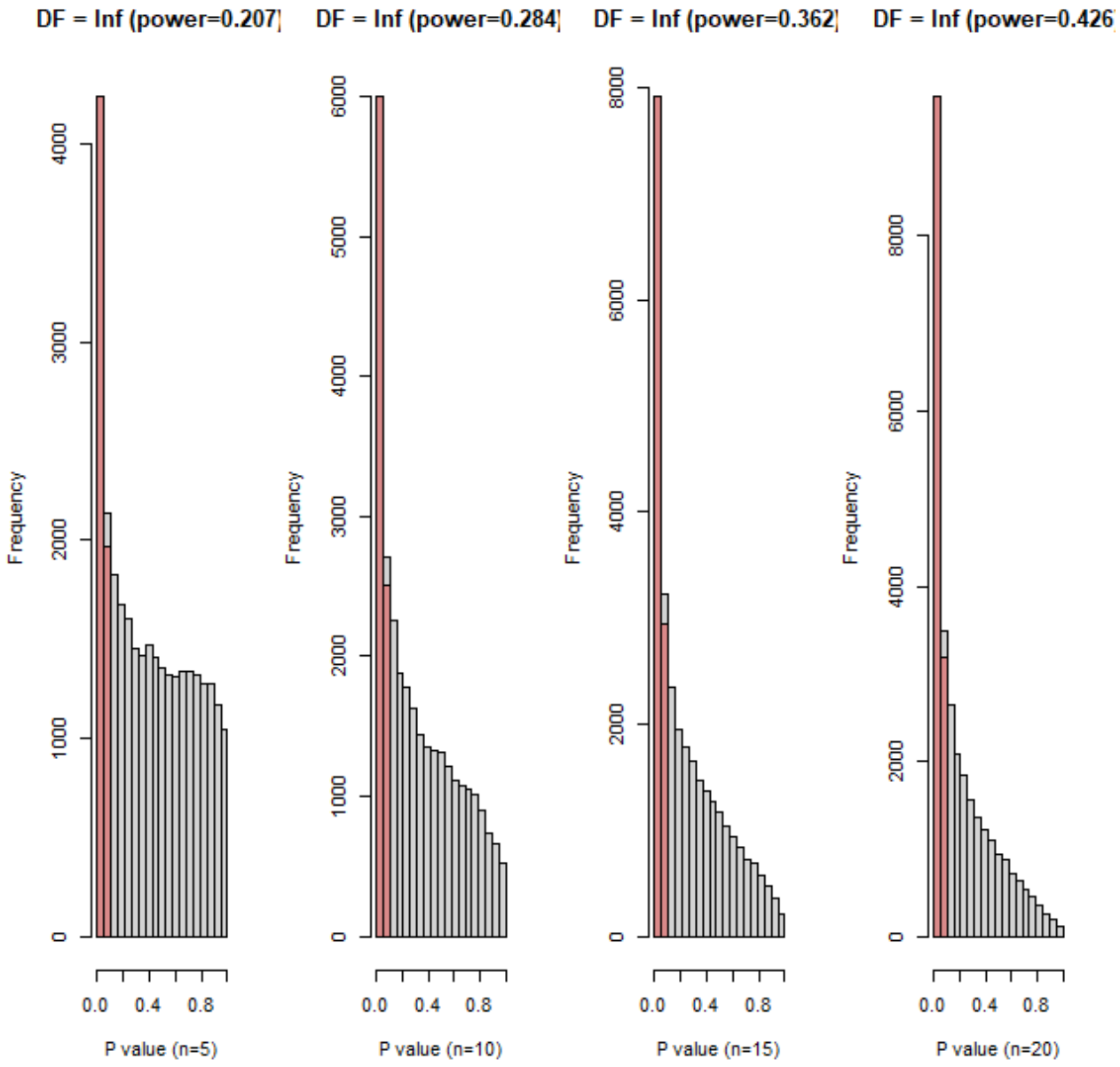


Figure 28

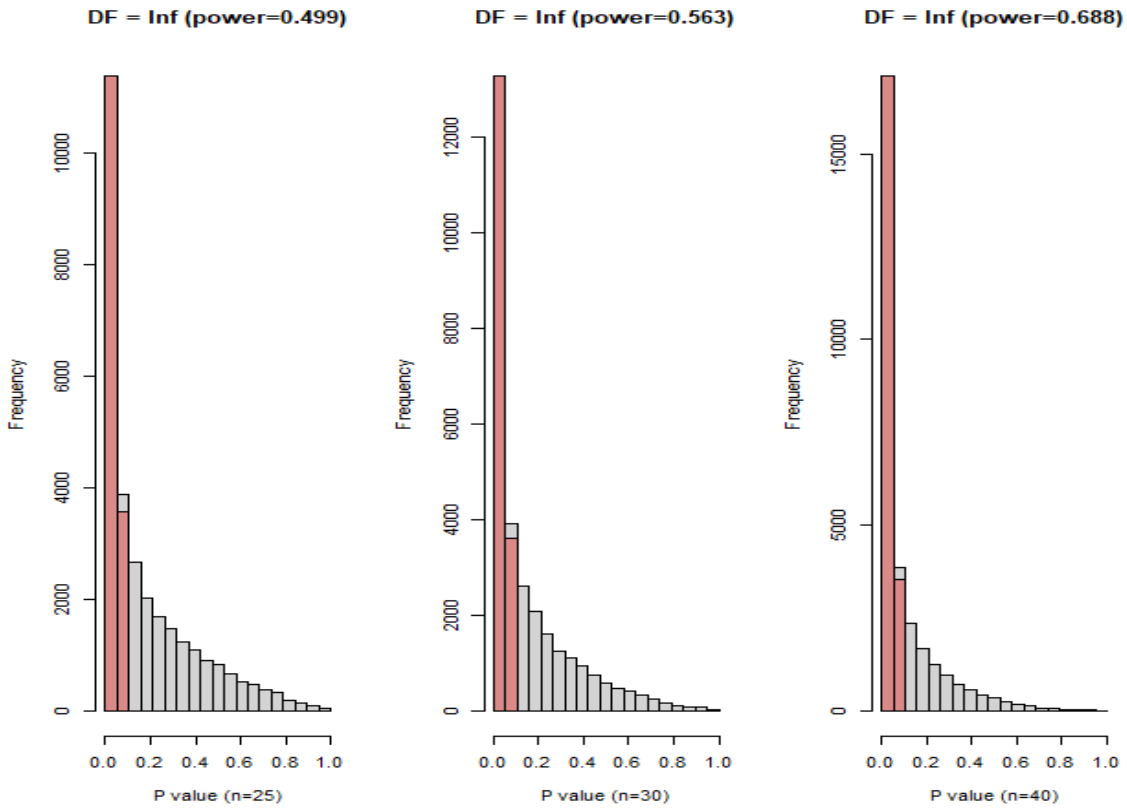


Figure 29

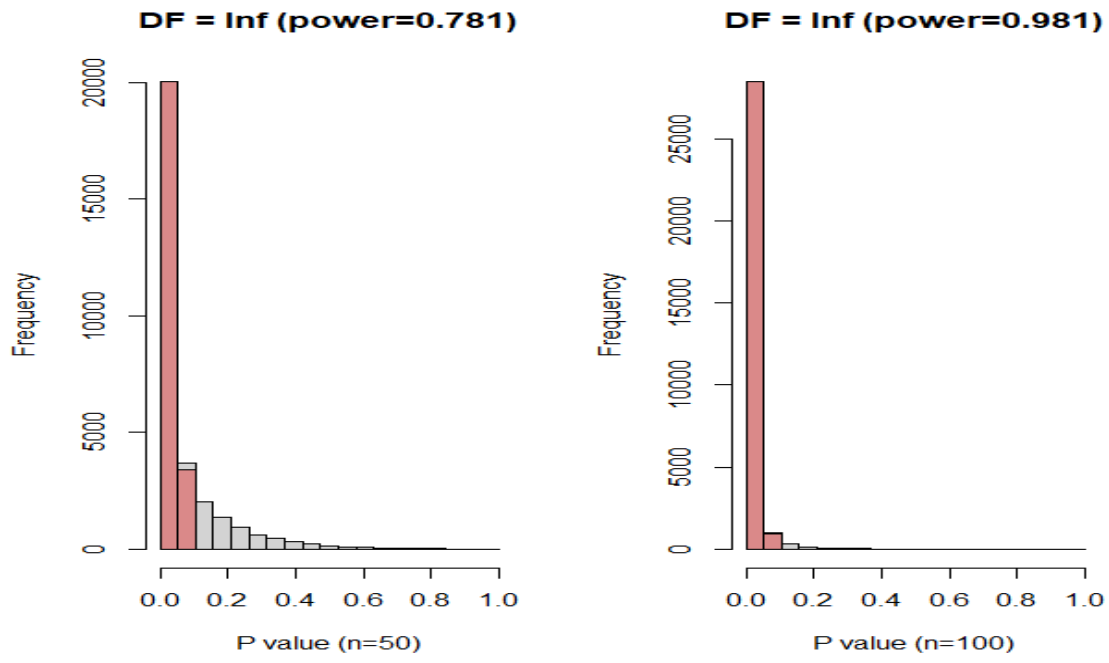


Figure 30

For $a_s = 0.75$

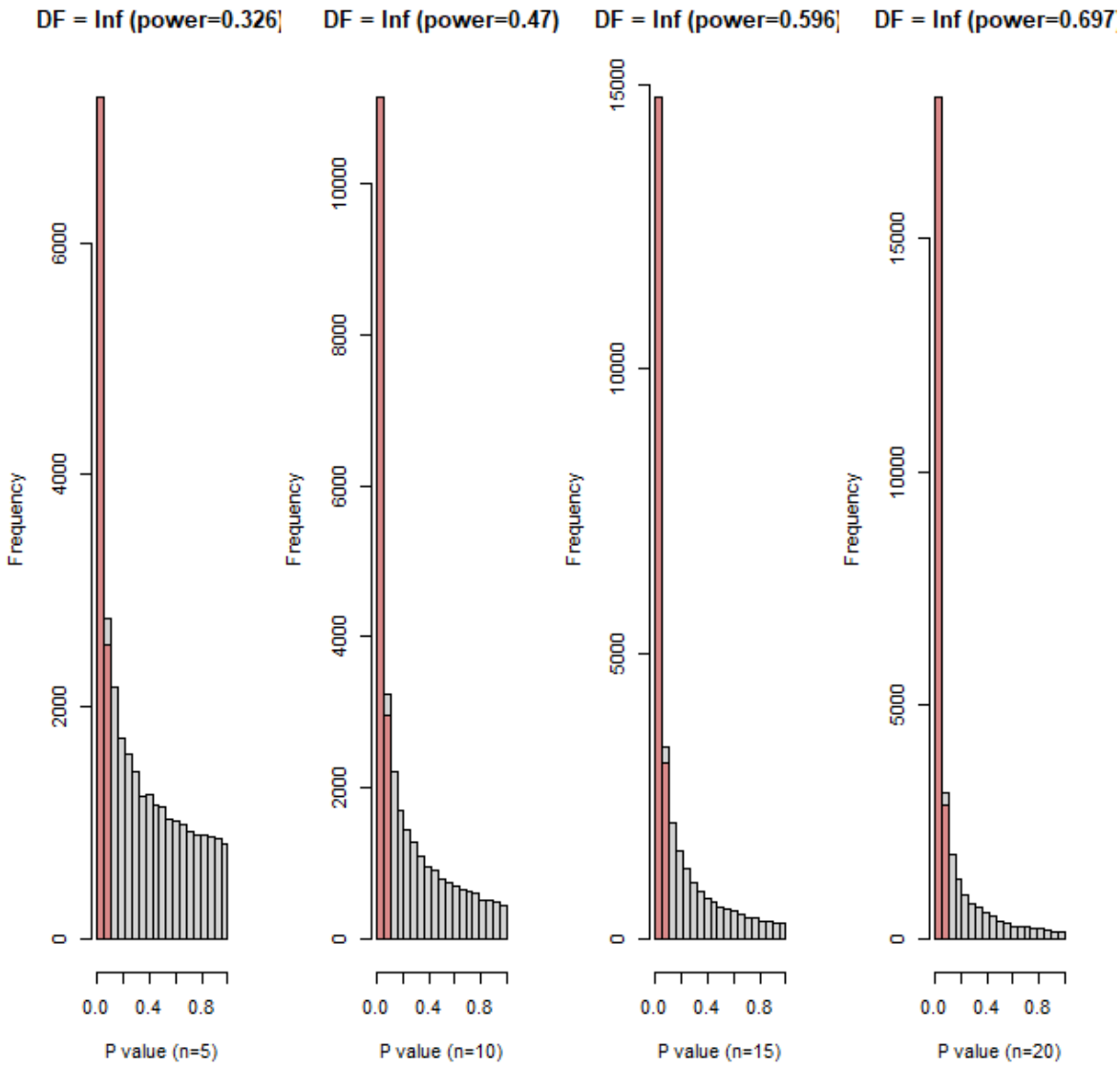


Figure 31

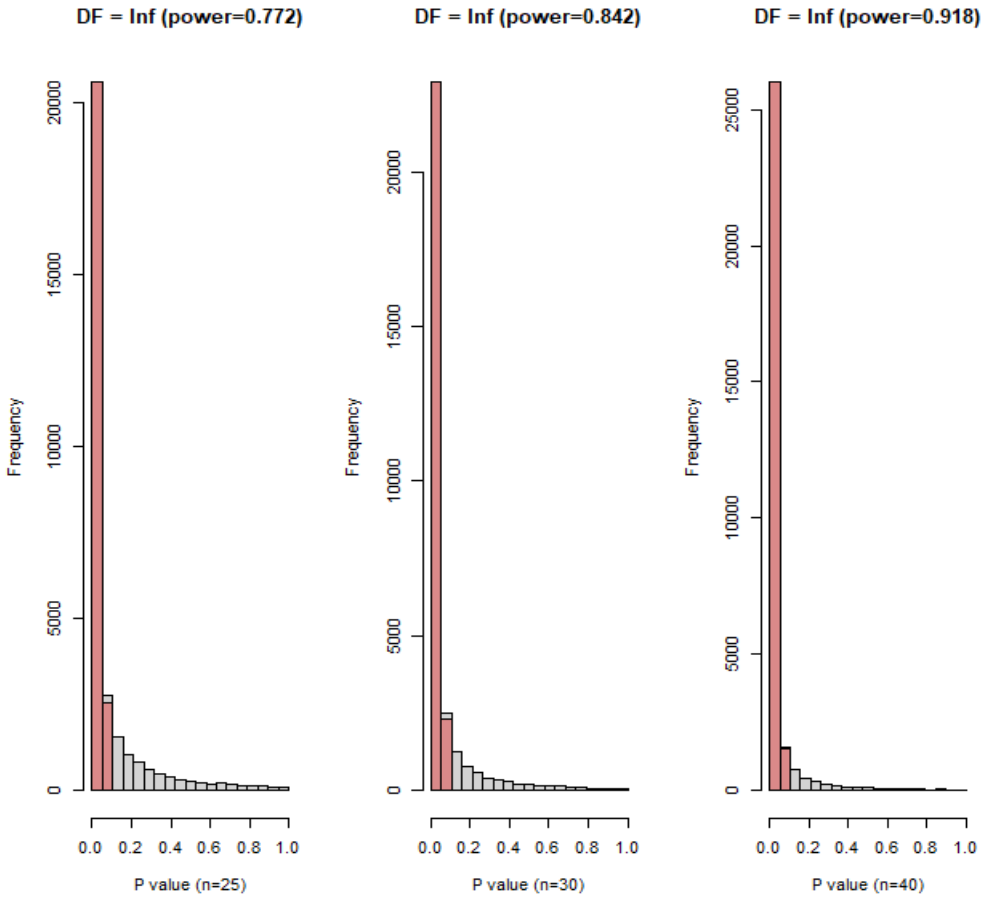


Figure 32

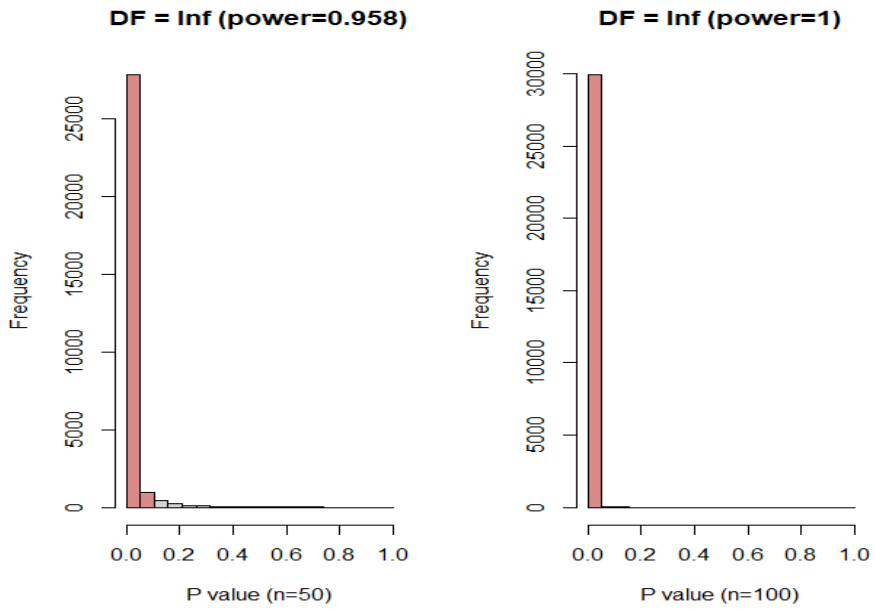


Figure 33

For $a_s = 1.00$

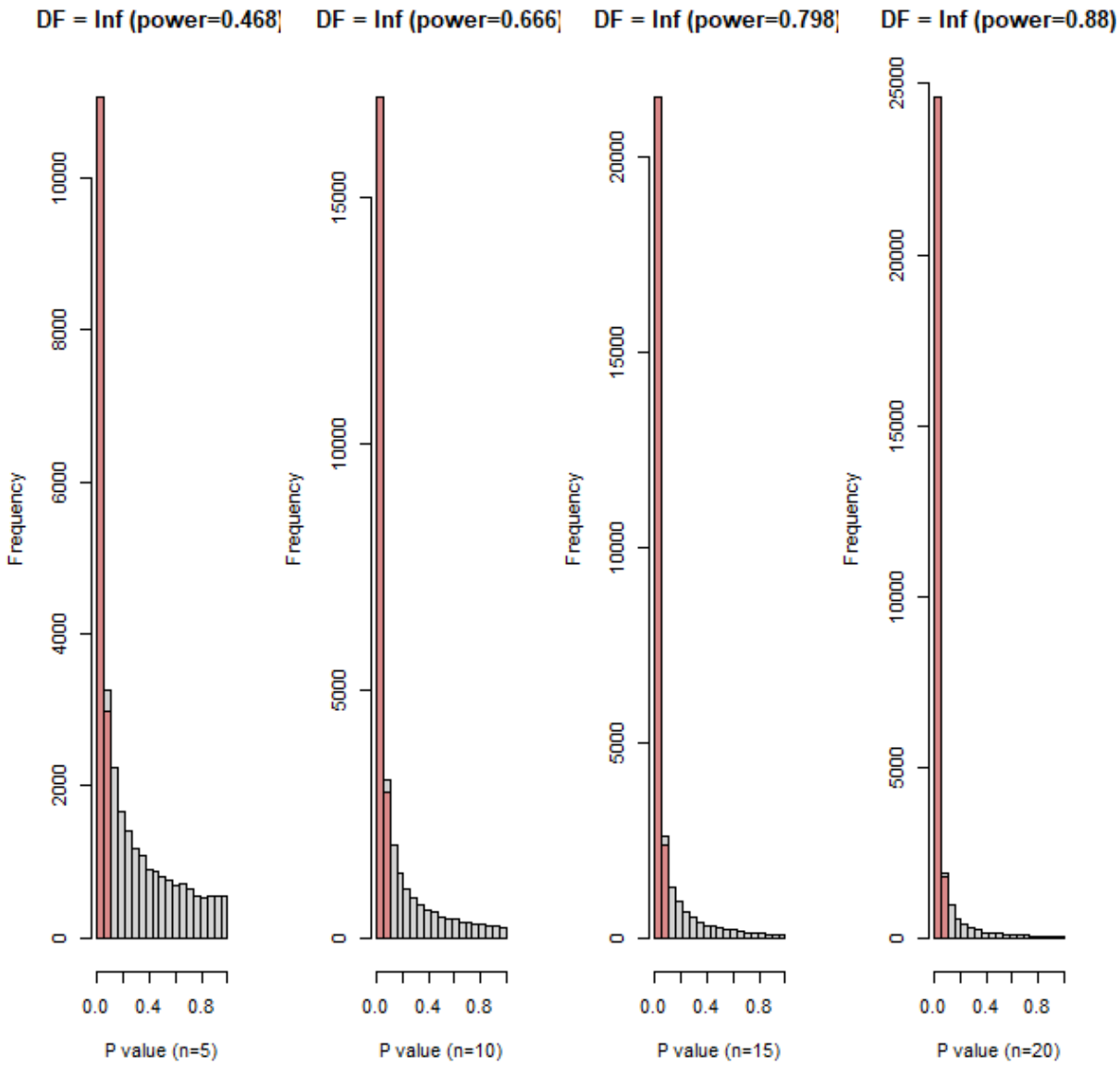


Figure 34

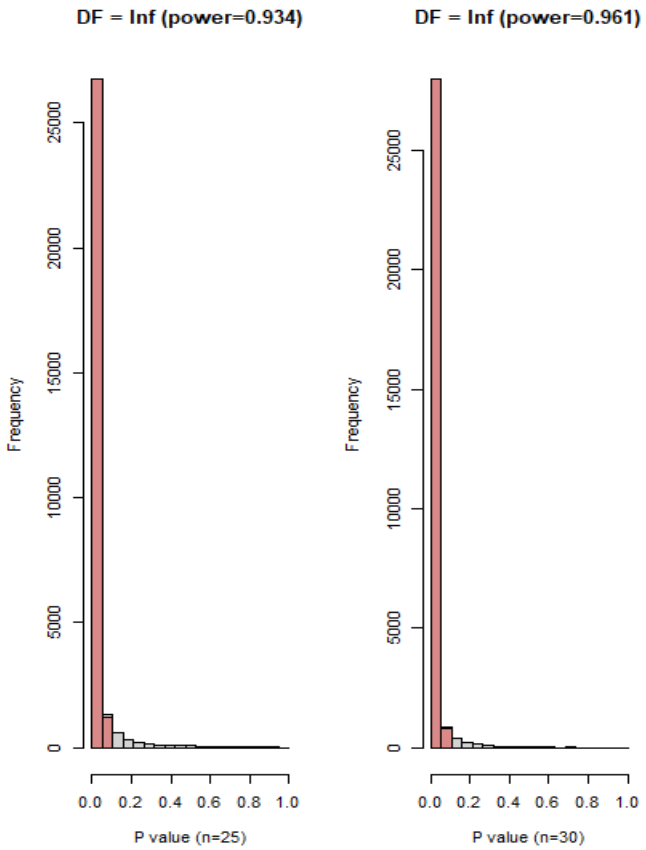


Figure 35

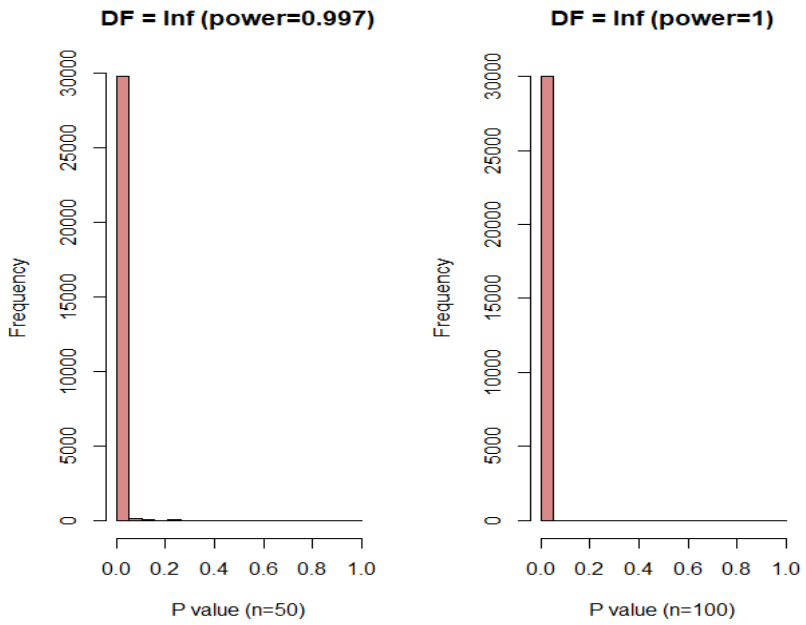


Figure 36

After $a_s > 1.0$ the plotted histograms of p-value and power are approximately the same because the p-value goes approximately to zero (0).

1.3.3. Power Comparisons of the Kolmogorov - Smirnov and Anderson - Darling Tests of Paper Hou.et.al

The plotted rejection rate of Kolmogorov – Smirnov test

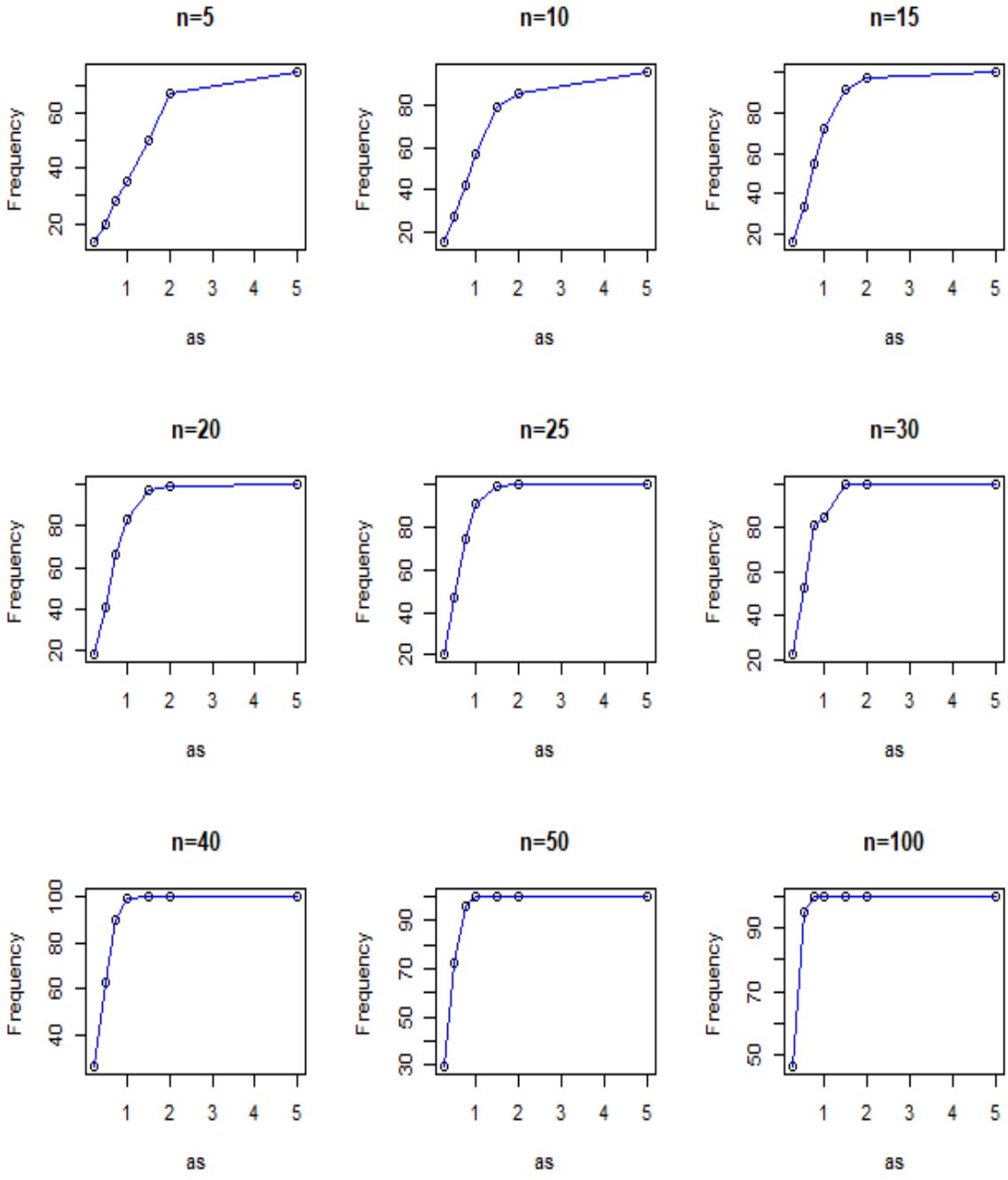


Figure 37

The plotted rejection rate of Anderson – Darling test

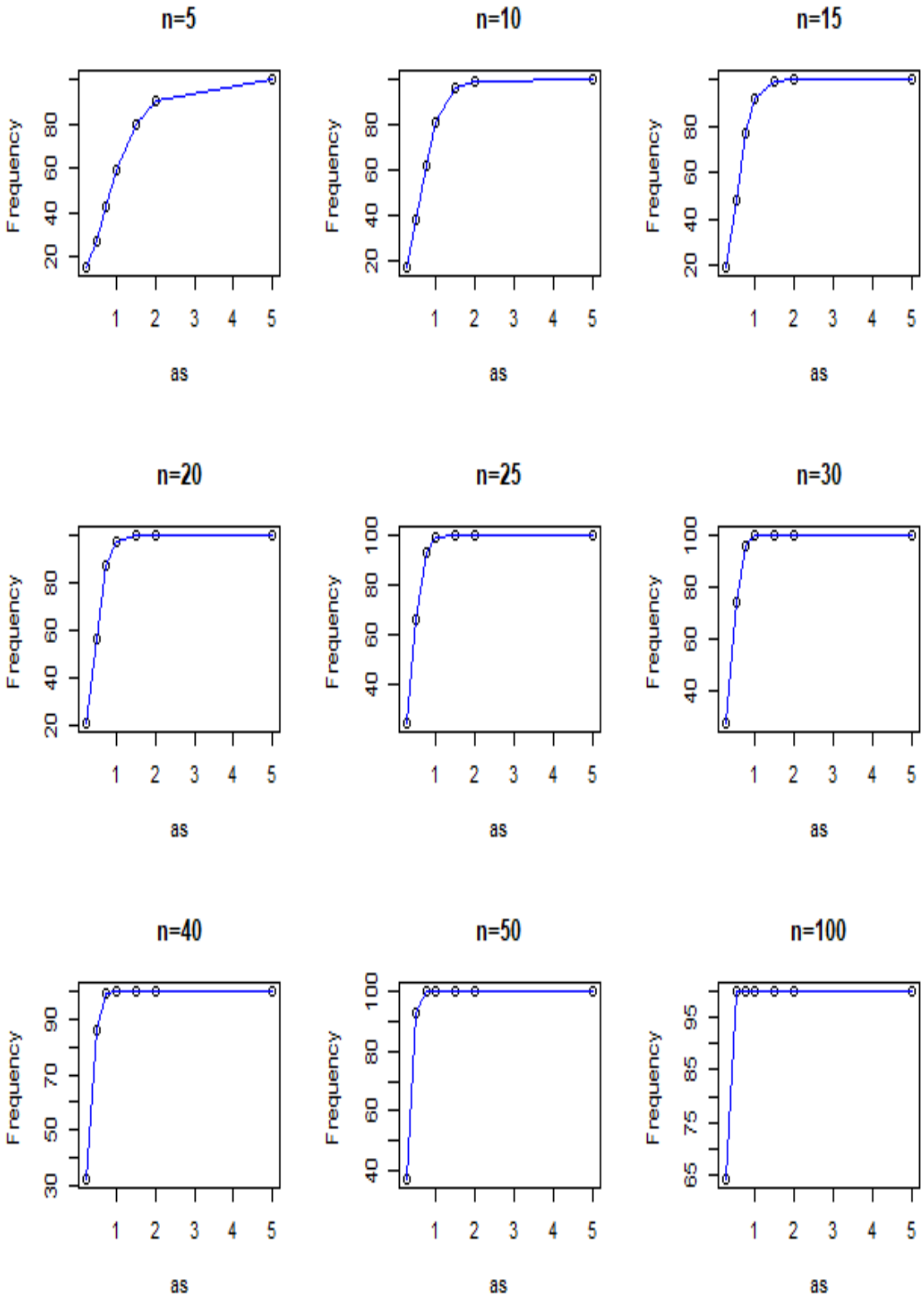


Figure 38

The results of the Power Comparisons of The Kolmogorov and A-D test Using Various Skewed Gaussian Distributions

Table 1

Results of the Power Comparisons of the Kolmogorov and A–D Tests Using Various Skewed Gaussian Distributions[a]

Test	α_S	n								
		5	10	15	20	25	30	40	50	100
A^2	0.25	15	17	19	21	24	27	32	37	64
D^*	0.25	13	15	16	18	20	22	26	29	46
A^2	0.50	27	38	48	57	66	74	86	93	100
D^*	0.50	20	27	34	41	47	53	63	72	95
A^2	0.75	43	62	77	87	93	96	99	100	100
D^*	0.75	28	42	55	66	74	81	90	96	100
A^2	1.0	59	81	92	97	99	100	100	100	100
D^*	1.0	35	57	72	83	91	85	99	100	100
A^2	1.5	80	96	99	100	100	100	100	100	100
D^*	1.5	50	79	91	97	99	100	100	100	100
A^2	2.0	90	99	100	100	100	100	100	100	100
D^*	2.0	67	86	97	99	100	100	100	100	100
A^2	5.0	100	100	100	100	100	100	100	100	100
D^*	5.0	75	96	100	100	100	100	100	100	100

Note. [a] This table gives the percentage of rejection based on the 10% critical values given in DA86.

We continue our analysis plotting the power of the two (2) tests in function with the sample size n for various shape parameters.

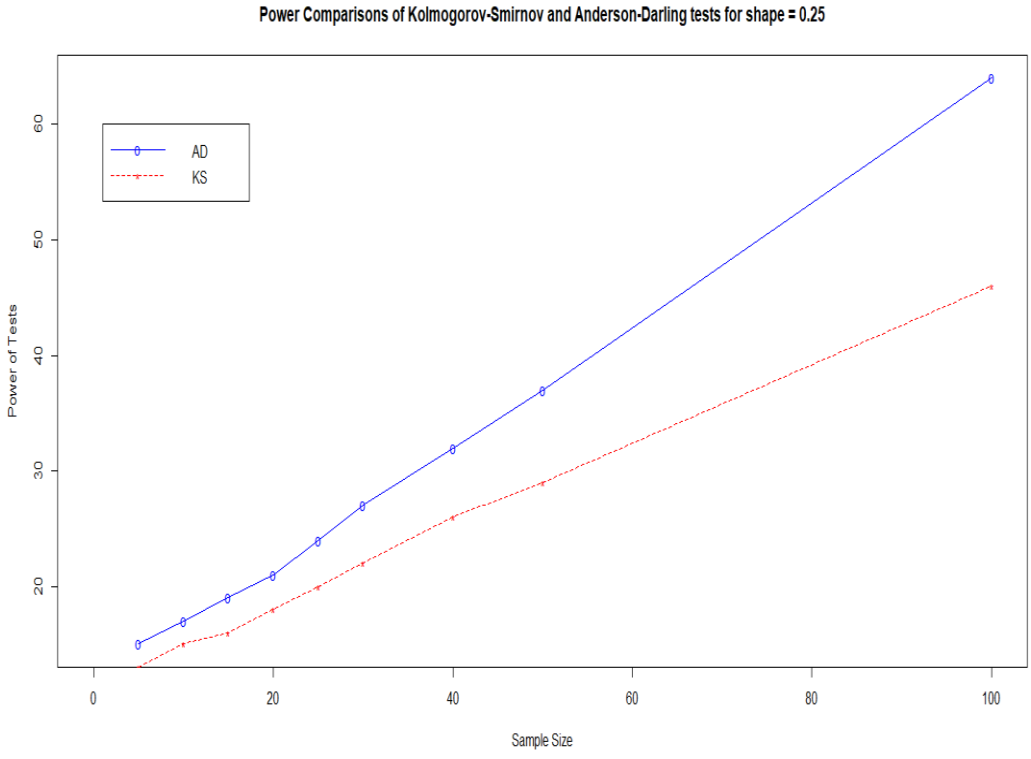


Figure 39

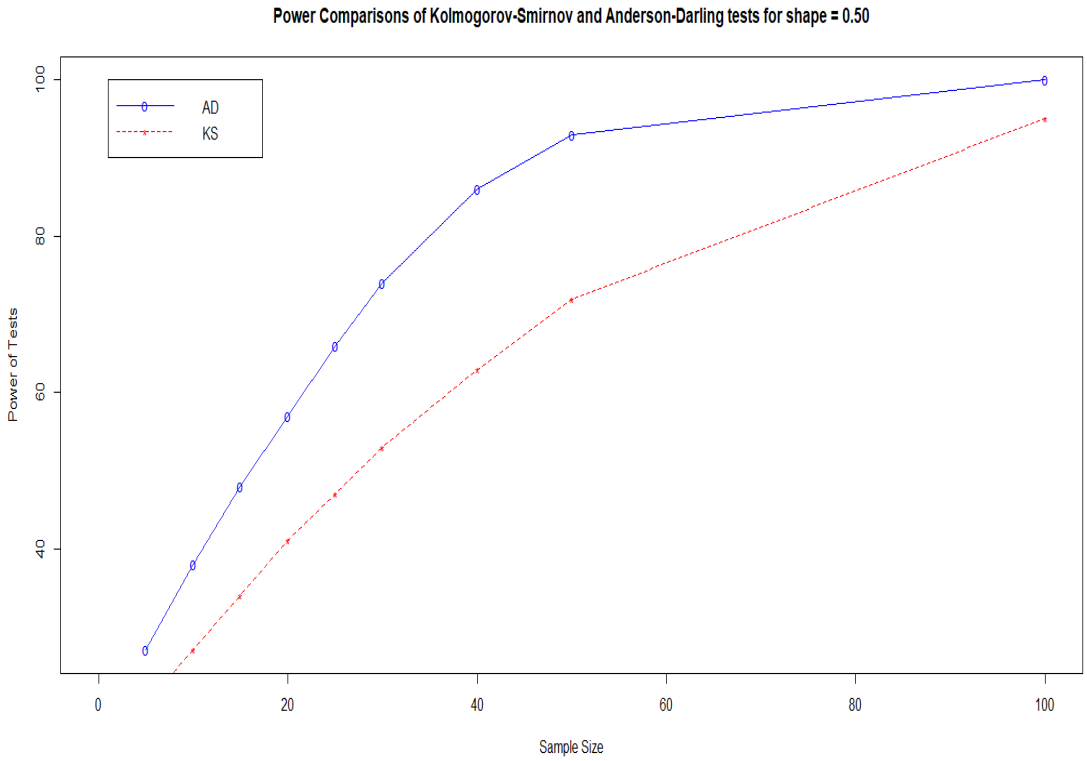


Figure 40

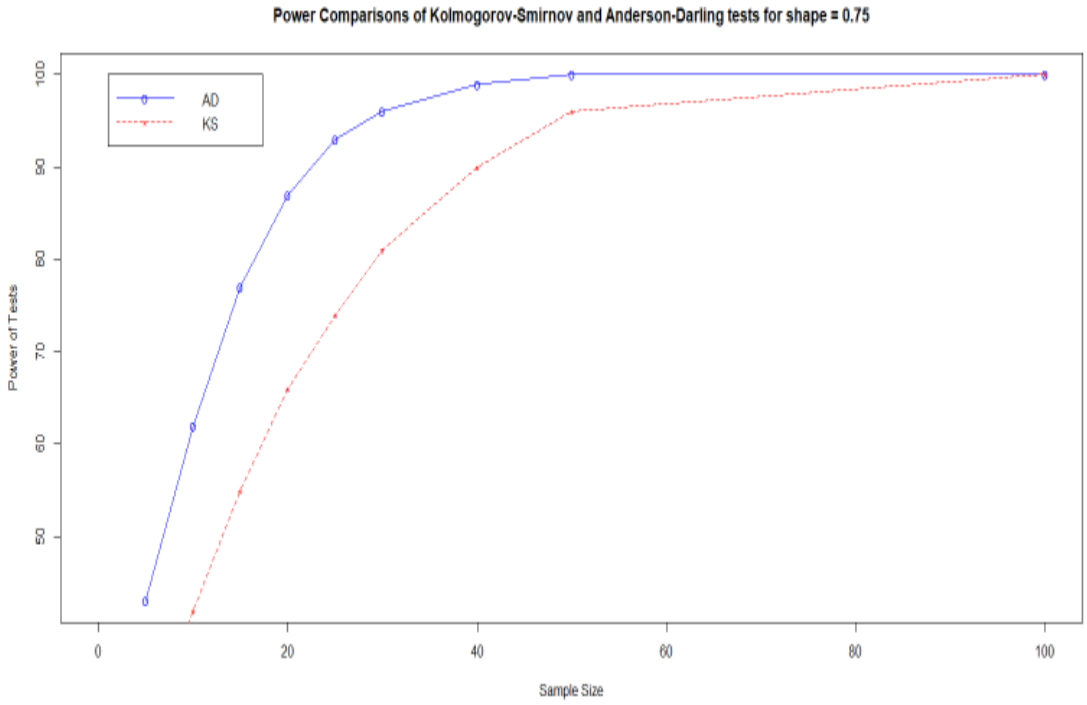


Figure 41

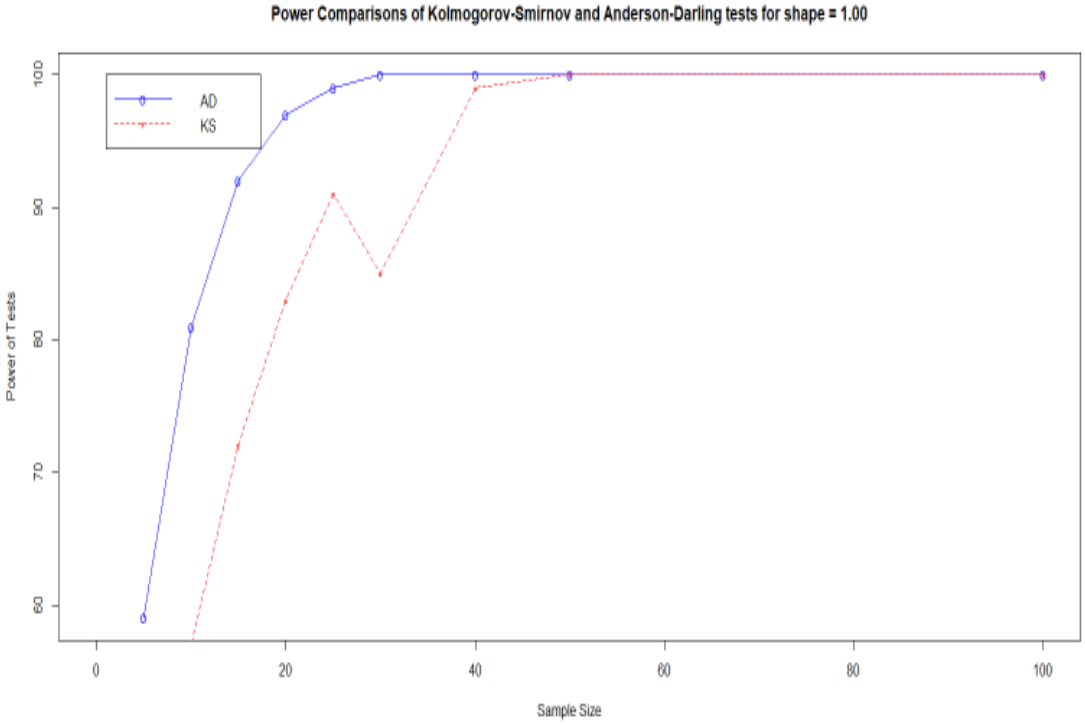


Figure 42

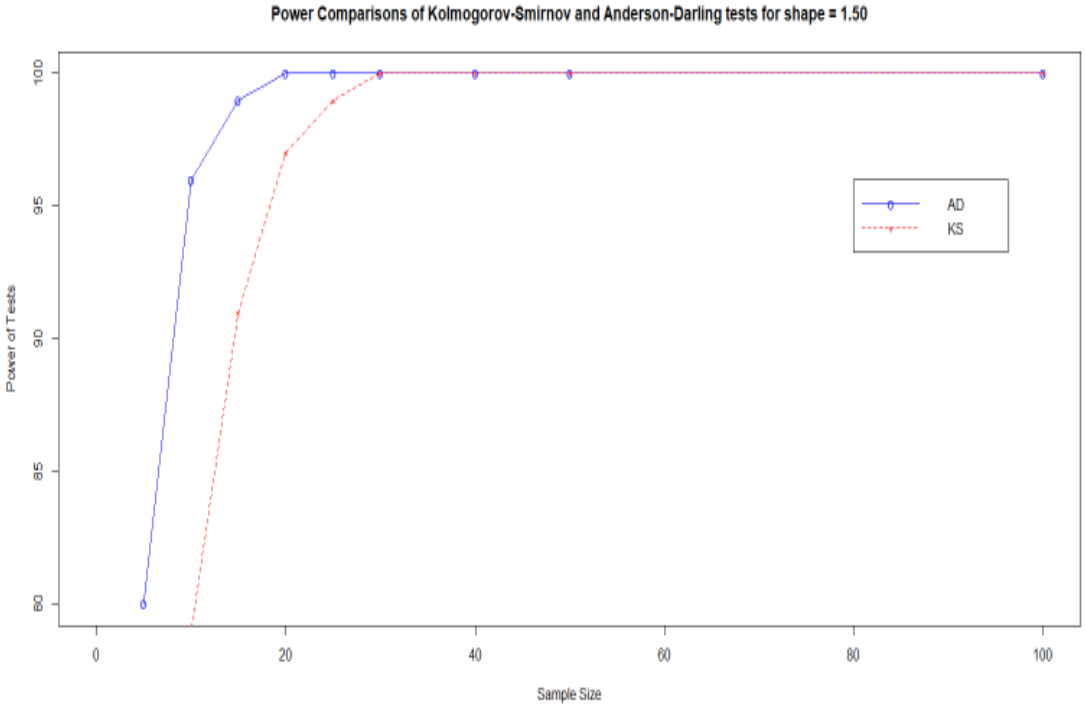


Figure 43

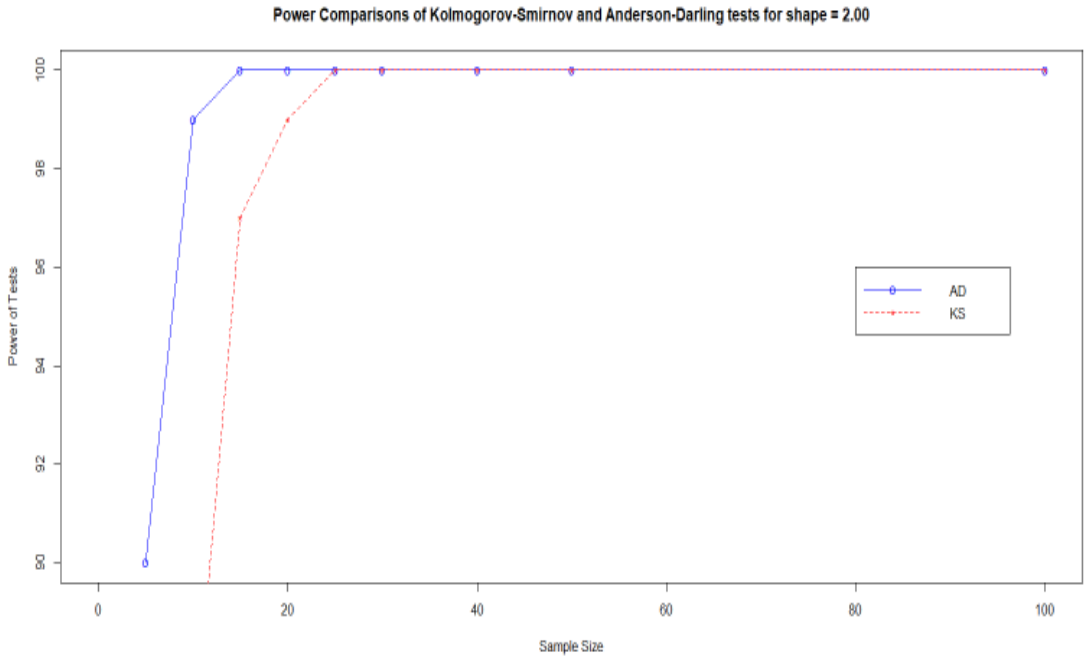


Figure 44

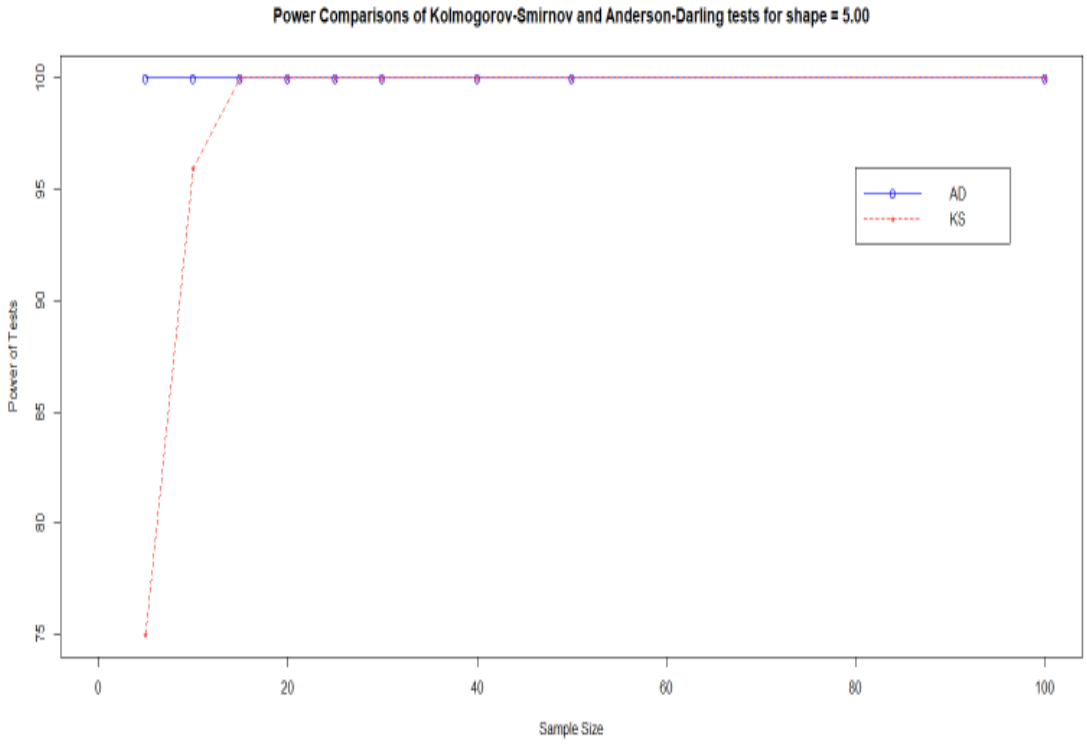


Figure 45

We understand that as the shape parameter increases the curves of KS and AD test going to converge each other.

So we conclude that the Anderson – Darling test is stricter than the Kolmogorov-Smirnov test, because the rejection rate of the A-D is higher than the K-S test.

1.3.4 Azzalini's package Skew Normal Distribution

We used the package of Azzalini for the power comparisons of Kolmogorov-Smirnov and Anderson-Darling tests

We have plotted the probability density function of a skew normal for different values of alpha parameter from the package of the Azzalini called "sn".

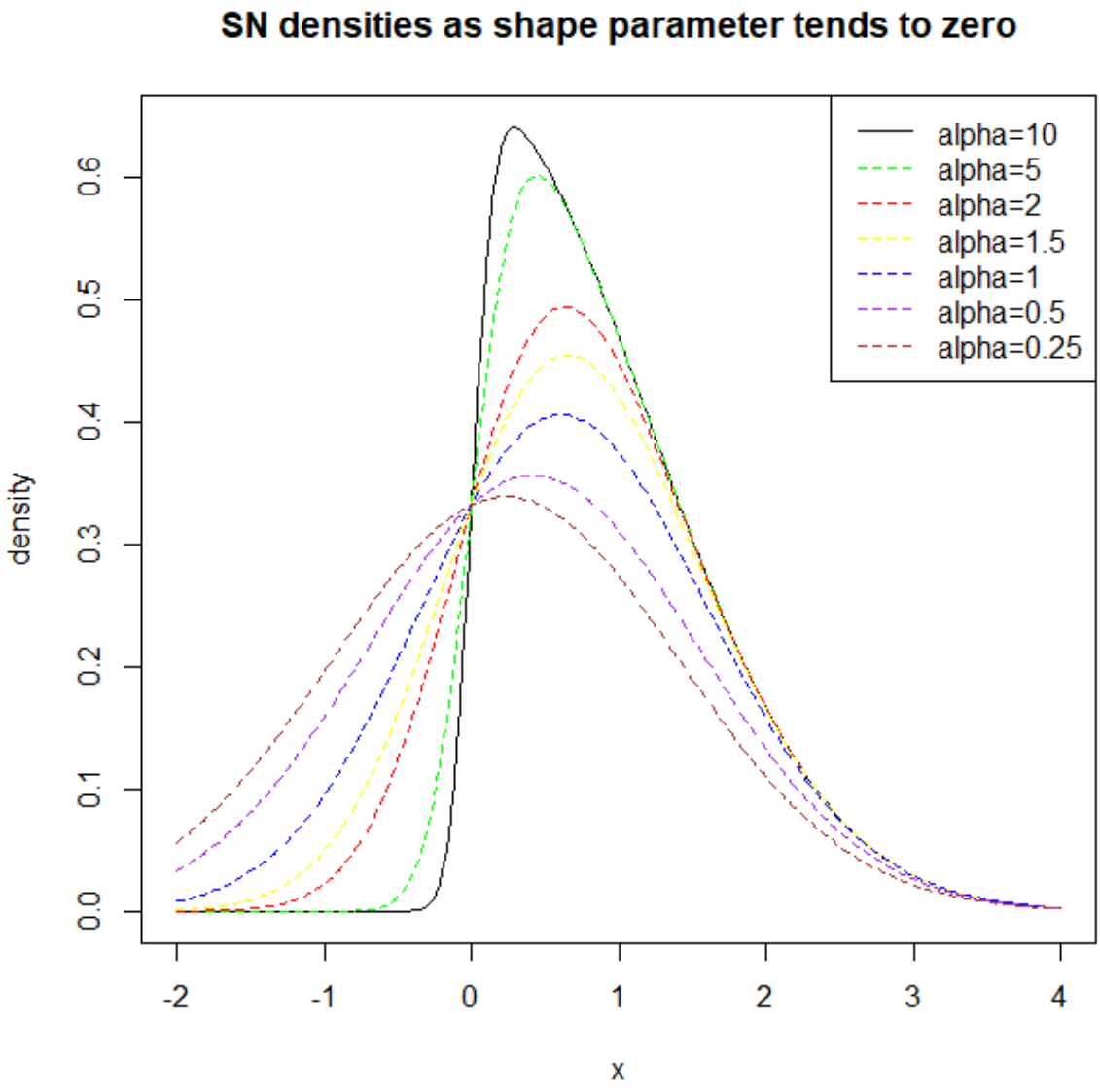


Figure 46

Also we have plotted the 2 D density distribution of skew normal.

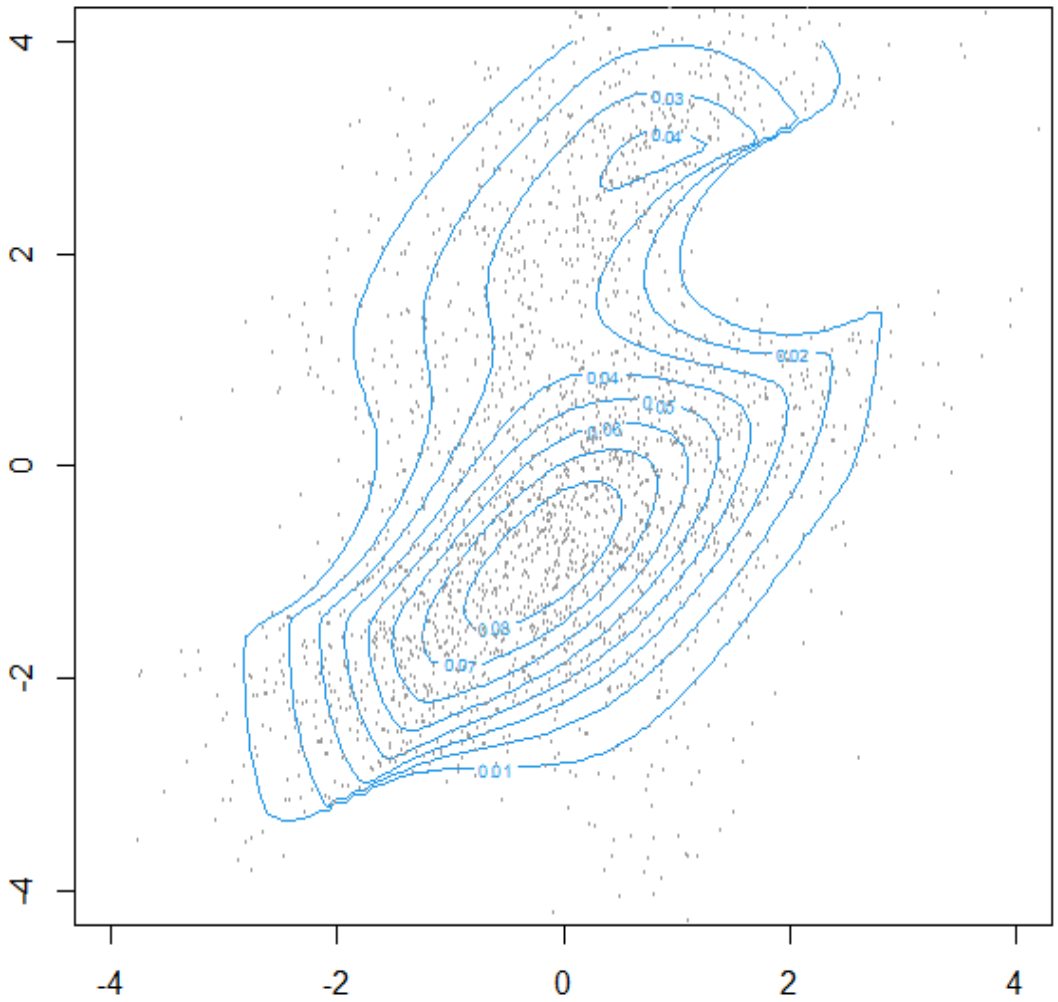


Figure 47

2. Applying the Tests to Data

2.1. The Sample of 2MASS

We study 416 galaxy systems with more than 7 members selected from the 2MASS catalog. We applied five well known normality tests to the velocity distributions of these systems to distinguish Gaussian and non-Gaussian clusters. Using controlled samples, we estimated type I and II errors for each test. We verified that individual tests minimize the chances of classifying a Gaussian system as non-Gaussian, while the Fisher's meta-analysis method, a procedure to combine p-values from several statistical tests, minimizes the chances of classifying a non-Gaussian system as Gaussian. Taking the positive elements of each method and also including a modality analysis of the velocity distribution, we defined objective criteria to split up the sample into Gaussian and non-Gaussian clusters. Our analysis indicates that 50-58% of groups have Gaussian distribution. We also found that some properties of galaxy clusters are significantly different between Gaussian and non-Gaussian systems. For instance, non-Gaussian clusters have larger radii and contain more galaxies than Gaussian clusters. Finally, we discussed the importance of choosing the adequate methodology to classify galaxy systems from their velocity distributions and also the dependence of the results on the criteria used to identify clusters in galaxy surveys.

In this paper, we study 416 galaxy systems selected from the 2MASS (Two Micron All Sky Survey Extended Source Catalog - Crook et al. (2007, Cat. J/ApJ/655/790)). We used just groups with more than 7 members to avoid severe sample size effects.

Also we represent below an image of data of members of groups of galaxies in 2MASS Survey.

Full	n	Group	RAJ2000 "h:m:s"	DEJ2000 "d:m:s"	HV km/s	Kmag mag	Dist Mpc
☐	L	1	00 04 33.6	+28 18 06	8785	10.62	118.79
☐	L	1	23 58 28.3	+28 02 03	9145	10.94	123.72
☐	L	1	23 59 43.6	+28 17 25	9073	10.71	122.74
☐	L	2	00 03 29.1	+27 21 06	7690	11.02	103.97
☐	L	2	00 05 48.3	+27 26 58	7531	10.95	101.81
☐	L	2	00 07 17.1	+27 40 42	7550	11.13	102.05
☐	L	3	00 04 57.8	+05 07 24	5357	11.19	72.17
☐	L	3	00 05 27.6	+05 13 20	5294	10.08	71.31
☐	L	3	00 06 40.3	+05 06 48	5371	11.22	72.34
☐	L	4	00 04 24.5	+31 28 19	4958	11.14	67.36
☐	L	4	00 07 19.5	+32 36 33	5074	9.95	68.93
☐	L	4	00 08 01.6	+33 04 15	4904	10.31	66.66

L	4	00 08 47.7	+33 26 00	4811	9.89	65.43
L	4	00 09 32.7	+33 18 31	4901	9.55	66.62
L	4	00 10 40.8	+32 58 59	4788	10.66	65.08
L	4	00 10 46.8	+33 21 10	4765	10.47	64.79
L	5	00 09 48.2	+27 49 56	8315	10.85	112.37
L	5	00 10 02.0	+28 12 36	8168	11.19	110.39
L	5	00 11 14.2	+28 54 23	8046	10.98	108.75
L	6	00 04 26.6	+47 29 25	5269	10.52	72.27

- The first column n describes which galaxy group is Low density contrast and High density contrast.
- The second column describes in which group belong the galaxies.
- The third column gives the right ascension in $hour:minurtes:seconds$
- The fourth column gives the declination in $hour:minurtes:seconds$
- The fifth column gives the Heliocentric velocity
- The sixth column gives the K-band magnitude
- The seventh column gives the Heliocentric Distance

We represent now some data of groups of galaxies

Group	ID	N200	sigma (km/s)	R200 (Mpc)	logM200 [Msun])	m12 (mag)	v200 (Mpc-3)
LDC	6	20	146.84	0.460	12.84	-0.10	2.936
LDC	12	18	300.75	0.939	13.77	-0.11	1.959
LDC	17	18	333.59	1.044	13.91	-0.29	1.822
LDC	18	8	59.47	0.186	11.66	-0.18	3.717
LDC	26	9	176.38	0.556	13.08	-2.74	2.340
LDC	31	8	71.36	0.225	11.90	-4.28	3.467
LDC	34	10	199.27	0.625	13.24	-0.01	2.235
LDC	35	11	128.24	0.400	12.66	0.00	2.859
LDC	36	8	160.51	0.502	12.96	-0.44	2.422
LDC	39	13	276.32	0.865	13.66	-0.01	1.925 …

We would like to quantify how many of these groups have more complex velocity distributions, potentially identifying merger products or systems in the early stages of virialization.

2.2. Estimation of Distribution Parameters μ and σ

The χ^2, Kolmogorov, and A–D statistics were developed under the assumption that all of the parameters of the underlying distribution were completely specified. Modifications to the statistics, with the use of Monte Carlo simulations, have been carried out to allow these tests to be applied to cases where the distribution parameters are not completely known a priori, but must be estimated from the data (DA86).The parameters required to define a Gaussian distribution are the mean μ, and the dispersion,σ. In the analysis described inSection3.3, μ is calculated using the standard mean, and the velocity dispersions are estimated with the Gapper algorithm, given by:

$$\sigma_{Gapper} = \frac{\sqrt{\pi}}{n(n-1)} \sum_{i=1}^{n} w_i g_i \quad , (12)$$

, where $w_i = i(n-i)$, $g_i = x_{i+1} - x_i$, here the ordered xi values are given by the observed radial velocities of the group members. For small number statistics, Beers et al. (1990) recommend the Gapper Estimator over the canonical rms standard deviation, as this algorithm is insensitive to outliers and thus more accurately reproduces the true dispersion of the system. To ensure that the Gapper Estimator is a more reliable method of computing the dispersion of a system, we perform Monte Carlo simulations of both the Gapper algorithm and the canonical rms standard deviation. We draw various sample sizes (n=5, 15, 20, 50) from a Gaussian random number generator (Galassi et al.2006) with the inputs μ=0.0 and σintrinsic=100 and then compute the dispersion using the two a for mentioned methods. The results are shown in Figure40, where we have plotted histograms of the output dispersions for each sample size. Figure40 shows that for then=5 case the canonical rms standard deviation underestimates the true dispersion by 25% and the distribution is heavily skewed to lower values. Although the distribution of the Gapper Estimator is also skewed, the peak of the distribution occurs at the true dispersion value of 100, indicating that this method is indeed insensitive to outliers. As we increase the sample size, n=15 and 20, we can see that the rms dispersion continues to underestimate the true velocity dispersion, but also that the two methods begin to converge. The histogram for then=50 case shows that the rms dispersion and the Gapper Estimator both correctly identify the true dispersion. Given the large n of the 2MASS groups we choose the rms algorithm to estimate the velocity dispersions of the groups.

We first show in Figure40 the Gapper Estimator plotted histograms for n = 5, 15, 20, 50. In Figure41 we show the RMS estimator plotted histograms for n = 5, 15, 20, 50. In Figure42 we show both of them because we want to do the comparisons of these estimators.

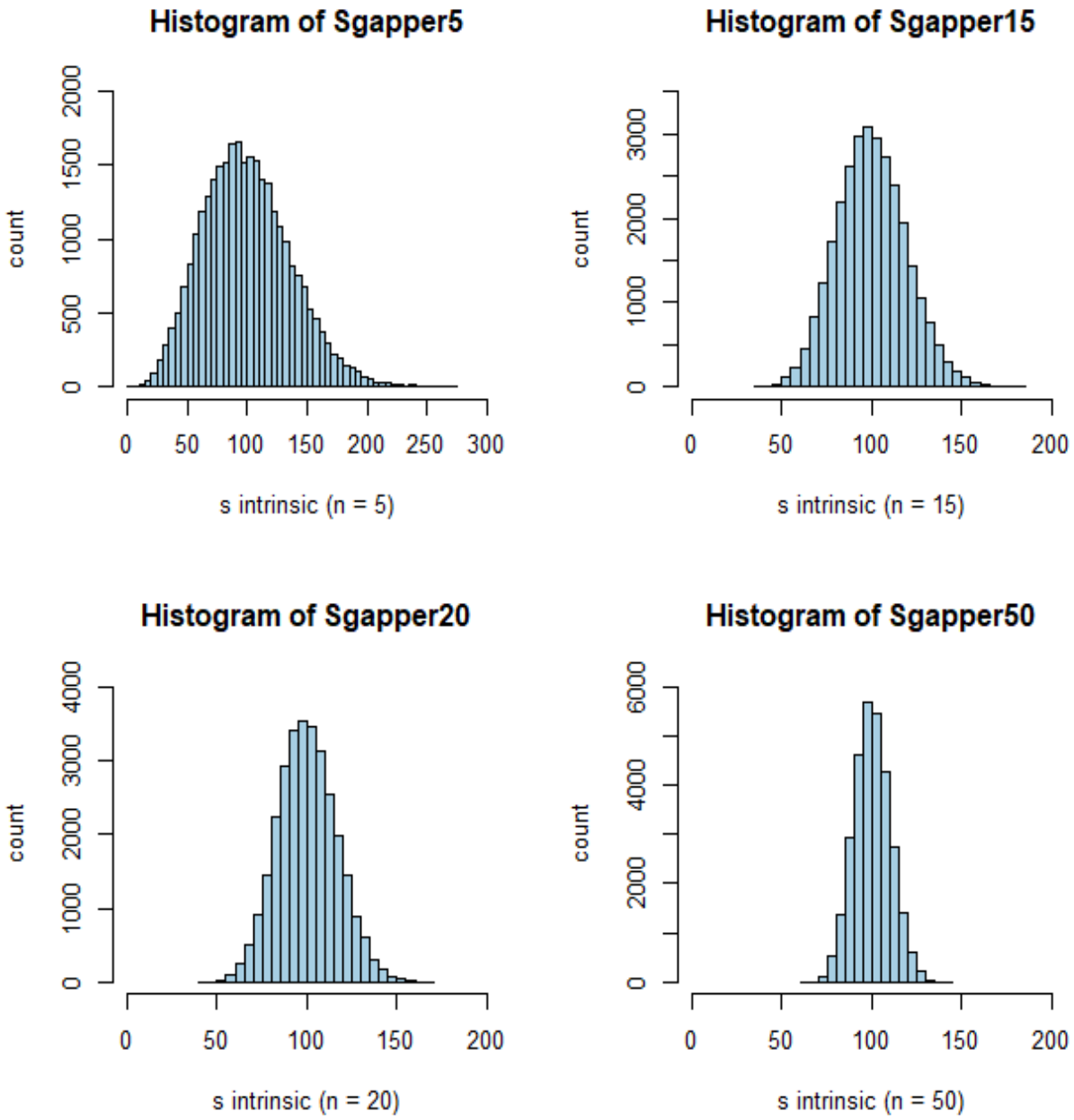


Figure 50 Monte Carlo simulations of the Gapper Estimator method for n=5 (top left),n=15 (top right),n=20 (bottom left), and n=50 (bottom right). For each histogram, we compute the Gapper dispersion using velocities generated from a random Gaussian distribution with input values of μ=0.0 and σ_{int}= 100. This process was done with 30,000 iterations.

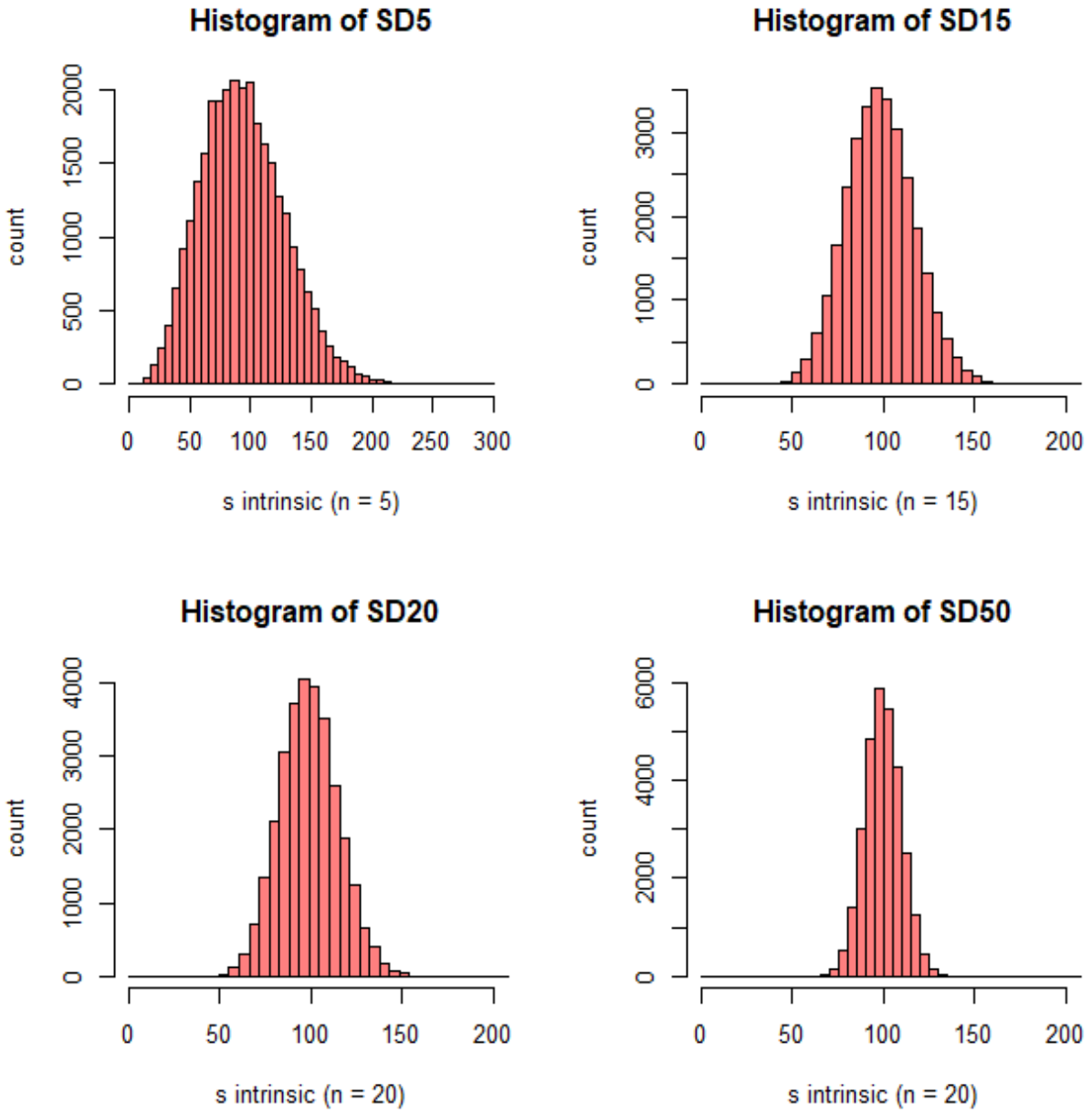


Figure 51 Monte Carlo simulations of the RMS Estimator method for n=5 (top left), n=15 (top right), n=20 (bottom left), and n=50 (bottom right). For each histogram, we compute the RMS dispersion using velocities generated from a random Gaussian distribution with input values of μ=0.0 and σ_{int}= 100. This process was done with 30,000 iterations.

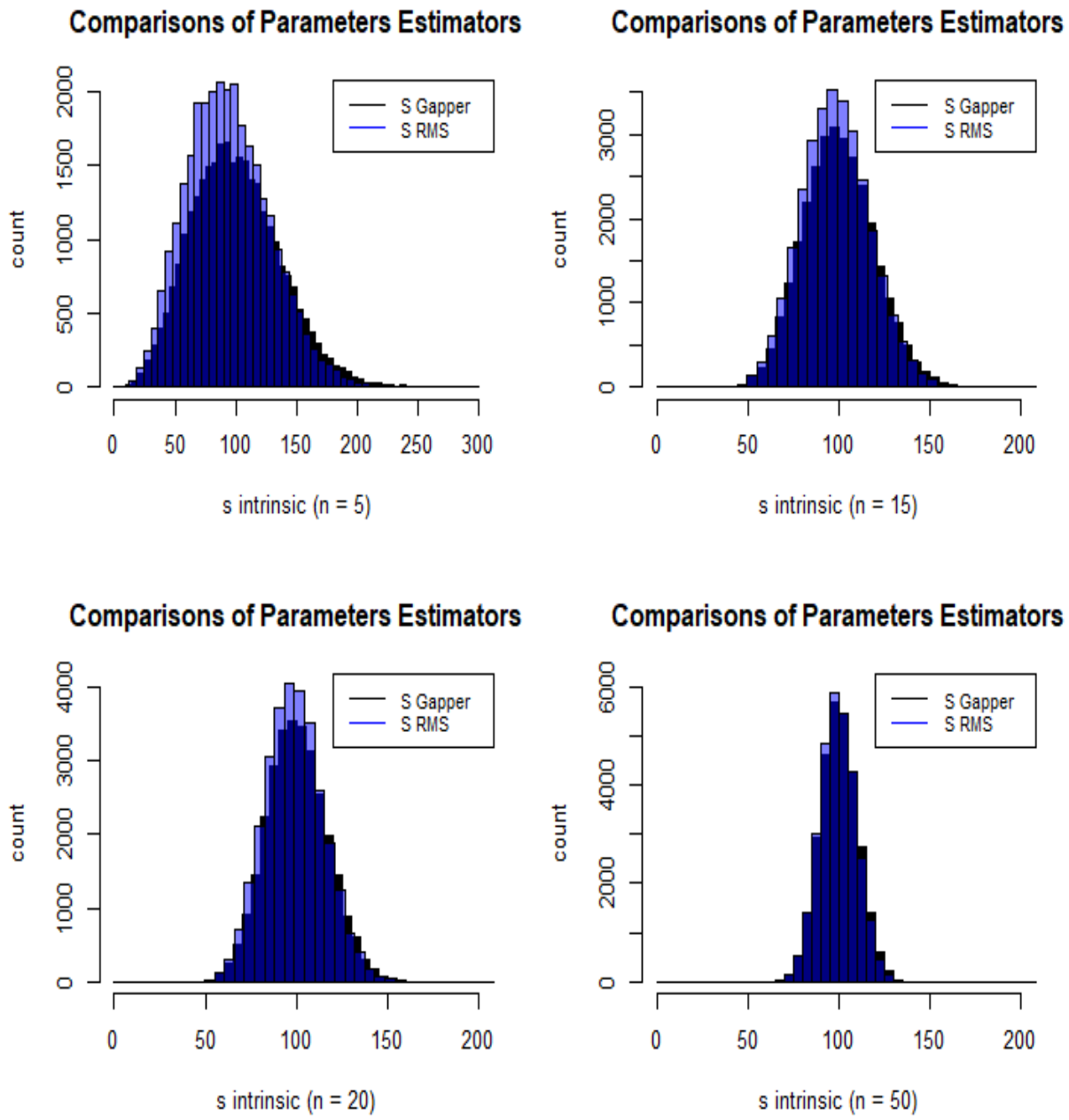


Figure 52 The blue histograms indicate rms dispersion values and the black histogram are values calculated using the Gapper algorithm. For n=5, the histogram for the rms method is skewed and underestimates the true dispersion by 20%. For the same sample size, the peak of the Gapper Estimator histogram is located at the true dispersion value of 100, even though the histogram is also skewed, indicating that the Gapper Estimator is insensitive to outliers. For then=50 case, the two methods begin to converge, with both the rms and Gapper algorithms picking out the true dispersion value.

2.3 Comparison of Tests

2.3.1. Statistical Analysis

We start the statistical analysis in the 2MASS sample.

Firstly we represent histograms of number of galaxies in groups.

In Figure43 we represent the histogram of $N_{members} \leq 50$

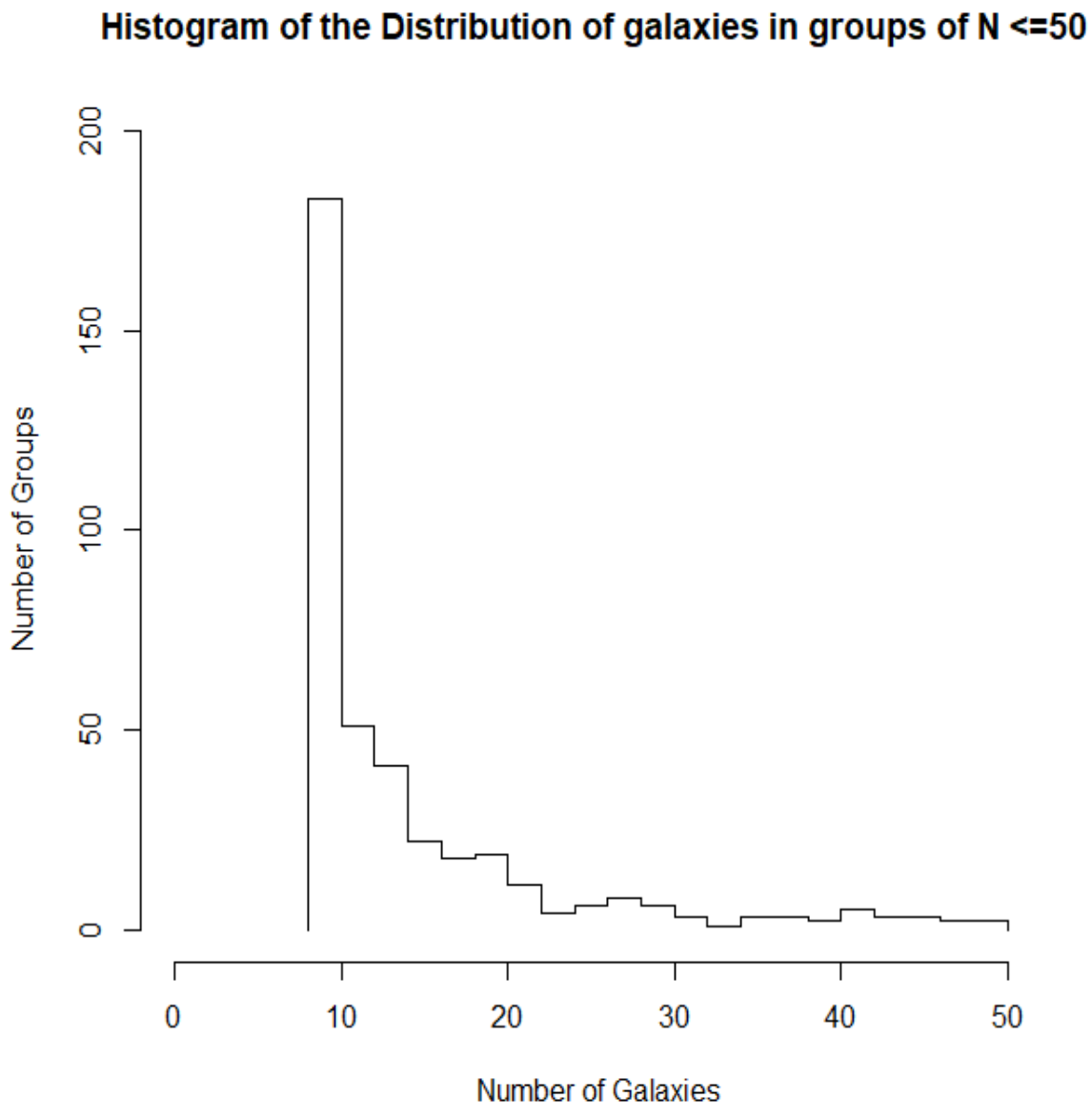


Figure 53 histogram of $N_{members} \leq 50$. We see that there many groups that include $N_{members} = 10, 11, 12$.

Also we distinguish the groups between the Low contrast (274 GROUPS) and the High contrast (142 GROUPS) and we start the normality tests for the velocity of the groups. We transform the velocities into standard values of $N(0,1)$.

We test the normality of velocity with five (5) methods for the Low contrast groups

- D' Agostino Normality Test

We found 34 Non - Normal groups and 240 Normal groups for $a = 0.10$.

Now, we will represent 4 histograms of velocities from normal and non-normal groups.

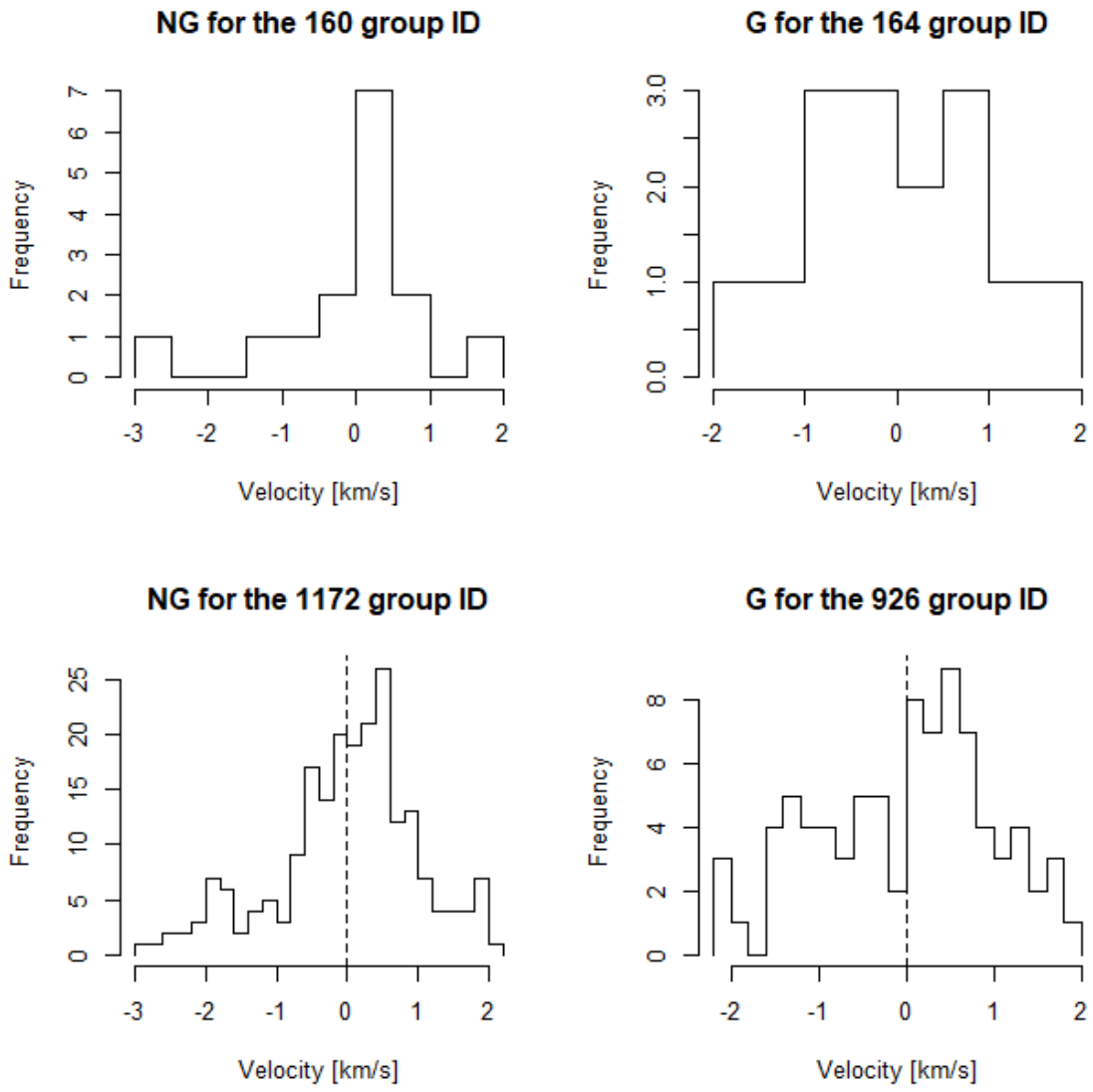


Figure 54 NG means Non-Gaussian Group and G means Gaussian group

- Shapiro – Wilk test for Normality

We found 39 Non - Normal groups and 235 Normal groups for $a = 0.10$.

Now, we will represent 4 histograms of velocities from normal and non-normal groups.

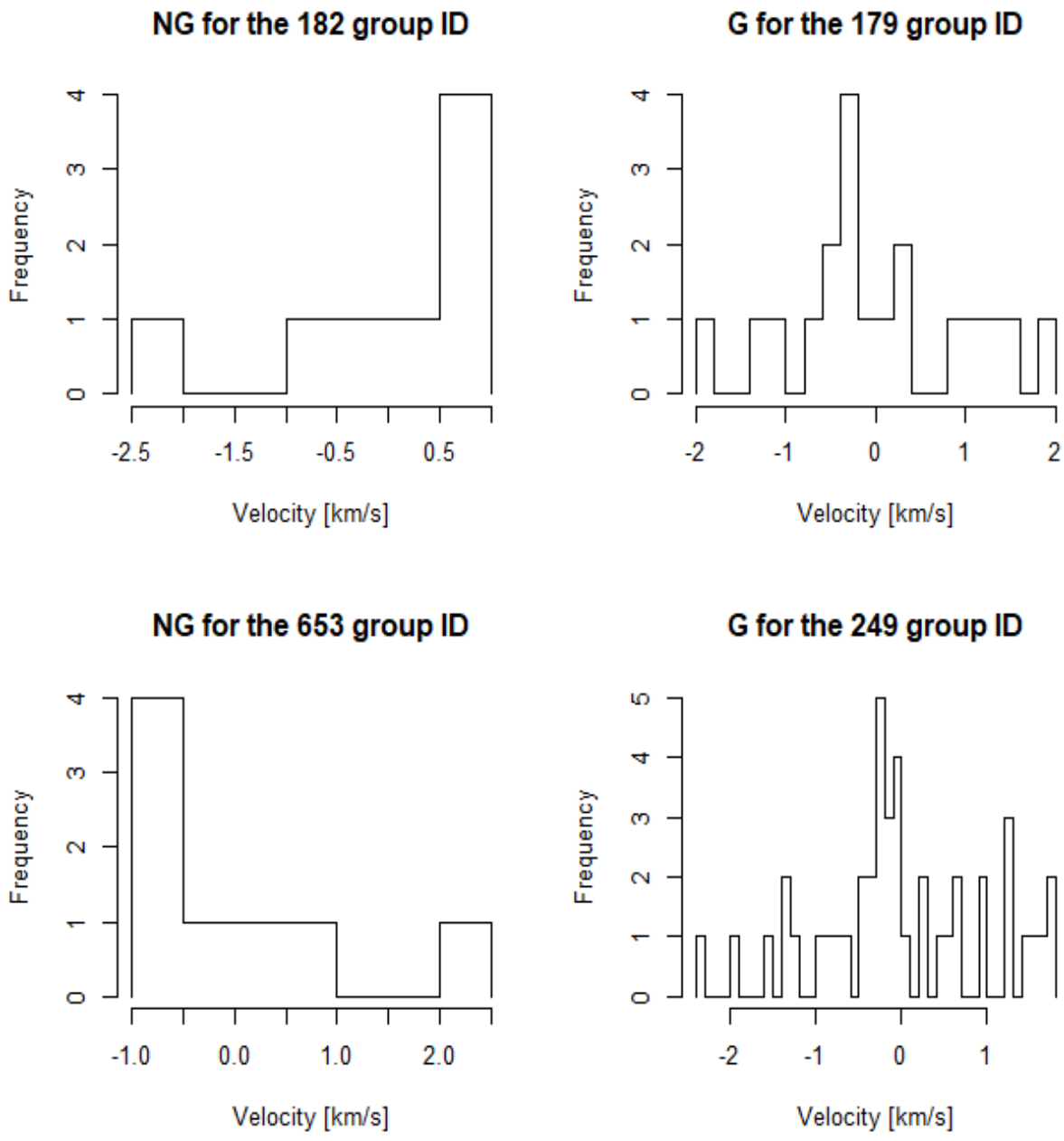


Figure 55 NG means Non-Gaussian Group and G means Gaussian group

- Anderson – Darling test for Normality

We found 42 Non - Normal groups and 232 Normal groups for $a = 0.10$.

Now, we will represent 4 histograms of velocities from normal and non-normal groups.

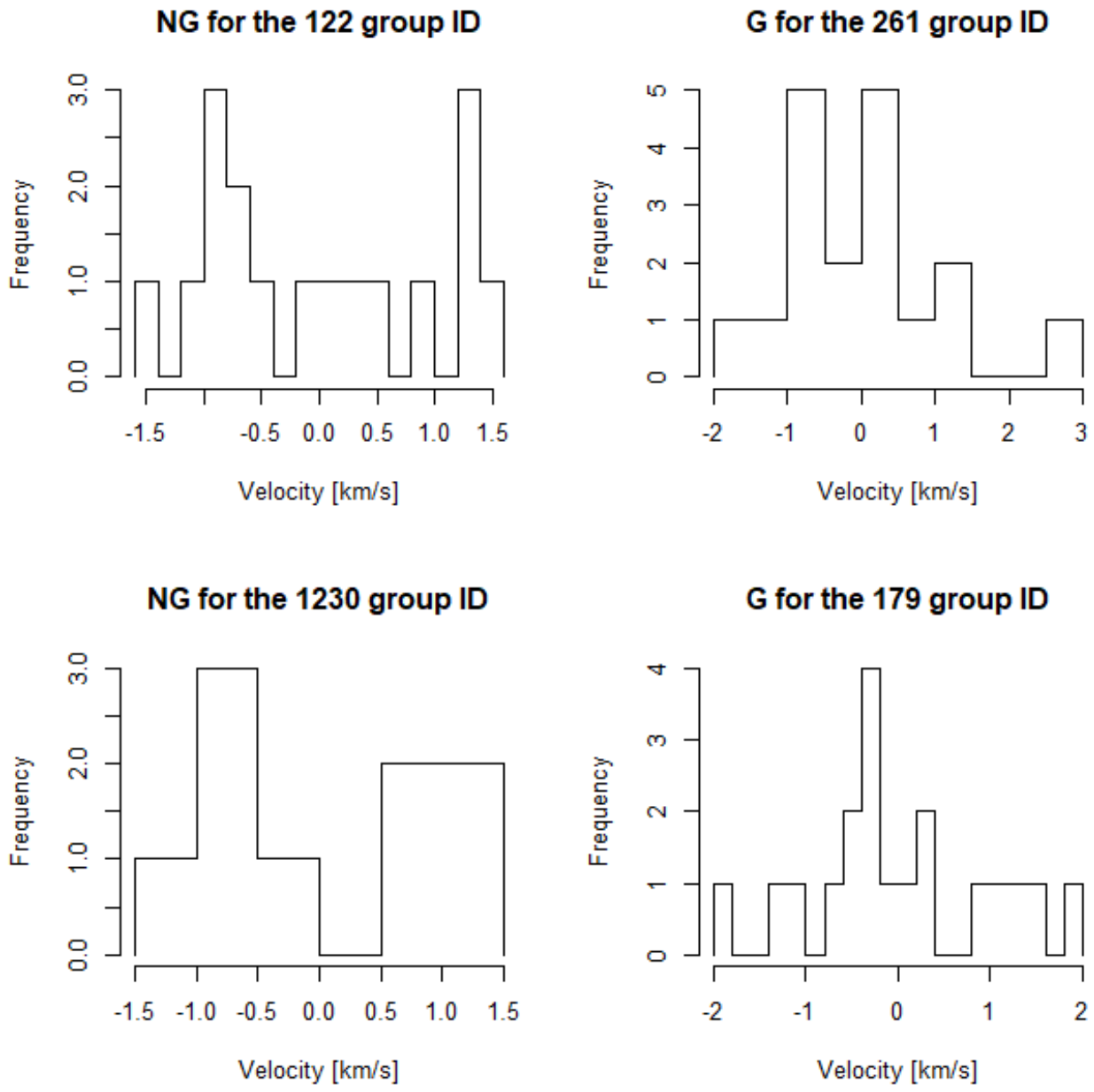


Figure 56 NG means Non-Gaussian Group and G means Gaussian group

- Lilliefors test for Normality

We found 38 Non - Normal groups and 236 Normal groups for $a = 0.10$.

Now, we will represent 4 histograms of velocities from normal and non-normal groups.

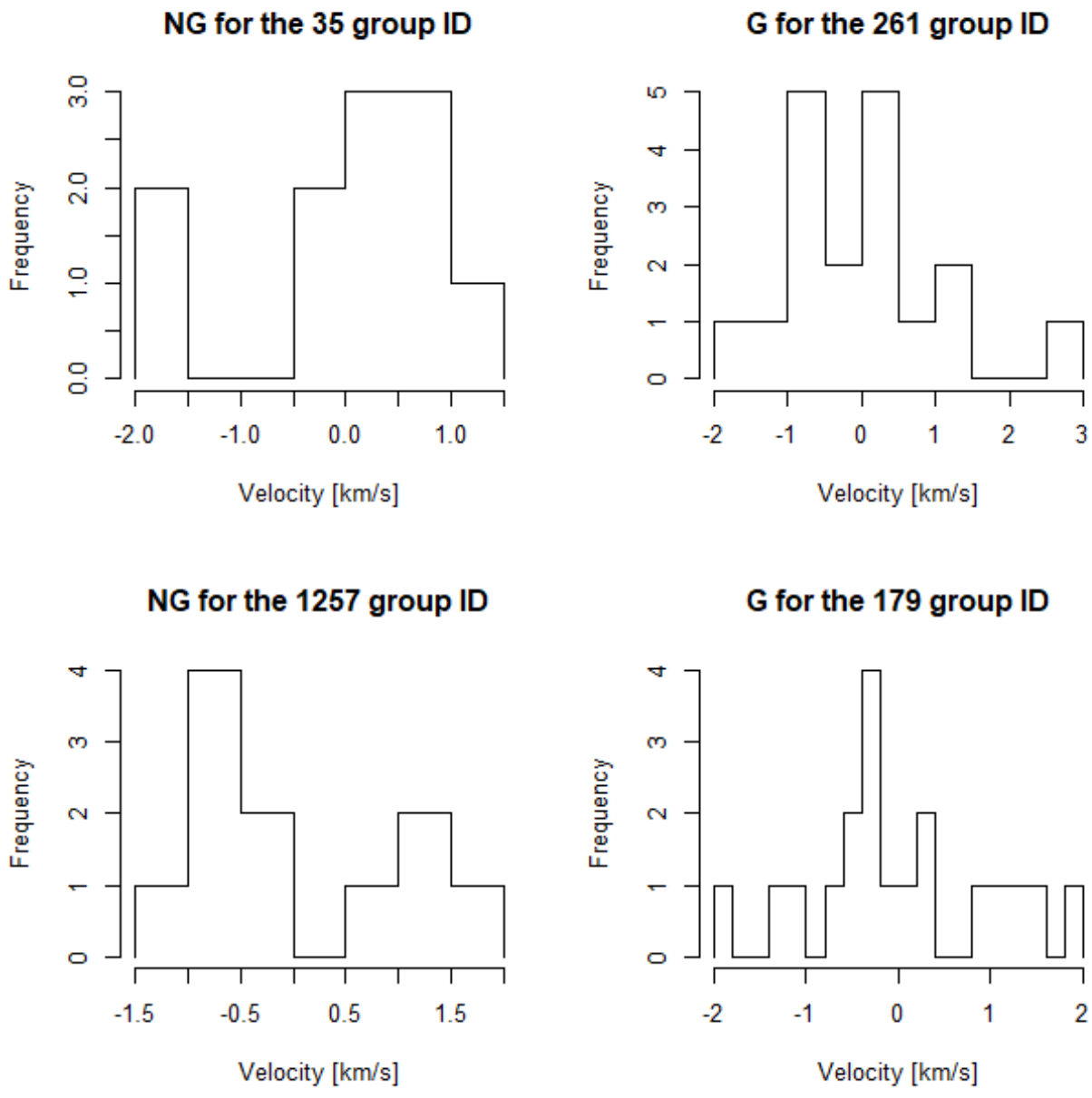


Figure 57 NG means Non-Gaussian Group and G means Gaussian group

- Pearson's test for Normality

We found 49 Non - Normal groups and 225 Normal groups for $a = 0.10$.

Now, we will represent 4 histograms of velocities from normal and non-normal groups.

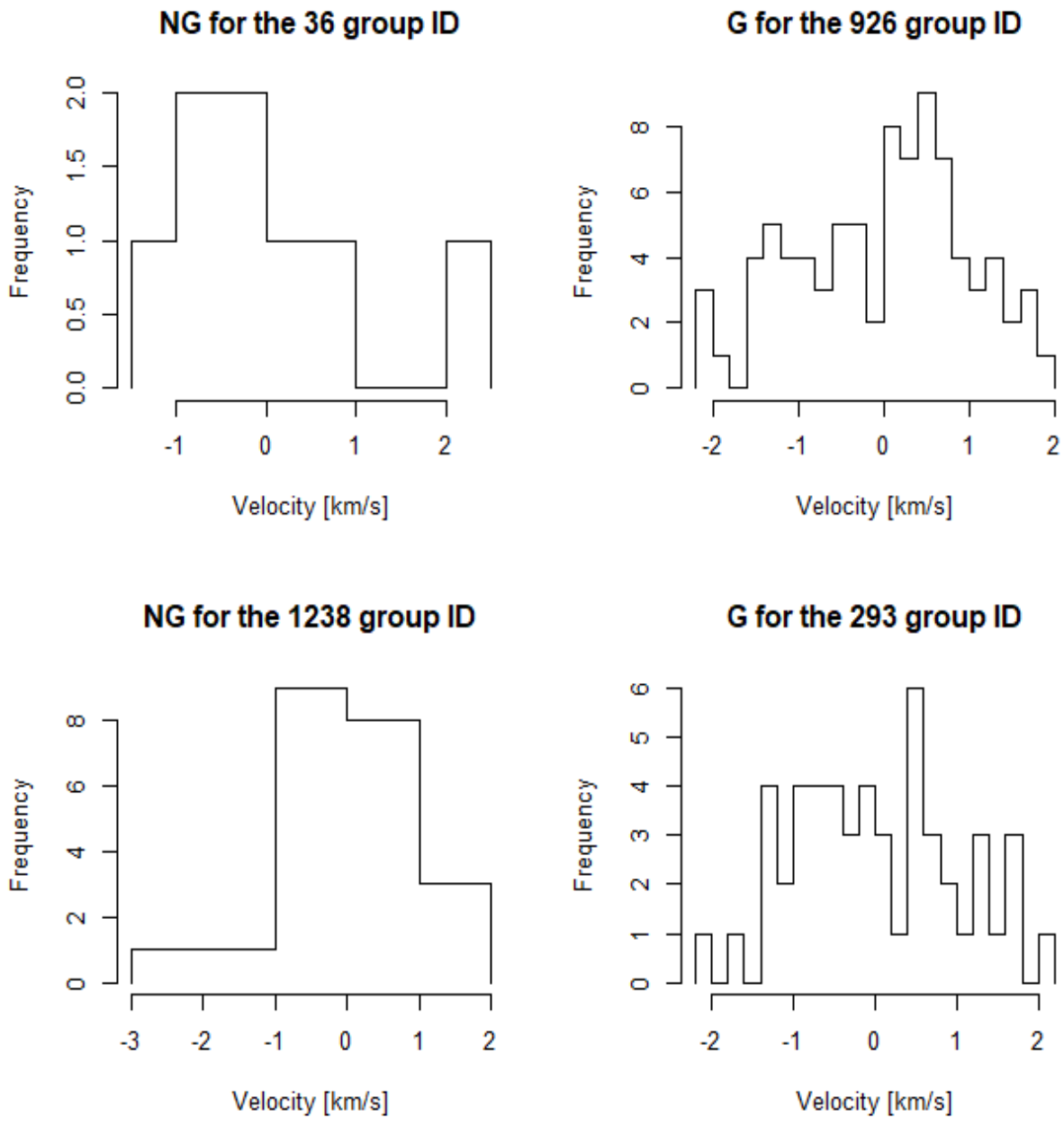


Figure 58 NG means Non-Gaussian Group and G means Gaussian group

Tests for Normality of Heliocentric Velocities				
/i-th Low Contrast Non - Gaussian Groups				
D'Agostino	Shapiro - Wilk	Anderson -Darling	Lilliefors	Pearson
[1,] 1	[1,] 8	[1,] 8	[1,] 8	[1,] 6
[2,] 3	[2,] 15	[2,] 15	[2,] 15	[2,] 9
[3,] 9	[3,] 17	[3,] 17	[3,] 17	[3,] 15
[4,] 14	[4,] 23	[4,] 23	[4,] 28	[4,] 17
[5,] 17	[5,] 34	[5,] 28	[5,] 34	[5,] 19
[6,] 23	[6,] 38	[6,] 31	[6,] 38	[6,] 25
[7,] 28	[7,] 47	[7,] 34	[7,] 57	[7,] 31
[8,] 34	[8,] 50	[8,] 38	[8,] 59	[8,] 34
[9,] 38	[9,] 51	[9,] 44	[9,] 64	[9,] 38
[10,] 47	[10,] 56	[10,] 51	[10,] 67	[10,] 44
[11,] 64	[11,] 59	[11,] 56	[11,] 72	[11,] 45
[12,] 77	[12,] 63	[12,] 57	[12,] 80	[12,] 49
[13,] 85	[13,] 64	[13,] 59	[13,] 85	[13,] 57
[14,] 95	[14,] 66	[14,] 64	[14,] 95	[14,] 59
[15,] 106	[15,] 72	[15,] 72	[15,] 106	[15,] 60
[16,] 128	[16,] 77	[16,] 77	[16,] 112	[16,] 69
[17,] 135	[17,] 106	[17,] 85	[17,] 135	[17,] 72
[18,] 138	[18,] 112	[18,] 98	[18,] 146	[18,] 77
[19,] 146	[19,] 128	[19,] 106	[19,] 149	[19,] 80
[20,] 147	[20,] 135	[20,] 112	[20,] 156	[20,] 82
[21,] 149	[21,] 146	[21,] 128	[21,] 160	[21,] 85
[22,] 159	[22,] 147	[22,] 135	[22,] 186	[22,] 89
[23,] 162	[23,] 149	[23,] 146	[23,] 189	[23,] 100
[24,] 178	[24,] 159	[24,] 149	[24,] 190	[24,] 106
[25,] 189	[25,] 160	[25,] 156	[25,] 194	[25,] 108
[26,] 190	[26,] 162	[26,] 160	[26,] 196	[26,] 113
[27,] 197	[27,] 178	[27,] 162	[27,] 197	[27,] 128
[28,] 204	[28,] 189	[28,] 189	[28,] 200	[28,] 135
[29,] 206	[29,] 190	[29,] 190	[29,] 204	[29,] 149
[30,] 219	[30,] 193	[30,] 193	[30,] 206	[30,] 151
[31,] 253	[31,] 194	[31,] 194	[31,] 219	[31,] 156
[32,] 256	[32,] 196	[32,] 196	[32,] 221	[32,] 157
[33,] 257	[33,] 197	[33,] 197	[33,] 223	[33,] 160
[34,] 264	[34,] 204	[34,] 204	[34,] 239	[34,] 174
[35,] NA	[35,] 206	[35,] 206	[35,] 241	[35,] 189
[36,] NA	[36,] 219	[36,] 219	[36,] 257	[36,] 190
[37,] NA	[37,] 241	[37,] 241	[37,] 270	[37,] 193
[38,] NA	[38,] 256	[38,] 256	[38,] 273	[38,] 194
[39,] NA	[39,] 257	[39,] 257	[39,] NA	[39,] 196
[40,] NA	[40,] NA	[40,] 260	[40,] NA	[40,] 197
[41,] NA	[41,] NA	[41,] 265	[41,] NA	[41,] 202
[42,] NA	[42,] NA	[42,] 270	[42,] NA	[42,] 206
[43,] NA	[43,] NA	[43,] NA	[43,] NA	[43,] 219
[44,] NA	[44,] NA	[44,] NA	[44,] NA	[44,] 221
[45,] NA	[45,] NA	[45,] NA	[45,] NA	[45,] 239
[46,] NA	[46,] NA	[46,] NA	[46,] NA	[46,] 241
[47,] NA	[47,] NA	[47,] NA	[47,] NA	[47,] 242
[48,] NA	[48,] NA	[48,] NA	[48,] NA	[48,] 257
[49,] NA	[49,] NA	[49,] NA	[49,] NA	[49,] 267

The above table show us which Low contrast groups are Non-Gaussian for each test we have performed. We see that the Pearson's test for normality is stricter (49 NG groups) than the others because it rejects more groups. Anderson-Darling test (42 NG groups) follows. The Shapiro-Wilk test takes the third place (39 NG groups). The Lilliefors takes the fourth place (38 NG groups) and the D'Agostino the last place (34 NG groups). Furthermore, in Figure49, 50, 51, 52 we will represent some comparisons of the tests. We have plotted histograms of velocities in order to see which of them are rejected or not for Gaussianity.

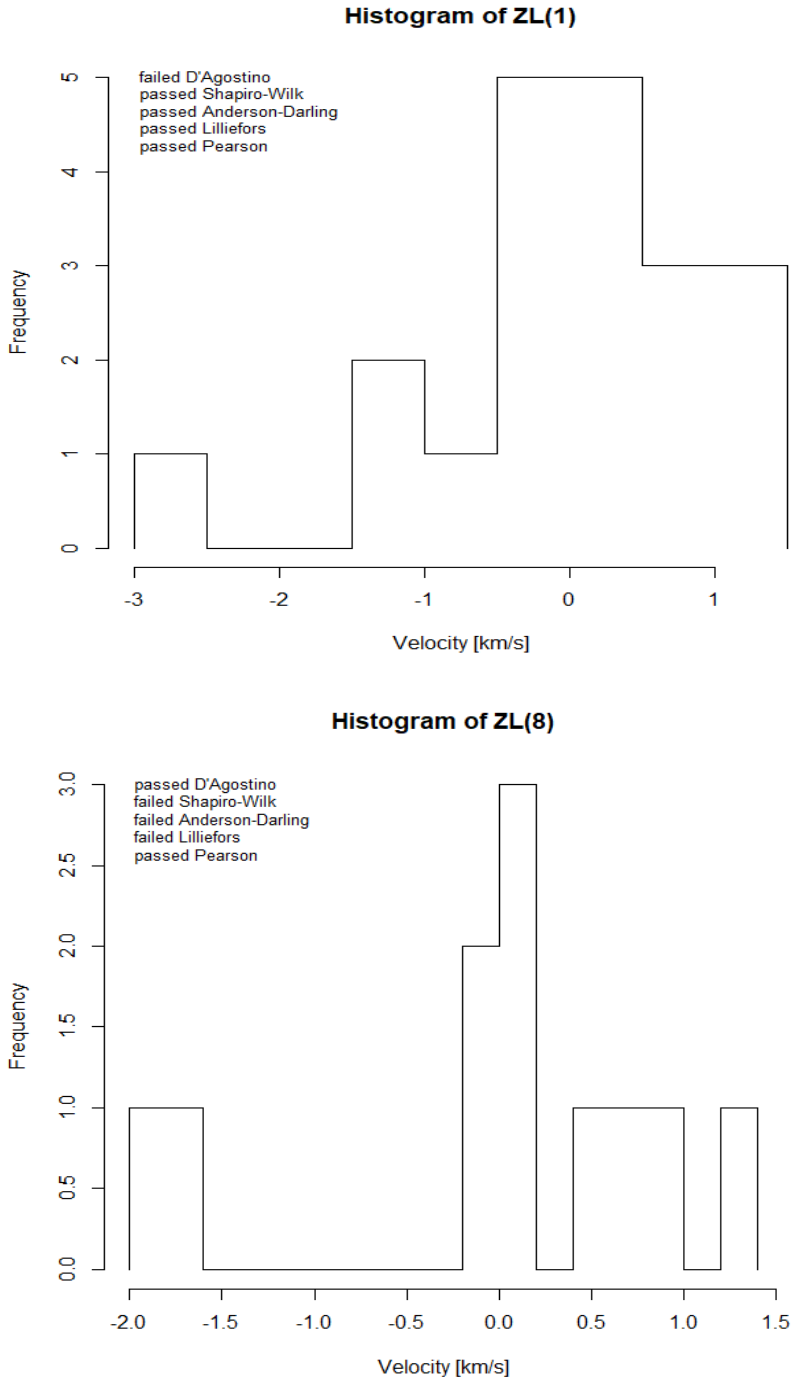


Figure 59

Figure 60

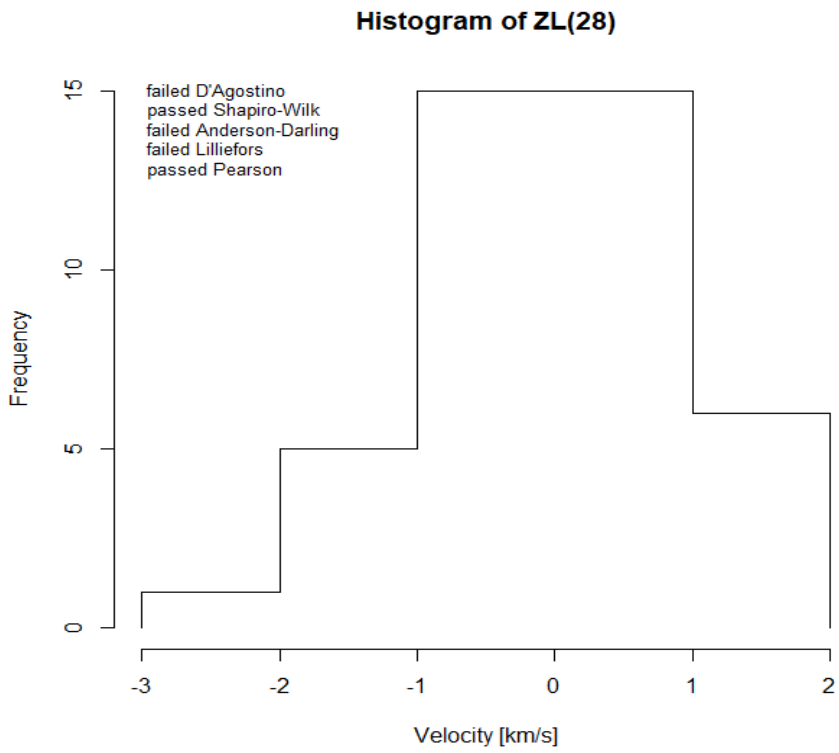


Figure 61

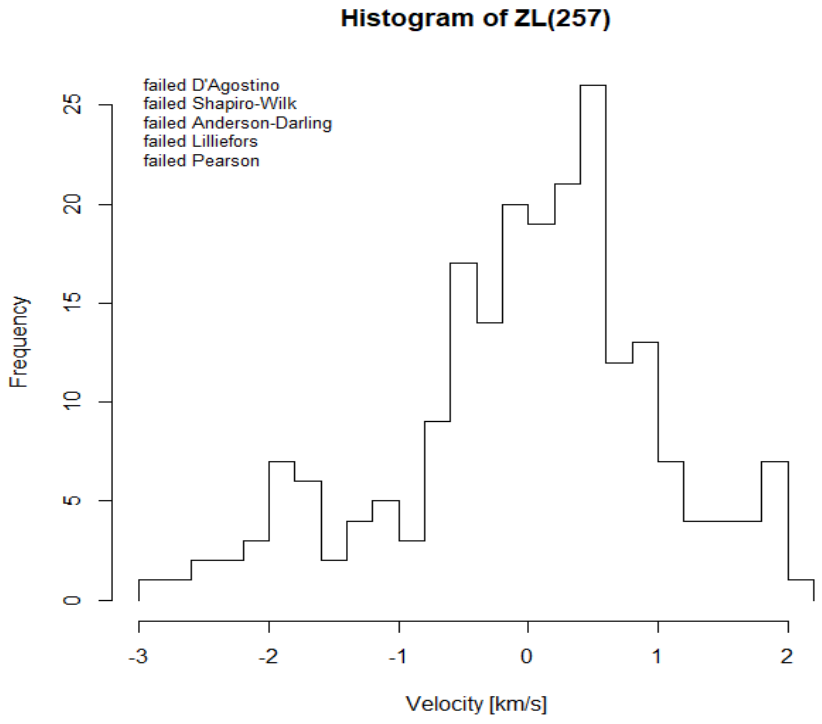


Figure 62

We want to test the Figure52. The group $1172\ (257)$ is non - Gaussian but we want to test more, because optically it seems like a normal density. In Figure53 we represent the Figure52 with the empirical curve and the theoretical curve of $Normal(0,1)$

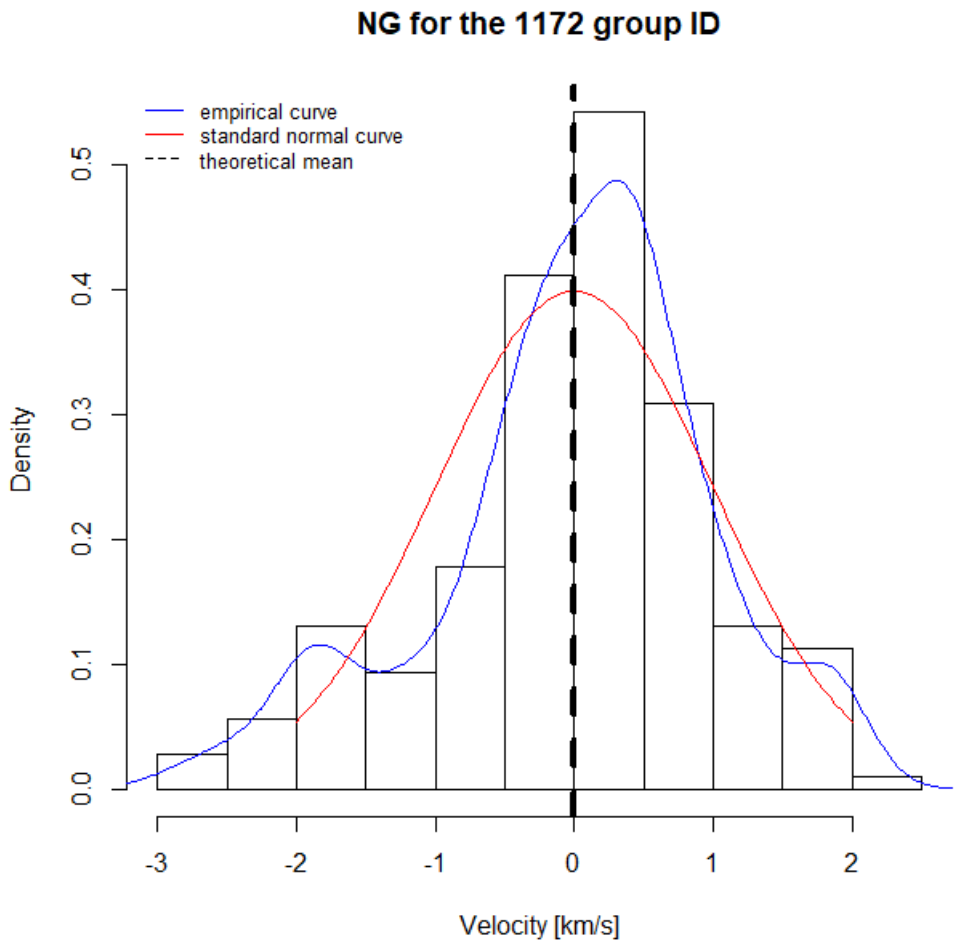


Figure 63

Now it is clear that the group 1172 is non-Gaussian, because the empirical curve is different from the theoretical curve. The dotted vertical line is the mean of the theoretical distribution. So, we see that the peak of the blue is not corresponding to the vertical line.

Now we will do the same for others Non-Gaussian groups, in order to understand the normality of the groups using the theoretical and empirical curves.

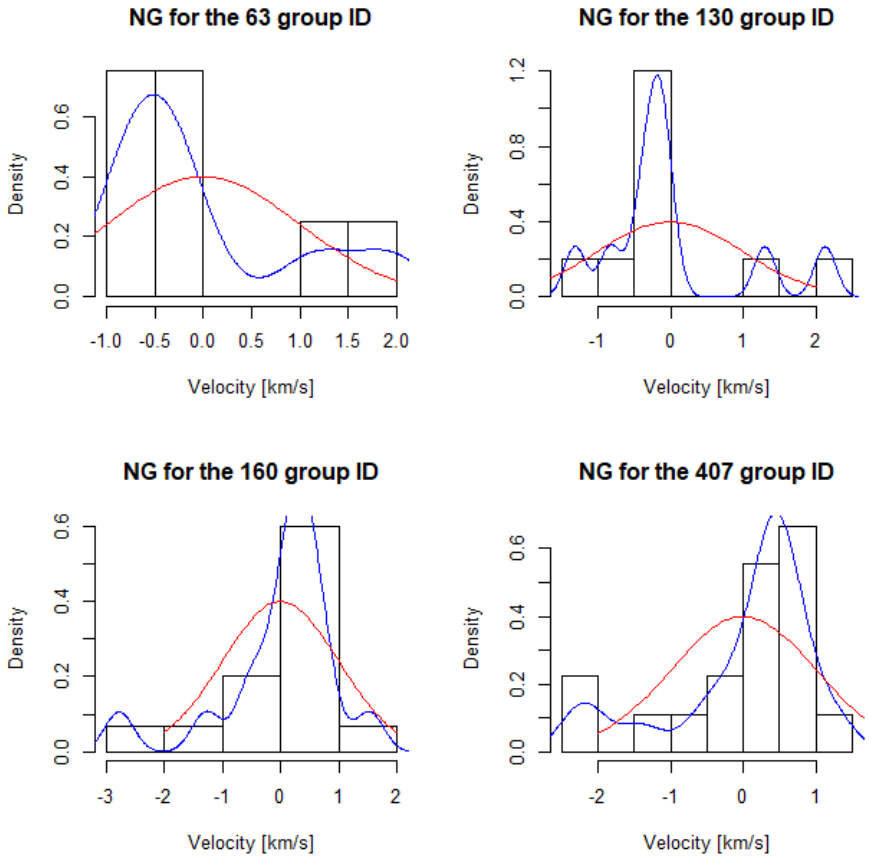


Figure 64

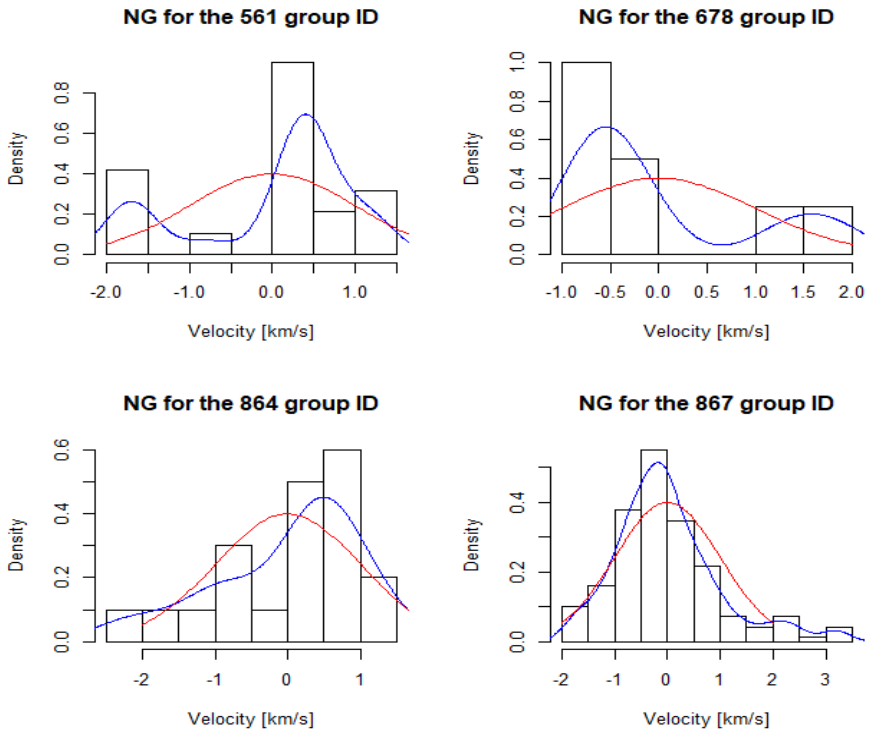


Figure 65

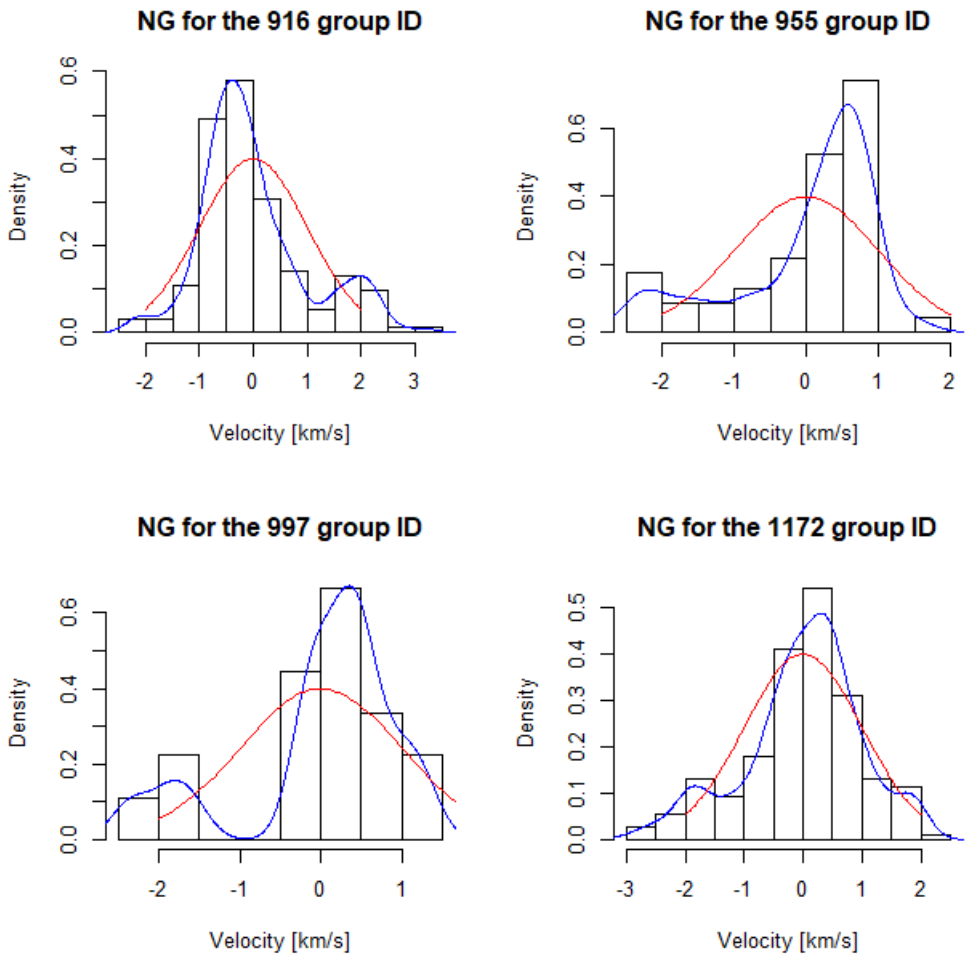


Figure 66

We notice that there are many declinations of the empirical and theoretical curves. For that reason these groups are non-normal. It is clear that the histograms are very helpful for identifying Gaussian and Non-Gaussian groups.

Velocity Dispersion Histograms

Now, we represent velocity dispersion and number of members per galaxy group histograms for Gaussian and Non-Gaussian groups for each test.

- D'Agostino Normality Test

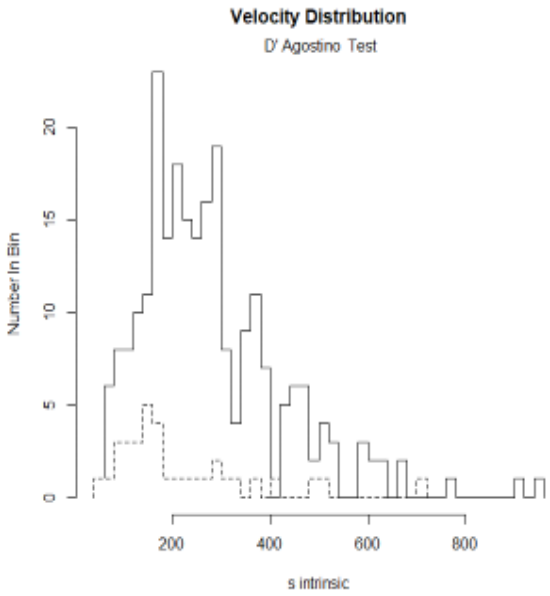


Figure 67 The dotted line indicate the non-normal groups and the solid line the normal groups.

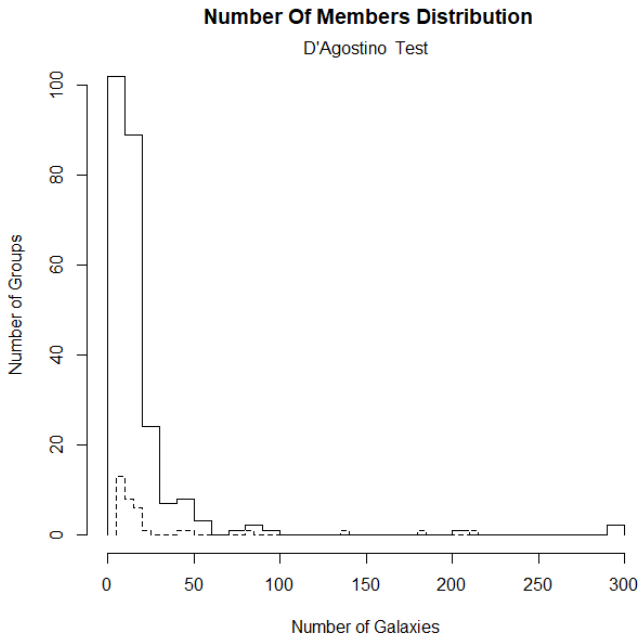


Figure 68 The dotted line indicate the non-normal groups and the solid line the normal groups.

- Shapiro-Wilk Normality Test

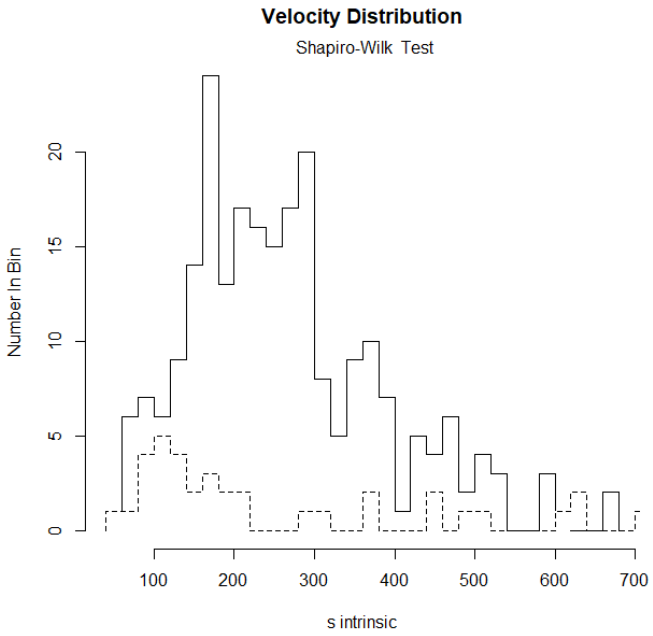


Figure 69 The dotted line indicate the non-normal groups and the solid line the normal groups.

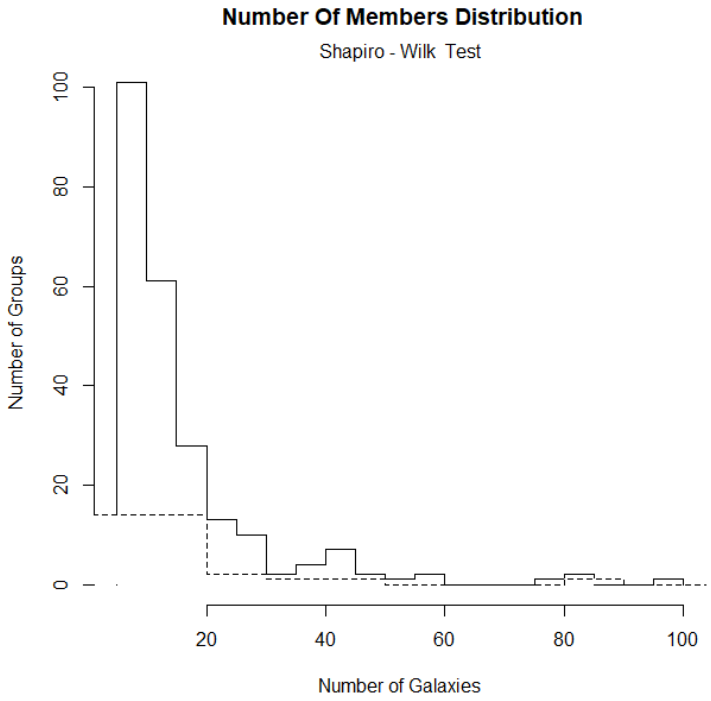


Figure 70 The dotted line indicate the non-normal groups and the solid line the normal groups.

- Anderson-Darling Normality Test

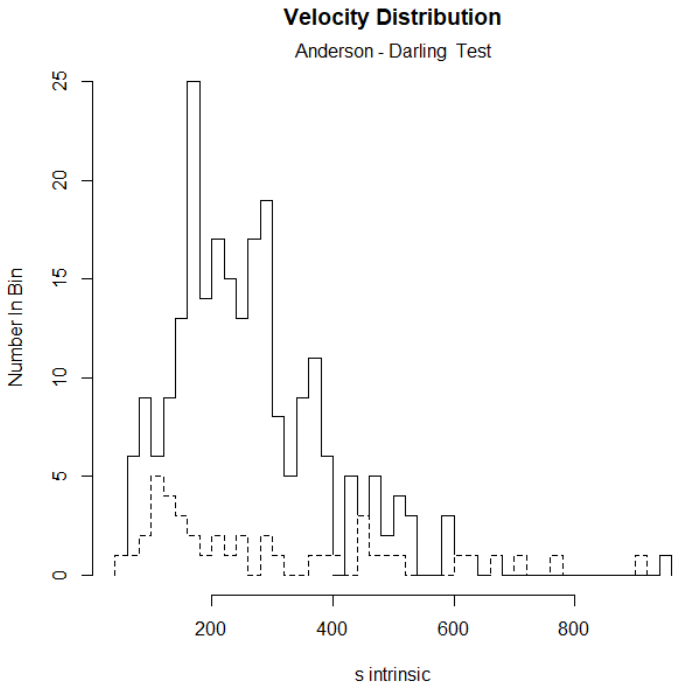


Figure 71 The dotted line indicate the non-normal groups and the solid line the normal groups.

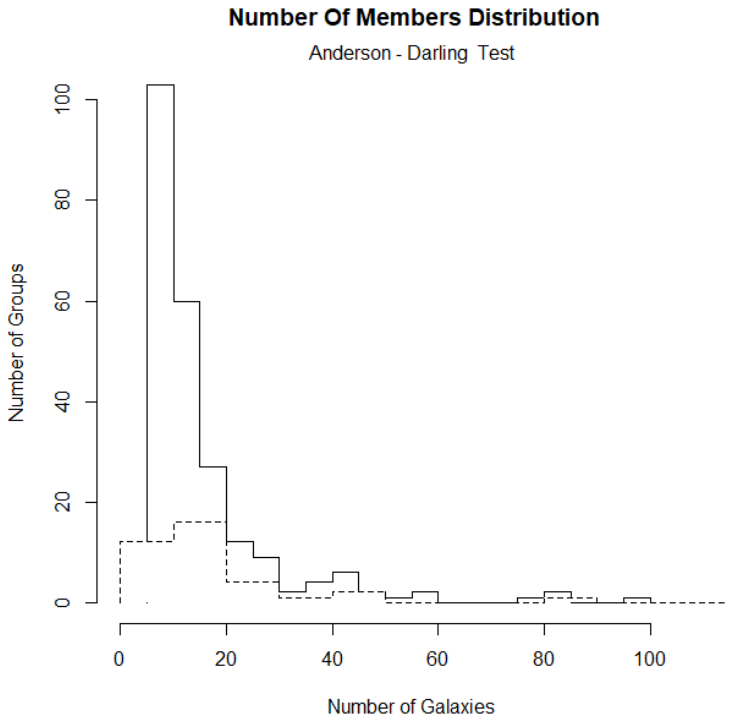


Figure 72 The dotted line indicate the non-normal groups and the solid line the normal groups

- Lilliefors Normality Test

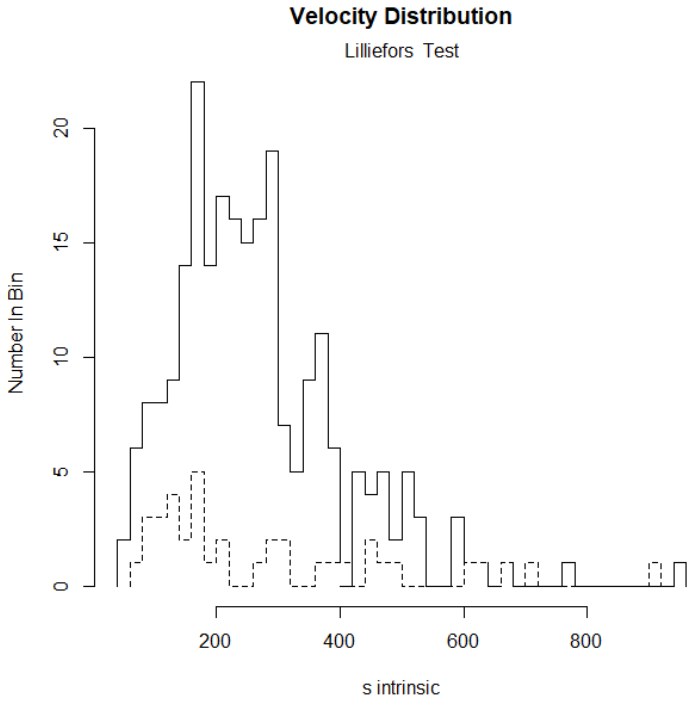


Figure 73 The dotted line indicate the non-normal groups and the solid line the normal groups.

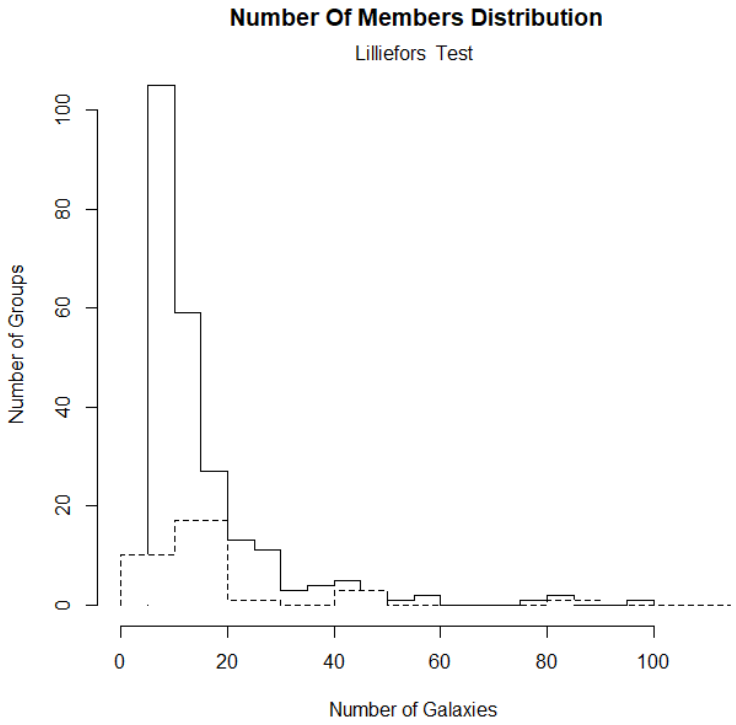


Figure 74 The dotted line indicate the non-normal groups and the solid line the normal groups.

- Pearson's Normality Test

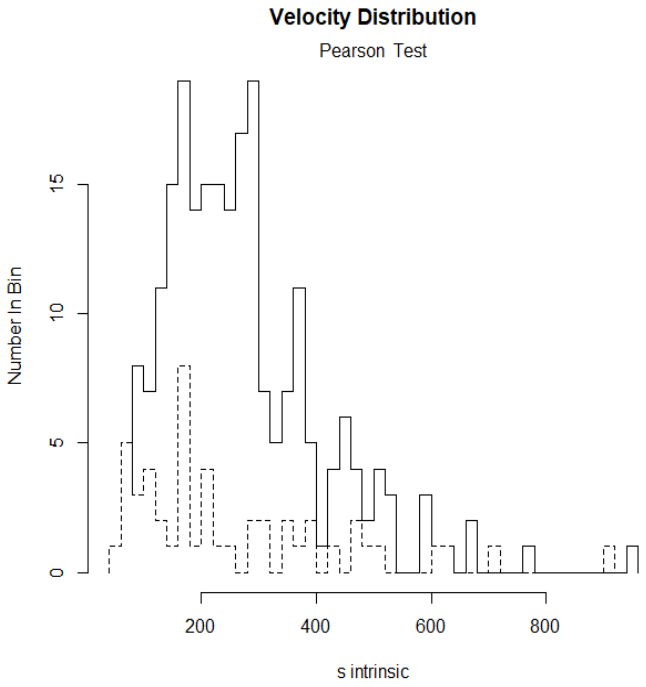


Figure 75 The dotted line indicate the non-normal groups and the solid line the normal groups.

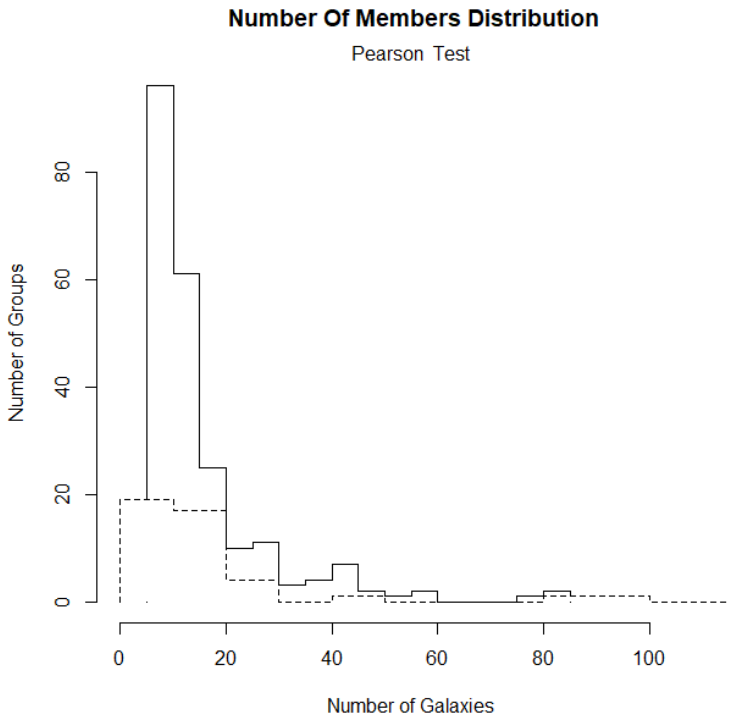


Figure 76 The dotted line indicate the non-normal groups and the solid line the normal groups.

We represent the results of the five tests for the Low Contrast Groups in a table called Table2.L

Normality Tests	Number Of Groups Member	Number Of Failed Groups	Percent Of Failed Groups	Significance Level (α)
D'Agostino	274	34	0.1240876	0,1
Shapiro-Wilk	274	39	0.1423358	0,1
Anderson-Darling	274	42	0.1532847	0,1
Lilliefors	274	38	0.1386861	0,1
Pearson's	274	49	0.1788321	0,1

Next we will show two (2) tables that informs us about the non-normal and normal groups. Particularly it shows their number of galaxies of Gaussian and Non-Gaussian groups and their virial radius for both of them.

Table4.1.L

Normalit Tests	Number Of Galaxies for Non-Gaussian Groups	Number Of Galaxies for Gaussian Groups
D'Agostino	31	20
Shapiro-Wilk	48,41026	16
Anderson-Darling	47,47619	16,18103
Lilliefors	51	16,22034
Pearson's	42,61224	16,26667

Table4.2.L

Normalit Tests	Virial Radius Of Galaxies for Non-Gaussian Groups	Virial Radius Of Galaxies for Gaussian Groups
D'Agostino	0.6827353	0.8645583
Shapiro-Wilk	0.9313590	0.8271660
Anderson-Darling	0.9471429	0.8229612
Lilliefors	0.9079211	0.8313814
Pearson's	0.8116735	0.8486000

We conclude that the non-Gaussian groups have more members and bigger radius than the Gaussian. So, the Non-Gaussian groups are more complex than the Gaussian.

We represent in this paper the mass distribution of the low contrast groups. Firstly, in Figure67 we will show the mass distribution of the groups with $N_{members} \leq 20$ and $N_{members} > 20$, because we want to test what types of Groups have high mass and what type have low mass.

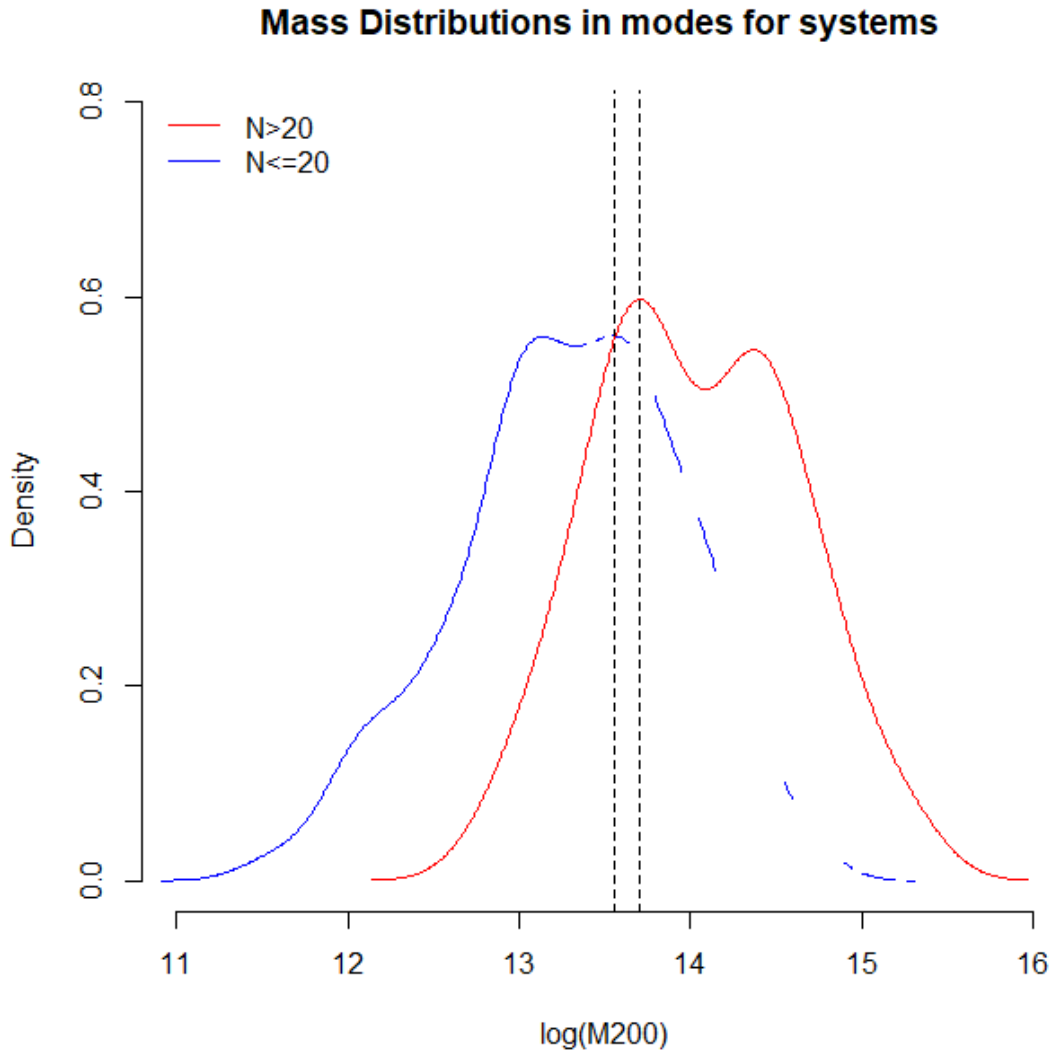


Figure 77 We see that the peaks of the two curves are too close, so the groups with N>20 have the same mass with the groups with N<=20. We understand that there are many groups with small N and big mass and many groups with large N and small mass. We can imagine a group that contains new galaxies and a group that contains old galaxies. (The dotted lines indicate the peak of each curve.)

Next, we do comparisons of the Mass Distributions between Gaussian and Non-Gaussian groups. We do five (5) comparisons because we used five (5) normality tests, in order to comprehend if the tests we did for the velocities correspond to the mass distributions .We understand that the Non-Gaussian groups have masses significantly higher than Gaussian groups.

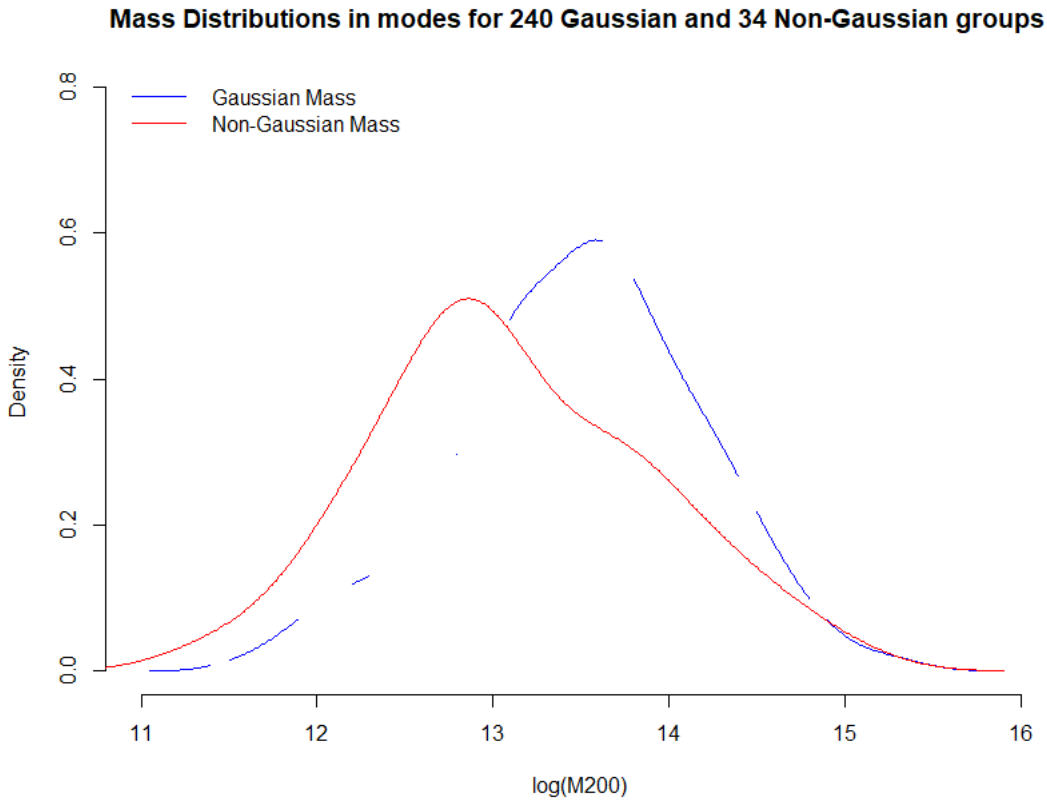


Figure 78 Density probability comparisons among systems (Gaussian and Non-Gaussian)

We observe from Figure68 that the masses of Non-Gaussian groups are higher than the Gaussian groups, even if the number of non-Gaussian is lower than the number of Gaussian (240G-34NG).

Also we calculate their Median

$Median_{Gaussian} = 13.53$, Gaussians cover 87.60% of all groups

$Median_{Non-Gaussian} = 12.9$, Non-Gaussians cover 12.40% of all groups

Also we demonstrate comparison of the mass distributions according to the different dynamical stages of the groups.

- D'Agostino Test

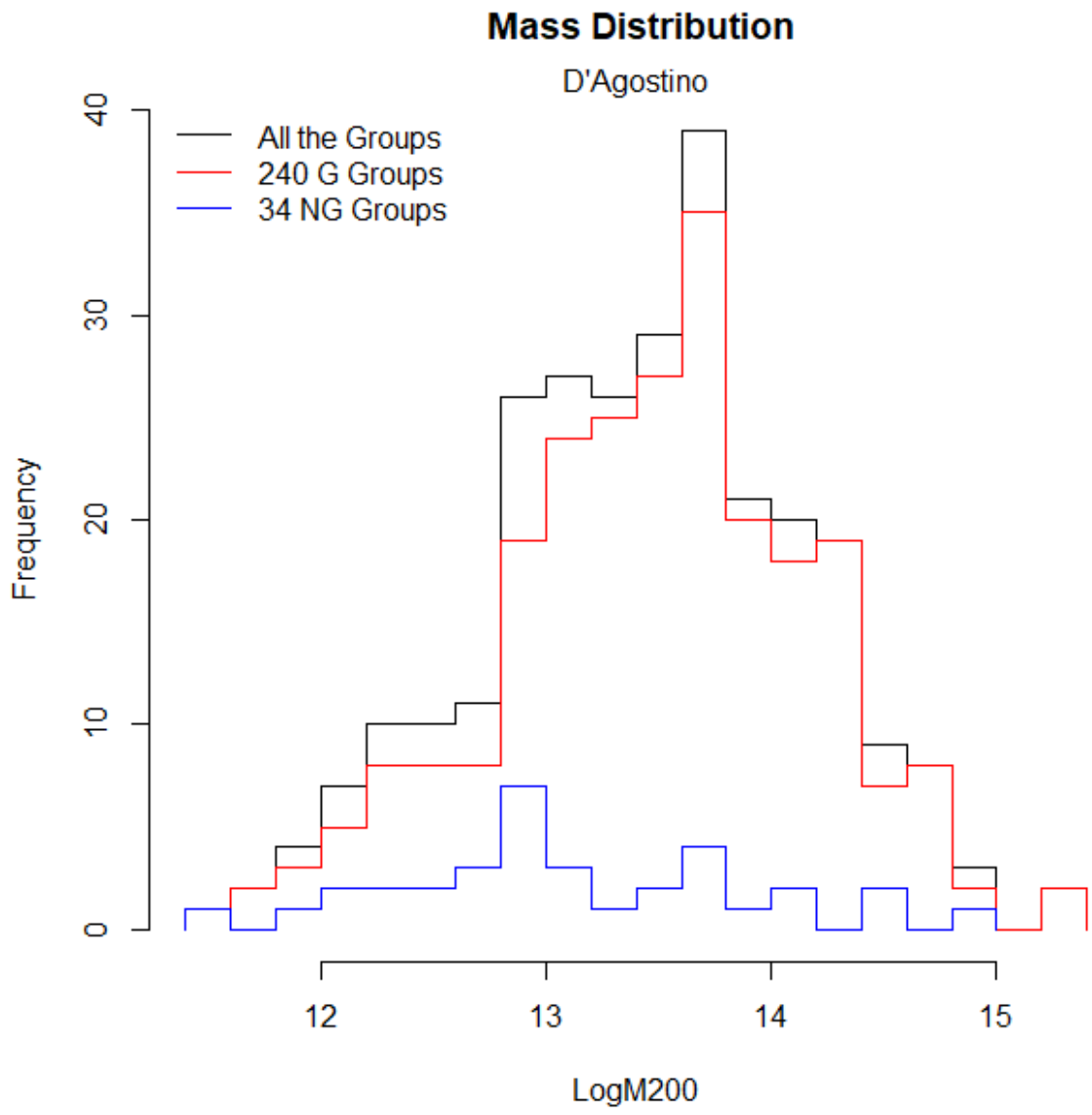


Figure 78

- Shapiro – Wilk Test

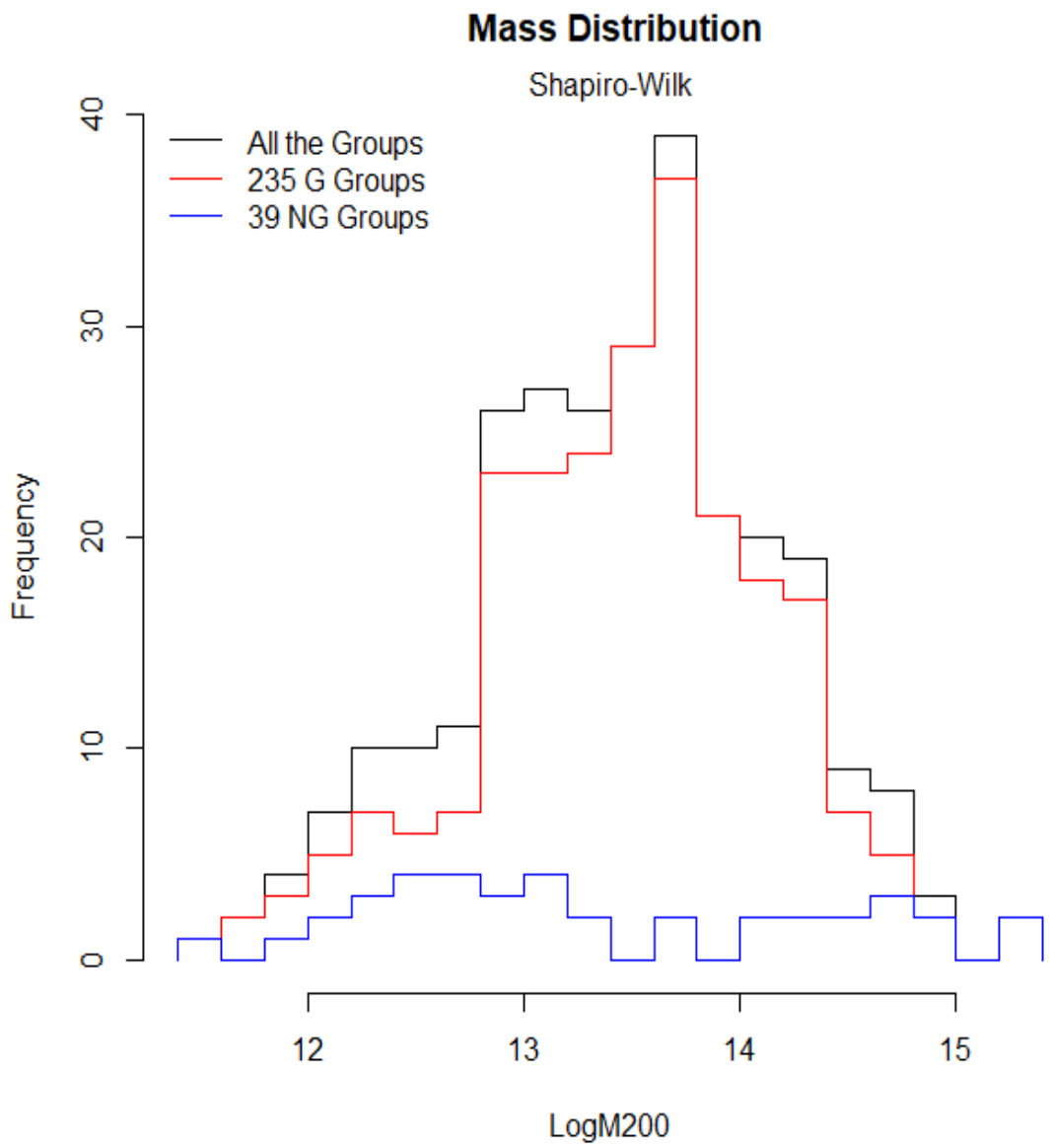


Figure 79

- Anderson-Darling Test

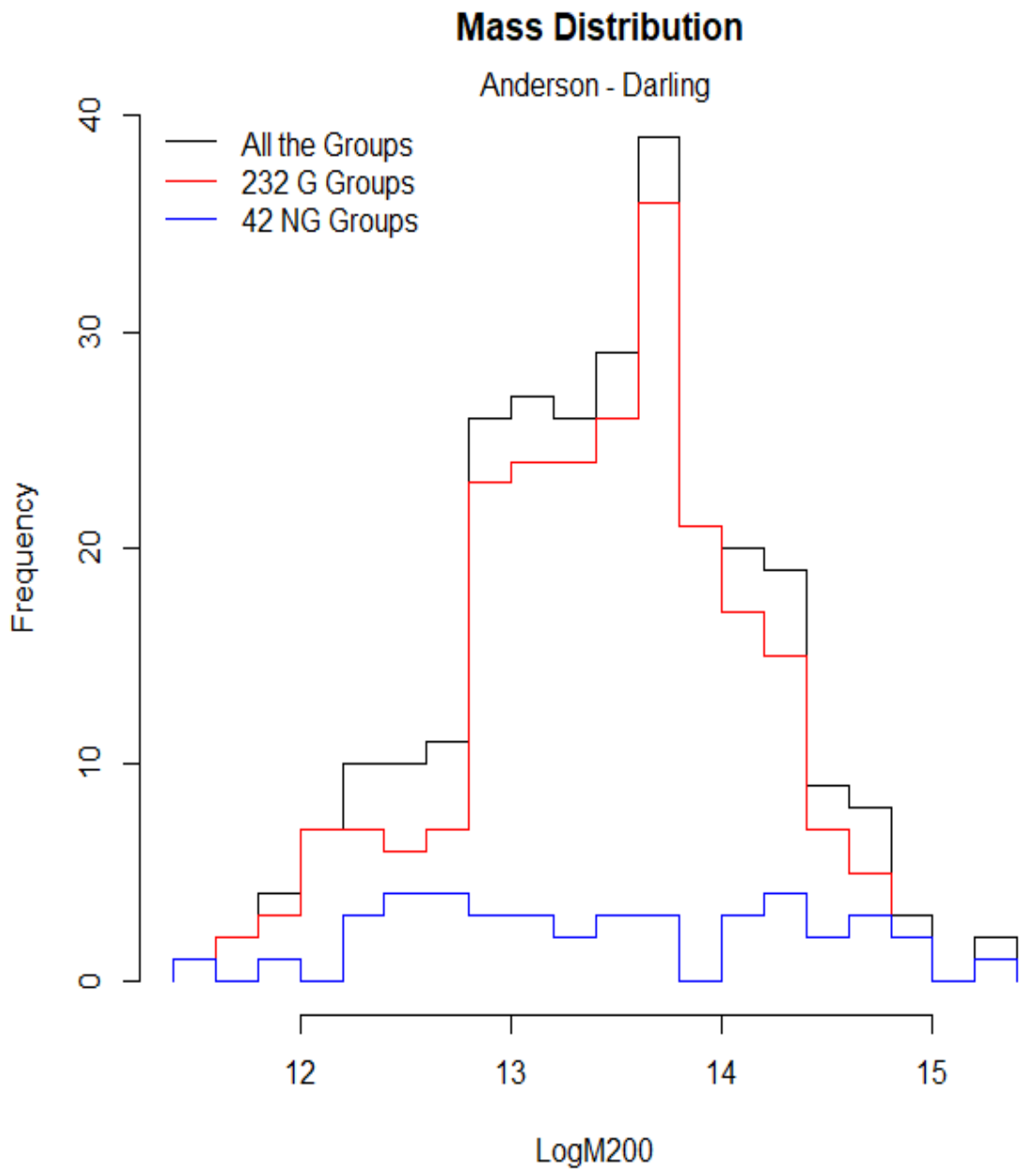


Figure 80

- Lilliefors Test

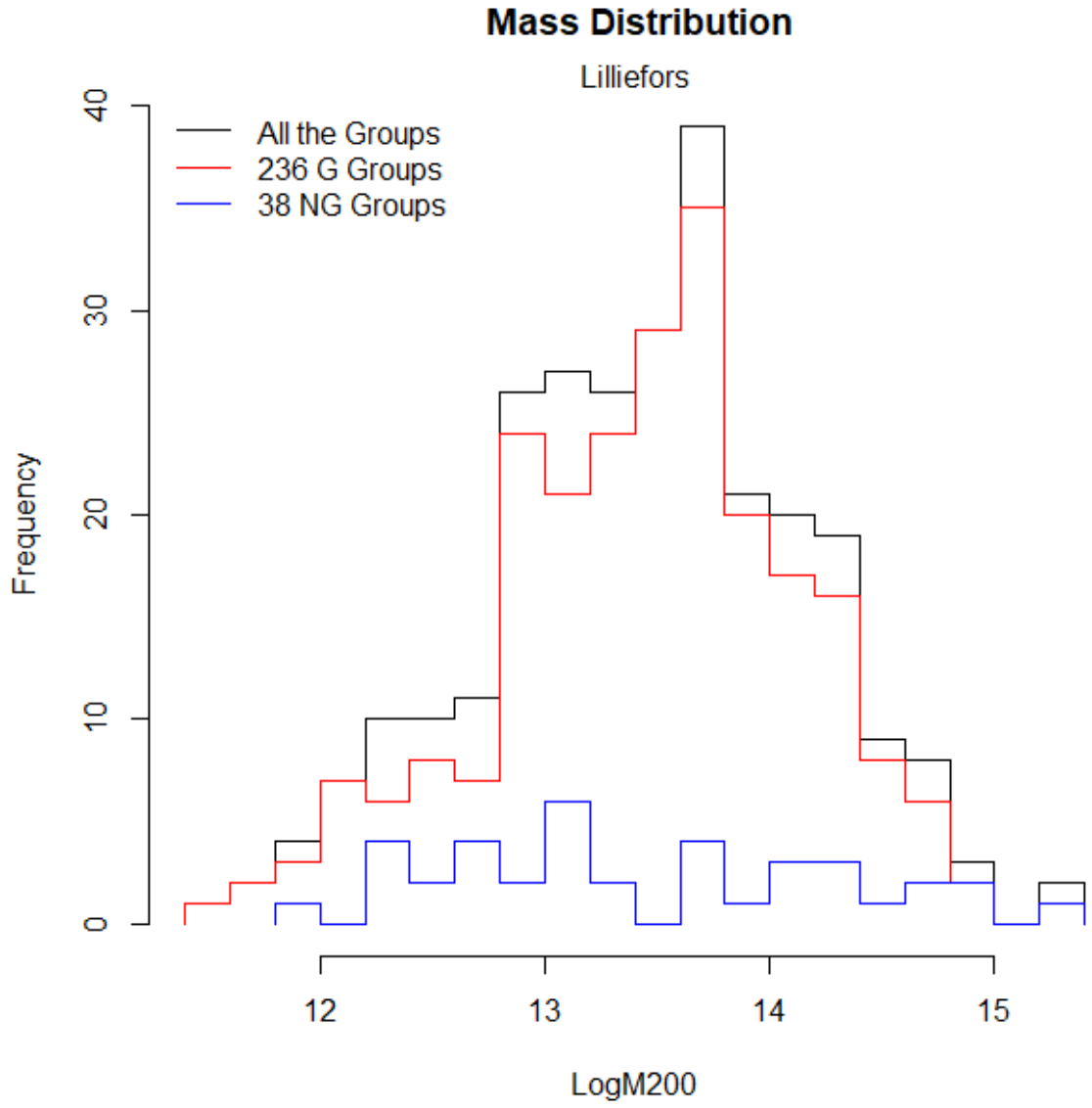


Figure 81

- Pearson's Test

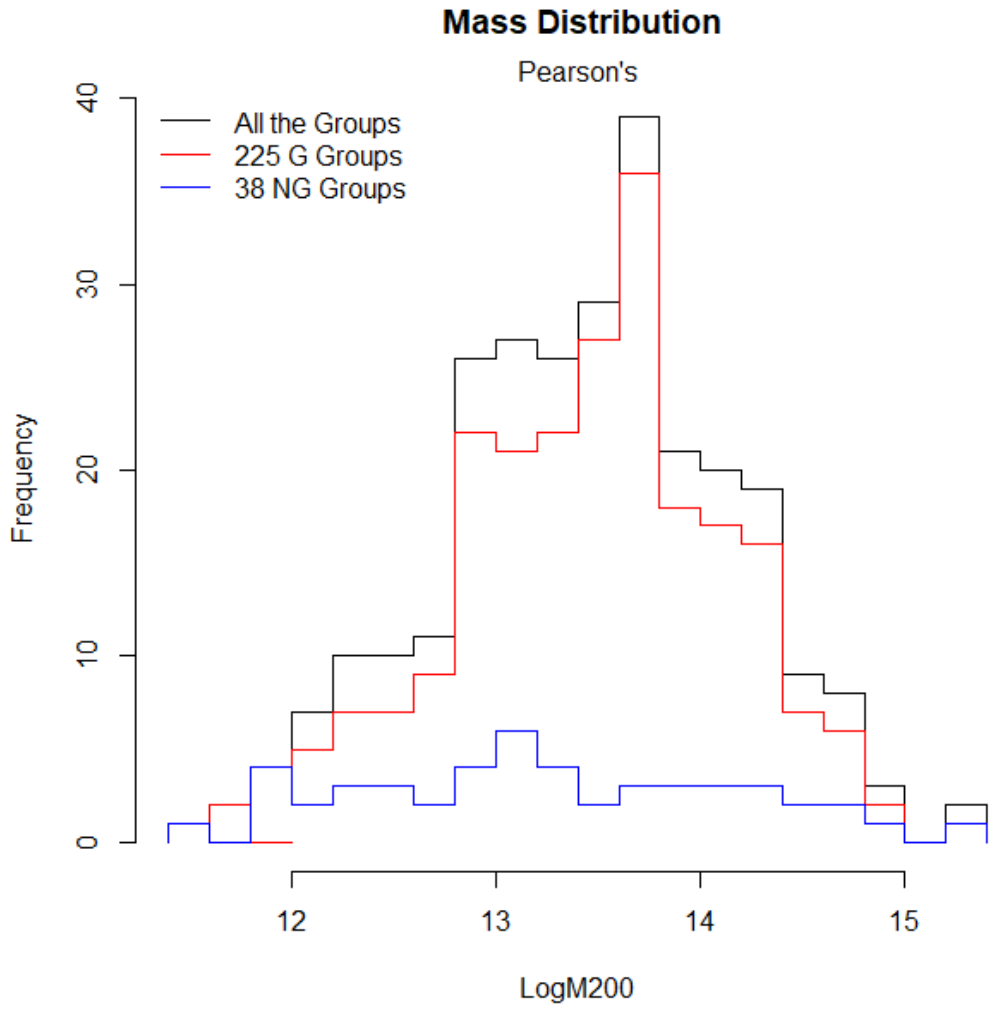


Figure 82

Magnitude Distribution

Our goal is to apply normality tests for the K-band magnitude of galaxy groups. We apply the Cramer-von Misses test for the Magnitude of the galaxy groups. We found that 106 groups are Non-Gaussian and 168 groups are Gaussian for their Magnitude for $a = 0.10$. In Figure73 we represent the empirical curve and the theoretical curve of the 6-th Non-Gaussian group.

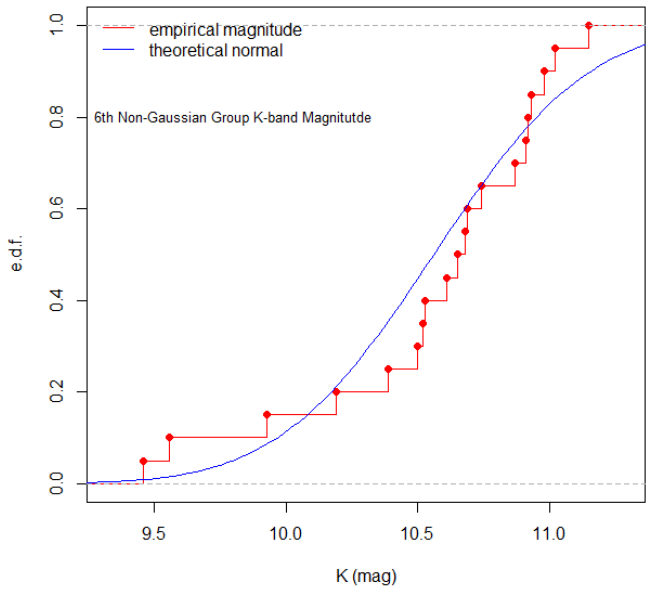


Figure 83

Next we demonstrate a QQ-plot to examine the normality of the 6-th group.

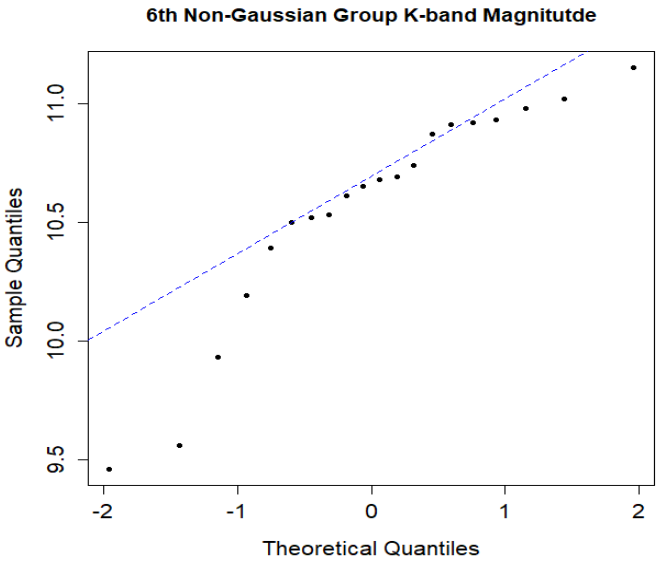


Figure 84

We do the same for the Gaussian 1249th group.

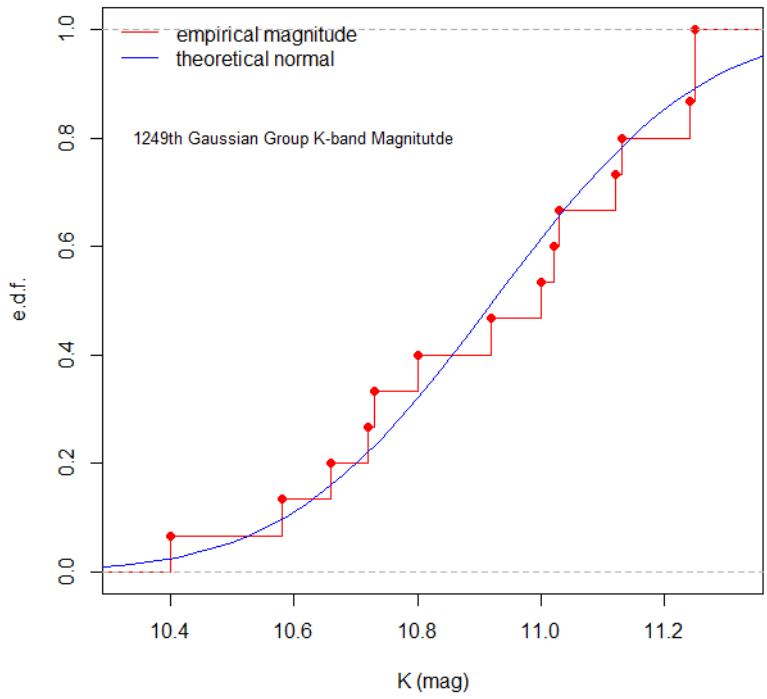


Figure 85

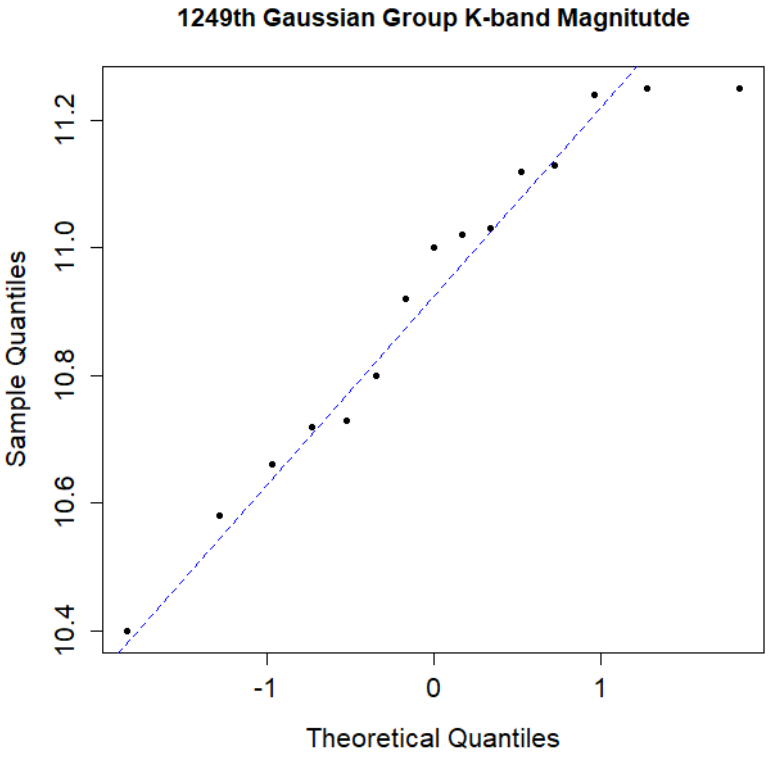


Figure 86

We do the same for the Non-Gaussian 1172th group.

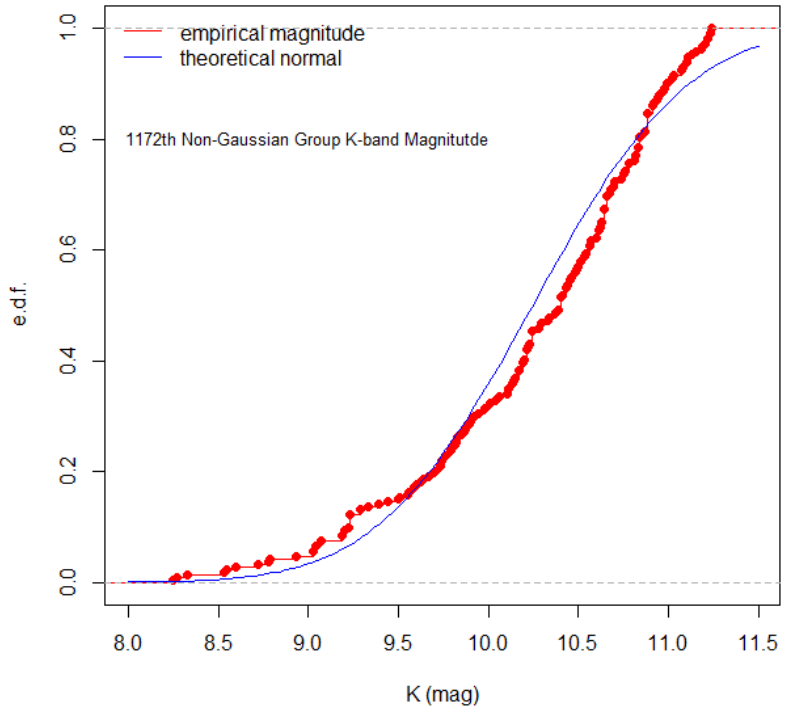


Figure 87

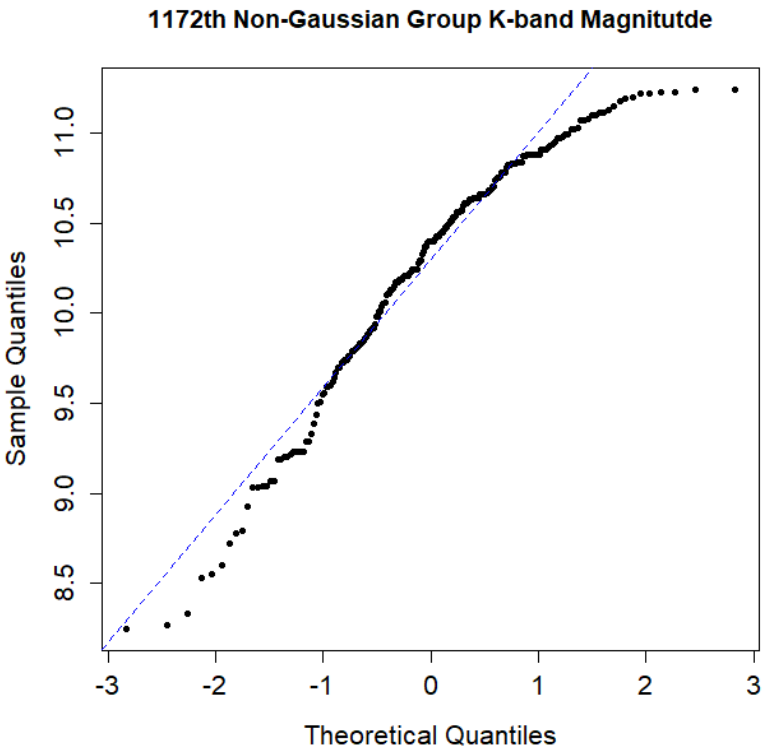


Figure 78

We do the same for the Non-Gaussian 312th group.

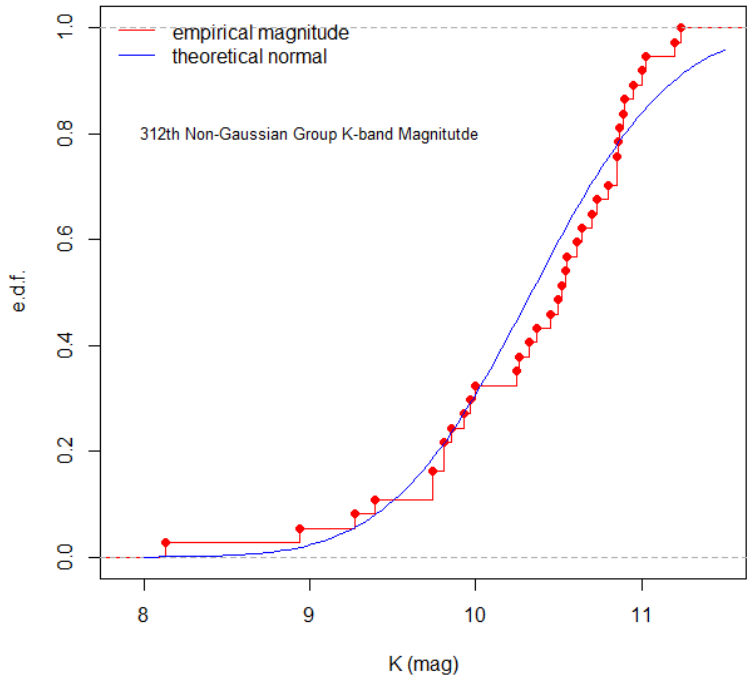


Figure 89

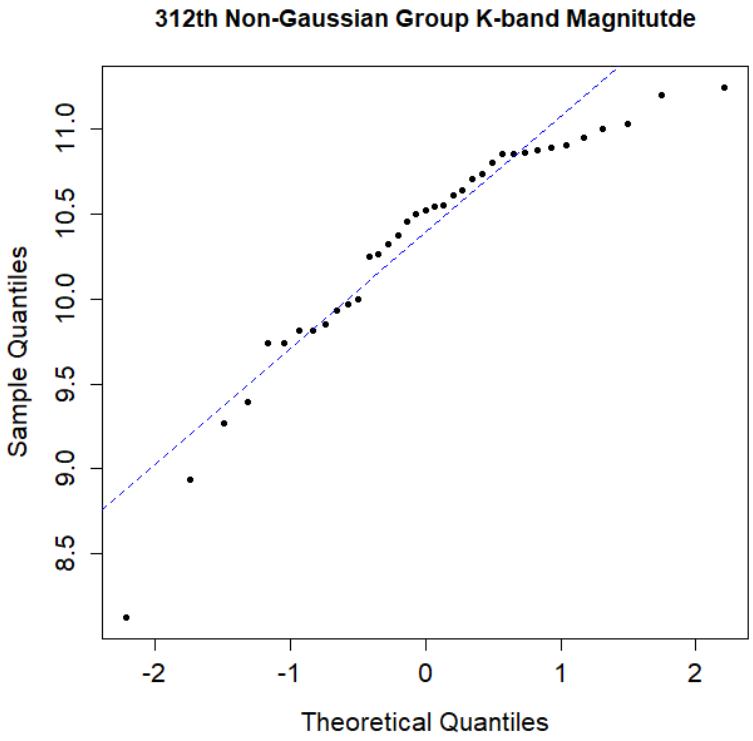


Figure 90

We do the same for the Gaussian 633th group.

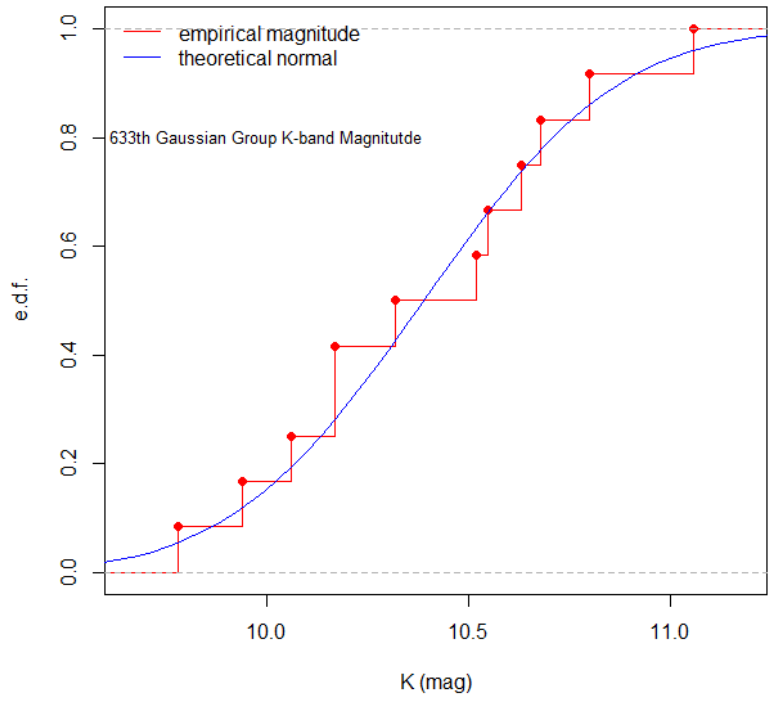


Figure 91

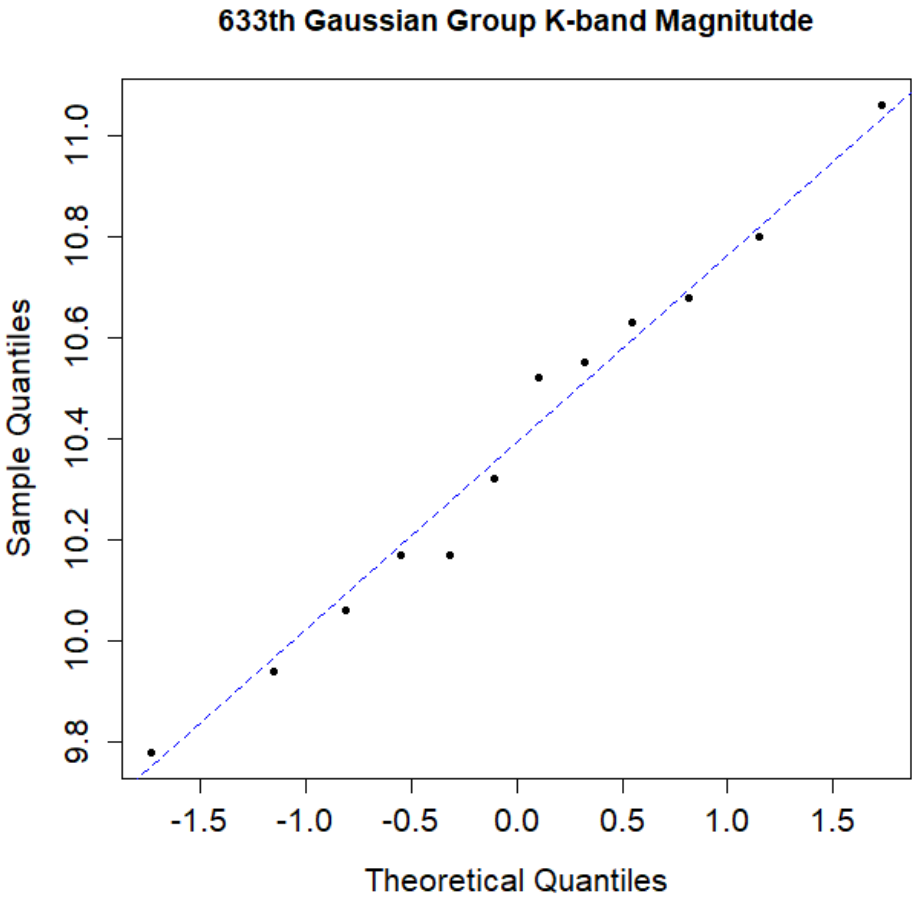

633th Gaussian Group K-band Magnitutde
Sample Quantiles
9.8
10.0
10.2
10.4
10.6
10.8
11.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles

Figure 92

We represent the Non-Gaussian Groups for the six tests. (Five (5) tests for velocities and one (1) for magnitude)

Table5.L

Tests for Normality					
/i-th Low Contrast Non - Gaussian Groups					
D'Agostino (velocities)	Shapiro – Wilk (velocities)	Anderson –Darling (velocities)	Lilliefors (velocities)	Pearson (velocities)	Cramer-von Misses (magnitude)
[1,] 1	[1,] 8	[1,] 8	[1,] 8	[1,] 6	[1,] 1
[2,] 3	[2,] 15	[2,] 15	[2,] 15	[2,] 9	[2,] 2
[3,] 9	[3,] 17	[3,] 17	[3,] 17	[3,] 15	[3,] 3
[4,] 14	[4,] 23	[4,] 23	[4,] 28	[4,] 17	[4,] 8
[5,] 17	[5,] 34	[5,] 28	[5,] 34	[5,] 19	[5,] 9
[6,] 23	[6,] 38	[6,] 31	[6,] 38	[6,] 25	[6,] 10
[7,] 28	[7,] 47	[7,] 34	[7,] 57	[7,] 31	[7,] 17
[8,] 34	[8,] 50	[8,] 38	[8,] 59	[8,] 34	[8,] 19
[9,] 38	[9,] 51	[9,] 44	[9,] 64	[9,] 38	[9,] 20
[10,] 47	[10,] 56	[10,] 51	[10,] 67	[10,] 44	[10,] 22
[11,] 64	[11,] 59	[11,] 56	[11,] 72	[11,] 45	[11,] 24
[12,] 77	[12,] 63	[12,] 57	[12,] 80	[12,] 49	[12,] 28
[13,] 85	[13,] 64	[13,] 59	[13,] 85	[13,] 57	[13,] 30
[14,] 95	[14,] 66	[14,] 64	[14,] 95	[14,] 59	[14,] 33
[15,] 106	[15,] 72	[15,] 72	[15,] 106	[15,] 60	[15,] 36
[16,] 128	[16,] 77	[16,] 77	[16,] 112	[16,] 69	[16,] 37
[17,] 135	[17,] 106	[17,] 85	[17,] 135	[17,] 72	[17,] 38
[18,] 138	[18,] 112	[18,] 98	[18,] 146	[18,] 77	[18,] 39
[19,] 146	[19,] 128	[19,] 106	[19,] 149	[19,] 80	[19,] 43
[20,] 147	[20,] 135	[20,] 112	[20,] 156	[20,] 82	[20,] 45
[21,] 149	[21,] 146	[21,] 128	[21,] 160	[21,] 85	[21,] 51
[22,] 159	[22,] 147	[22,] 135	[22,] 186	[22,] 89	[22,] 54
[23,] 162	[23,] 149	[23,] 146	[23,] 189	[23,] 100	[23,] 55
[24,] 178	[24,] 159	[24,] 149	[24,] 190	[24,] 106	[24,] 56
[25,] 189	[25,] 160	[25,] 156	[25,] 194	[25,] 108	[25,] 57
[26,] 190	[26,] 162	[26,] 160	[26,] 196	[26,] 113	[26,] 58
[27,] 197	[27,] 178	[27,] 162	[27,] 197	[27,] 128	[27,] 59
[28,] 204	[28,] 189	[28,] 189	[28,] 200	[28,] 135	[28,] 61
[29,] 206	[29,] 190	[29,] 190	[29,] 204	[29,] 149	[29,] 65
[30,] 219	[30,] 193	[30,] 193	[30,] 206	[30,] 151	[30,] 68
[31,] 253	[31,] 194	[31,] 194	[31,] 219	[31,] 156	[31,] 69
[32,] 256	[32,] 196	[32,] 196	[32,] 221	[32,] 157	[32,] 73
[33,] 257	[33,] 197	[33,] 197	[33,] 223	[33,] 160	[33,] 74
[34,] 264	[34,] 204	[34,] 204	[34,] 239	[34,] 174	[34,] 75
[35,] NA	[35,] 206	[35,] 206	[35,] 241	[35,] 189	[35,] 77
[36,] NA	[36,] 219	[36,] 219	[36,] 257	[36,] 190	[36,] 81
[37,] NA	[37,] 241	[37,] 241	[37,] 270	[37,] 193	[37,] 84
[38,] NA	[38,] 256	[38,] 256	[38,] 273	[38,] 194	[38,] 87
[39,] NA	[39,] 257	[39,] 257	[39,] NA	[39,] 196	[39,] 90
[40,] NA	[40,] NA	[40,] 260	[40,] NA	[40,] 197	[40,] 92
[41,] NA	[41,] NA	[41,] 265	[41,] NA	[41,] 202	[41,] 93
[42,] NA	[42,] NA	[42,] 270	[42,] NA	[42,] 206	[42,] 94
[43,] NA	[43,] NA	[43,] NA	[43,] NA	[43,] 219	[43,] 99
[44,] NA	[44,] NA	[44,] NA	[44,] NA	[44,] 221	[44,] 102
[45,] NA	[45,] NA	[45,] NA	[45,] NA	[45,] 239	[45,] 105
[46,] NA	[46,] NA	[46,] NA	[46,] NA	[46,] 241	[46,] 107
[47,] NA	[47,] NA	[47,] NA	[47,] NA	[47,] 242	[47,] 110
[48,] NA	[48,] NA	[48,] NA	[48,] NA	[48,] 257	[48,] 111
[49,] NA	[49,] NA	[49,] NA	[49,] NA	[49,] 267	[49,] 117
[50,] NA	[50,] NA	[50,] NA	[50,] NA	[50,] NA	[50,] 118
[51,] NA	[51,] NA	[51,] NA	[51,] NA	[51,] NA	[51,] 120
[52,] NA	[52,] NA	[52,] NA	[52,] NA	[52,] NA	[52,] 121
[53,] NA	[53,] NA	[53,] NA	[53,] NA	[53,] NA	[53,] 123
[54,] NA	[54,] NA	[54,] NA	[54,] NA	[54,] NA	[54,] 124
[55,] NA	[55,] NA	[55,] NA	[55,] NA	[55,] NA	[55,] 129
[56,] NA	[56,] NA	[56,] NA	[56,] NA	[56,] NA	[56,] 130
[57,] NA	[57,] NA	[57,] NA	[57,] NA	[57,] NA	[57,] 135
[58,] NA	[58,] NA	[58,] NA	[58,] NA	[58,] NA	[58,] 140
[59,] NA	[59,] NA	[59,] NA	[59,] NA	[59,] NA	[59,] 141
[60,] NA	[60,] NA	[60,] NA	[60,] NA	[60,] NA	[60,] 142
[61,] NA	[61,] NA	[61,] NA	[61,] NA	[61,] NA	[61,] 143

[62,] NA	[62,] NA	[62,] NA	[62,] NA	[62,] NA	[62,] 153
[63,] NA	[63,] NA	[63,] NA	[63,] NA	[63,] NA	[63,] 159
[64,] NA	[64,] NA	[64,] NA	[64,] NA	[64,] NA	[64,] 160
[65,] NA	[65,] NA	[65,] NA	[65,] NA	[65,] NA	[65,] 163
[66,] NA	[66,] NA	[66,] NA	[66,] NA	[66,] NA	[66,] 166
[67,] NA	[67,] NA	[67,] NA	[67,] NA	[67,] NA	[67,] 167
[68,] NA	[68,] NA	[68,] NA	[68,] NA	[68,] NA	[68,] 168
[69,] NA	[69,] NA	[69,] NA	[69,] NA	[69,] NA	[69,] 170
[70,] NA	[70,] NA	[70,] NA	[70,] NA	[70,] NA	[70,] 172
[71,] NA	[71,] NA	[71,] NA	[71,] NA	[71,] NA	[71,] 175
[72,] NA	[72,] NA	[72,] NA	[72,] NA	[72,] NA	[72,] 177
[73,] NA	[73,] NA	[73,] NA	[73,] NA	[73,] NA	[73,] 182
[74,] NA	[74,] NA	[74,] NA	[74,] NA	[74,] NA	[74,] 184
[75,] NA	[75,] NA	[75,] NA	[75,] NA	[75,] NA	[75,] 190
[76,] NA	[76,] NA	[76,] NA	[76,] NA	[76,] NA	[76,] 191
[77,] NA	[77,] NA	[77,] NA	[77,] NA	[77,] NA	[77,] 196
[78,] NA	[78,] NA	[78,] NA	[78,] NA	[78,] NA	[78,] 197
[79,] NA	[79,] NA	[79,] NA	[79,] NA	[79,] NA	[79,] 198
[80,] NA	[80,] NA	[80,] NA	[80,] NA	[80,] NA	[80,] 200
[81,] NA	[81,] NA	[81,] NA	[81,] NA	[81,] NA	[81,] 202
[82,] NA	[82,] NA	[82,] NA	[82,] NA	[82,] NA	[82,] 205
[83,] NA	[83,] NA	[83,] NA	[83,] NA	[83,] NA	[83,] 210
[84,] NA	[84,] NA	[84,] NA	[84,] NA	[84,] NA	[84,] 212
[85,] NA	[85,] NA	[85,] NA	[85,] NA	[85,] NA	[85,] 215
[86,] NA	[86,] NA	[86,] NA	[86,] NA	[86,] NA	[86,] 216
[87,] NA	[87,] NA	[87,] NA	[87,] NA	[87,] NA	[87,] 218
[88,] NA	[88,] NA	[88,] NA	[88,] NA	[88,] NA	[88,] 222
[89,] NA	[89,] NA	[89,] NA	[89,] NA	[89,] NA	[89,] 223
[90,] NA	[90,] NA	[90,] NA	[90,] NA	[90,] NA	[90,] 226
[91,] NA	[91,] NA	[91,] NA	[91,] NA	[91,] NA	[91,] 227
[92,] NA	[92,] NA	[92,] NA	[92,] NA	[92,] NA	[92,] 231
[93,] NA	[93,] NA	[93,] NA	[93,] NA	[93,] NA	[93,] 232
[94,] NA	[94,] NA	[94,] NA	[94,] NA	[94,] NA	[94,] 234
[95,] NA	[95,] NA	[95,] NA	[95,] NA	[95,] NA	[95,] 235
[96,] NA	[96,] NA	[96,] NA	[96,] NA	[96,] NA	[96,] 239
[97,] NA	[97,] NA	[97,] NA	[97,] NA	[97,] NA	[97,] 241
[98,] NA	[98,] NA	[98,] NA	[98,] NA	[98,] NA	[98,] 252
[99,] NA	[99,] NA	[99,] NA	[99,] NA	[99,] NA	[99,] 254
[100,] NA	[100,] NA	[100,] NA	[100,] NA	[100,] NA	[100,] 257
[101,] NA	[101,] NA	[101,] NA	[101,] NA	[101,] NA	[101,] 259
[102,] NA	[102,] NA	[102,] NA	[102,] NA	[102,] NA	[102,] 260
[103,] NA	[103,] NA	[103,] NA	[103,] NA	[103,] NA	[103,] 268
[104,] NA	[104,] NA	[104,] NA	[104,] NA	[104,] NA	[104,] 271
[105,] NA	[105,] NA	[105,] NA	[105,] NA	[105,] NA	[105,] 272
[106,] NA	[106,] NA	[106,] NA	[106,] NA	[106,] NA	[106,] 273

The intersection of the six tests is represented below:

Normality Tests
D'Agostino, Shapiro-Wilk, Anderson - Darling, Lilliefors, Pearson's, Cramer-von Mises
Non-Gaussian
17
38
135
190
197
257

Density Distributions

Here we test the normality of the density low contrast groups of the 2MASS sample. In Figure83 we plotted the density values of each group.

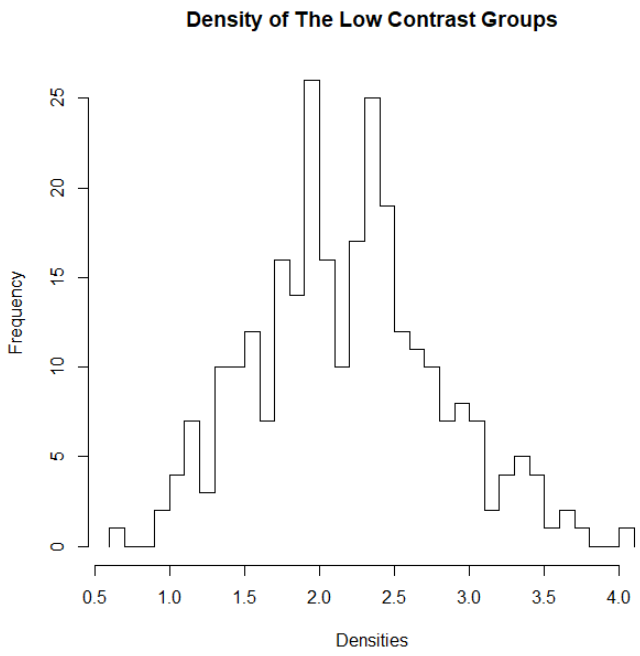


Figure 93

We continue with the theoretical and the empirical curves of the densities of the low contrast groups.

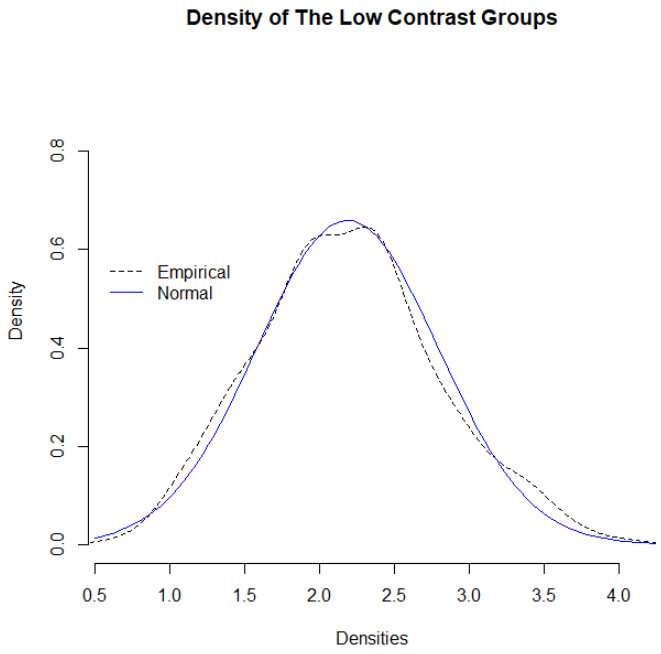


Figure 94

In Figure95 we do the same for the empirical distribution function and we compare the edf with the theoretical distribution function.

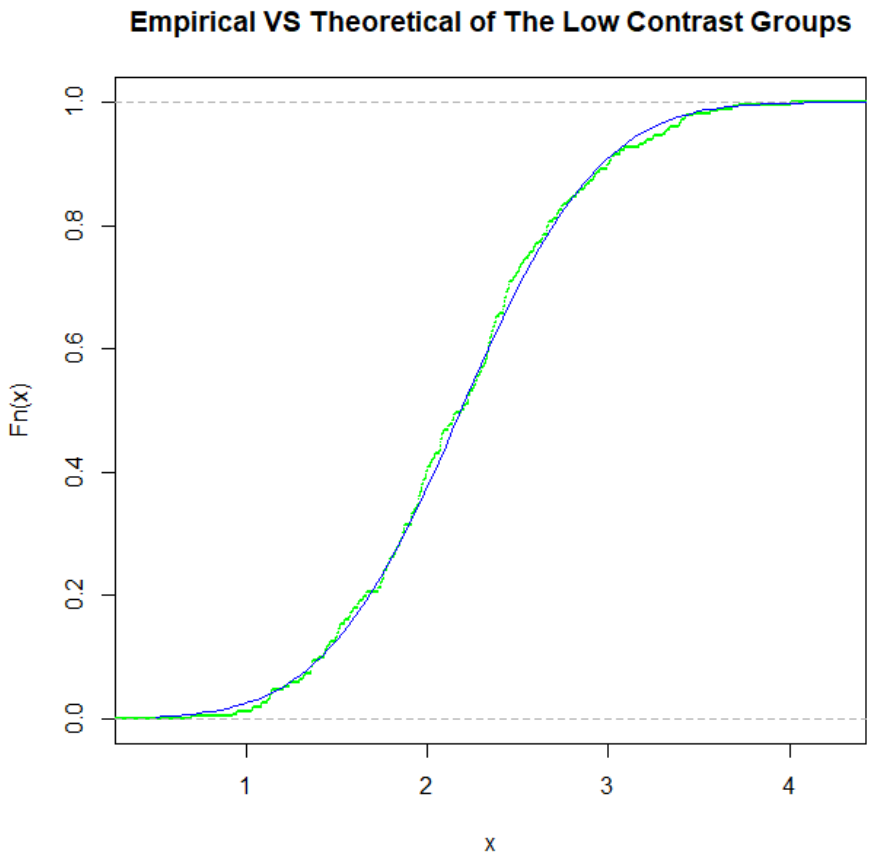


Figure 95

So we understand from the Figure84, 85 that the densities of the groups comes from a normal distribution with $mean = 2.187858$ and $standard\ deviation = 0.6042814$.

We can do many tests in R to examine the normality. We apply six (6) tests for examining the normality for all the groups' densities and we conclude that the densities of the groups come from normal distribution. The six tests are in agreement with the above proposition.

Table 6.L

Normality Tests	p.value	significance level α	passed/failed
A-D	0.32604089462296	0,1	passed
CVM	0.388885234716402	0,1	passed
LILLIE	0.39239130705169	0,1	passed
S-W	0.252480389695359	0,1	passed
PEARSON	0.344365181537625	0,1	passed
AGOSTINO	0.105096039726229	0,1	passed

Figure86 helps us understand better the p-values of the tests.

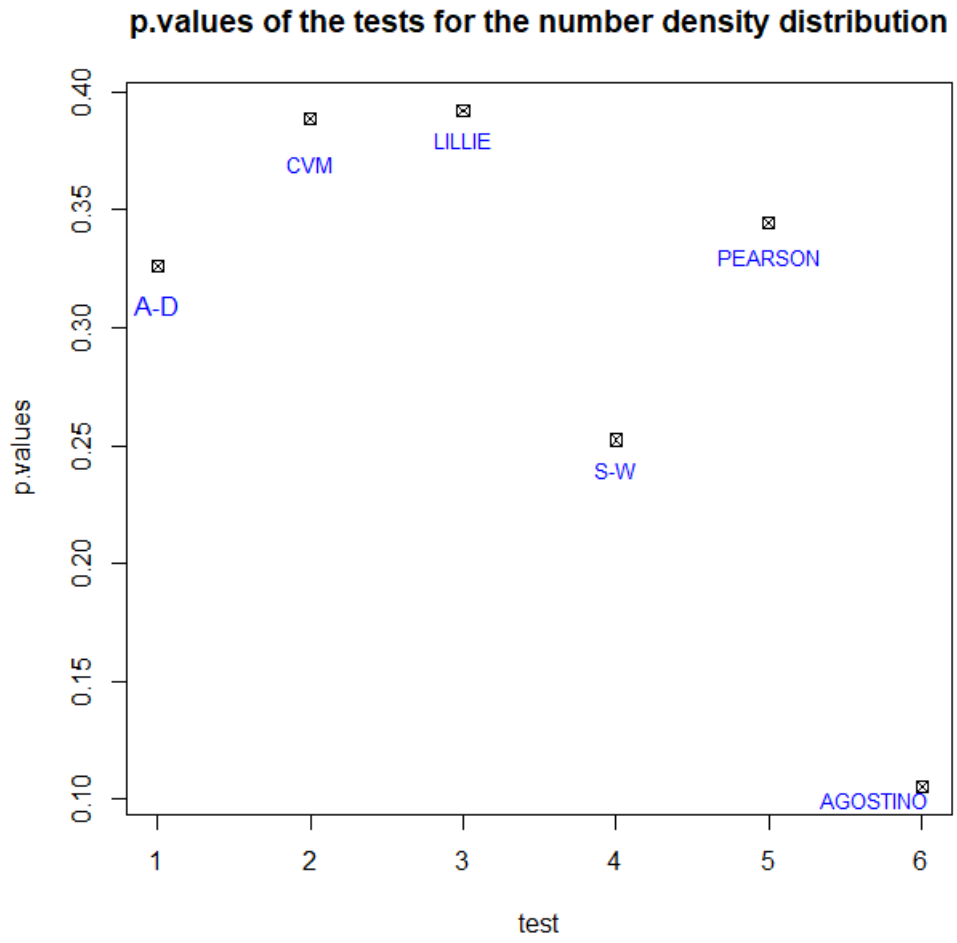


Figure 96

In the end we represent the coordinates of the group galaxies for the 2MASS sample. We have plotted the Non-Gaussian and the Gaussian Low Contrast groups and the High Contrast groups.

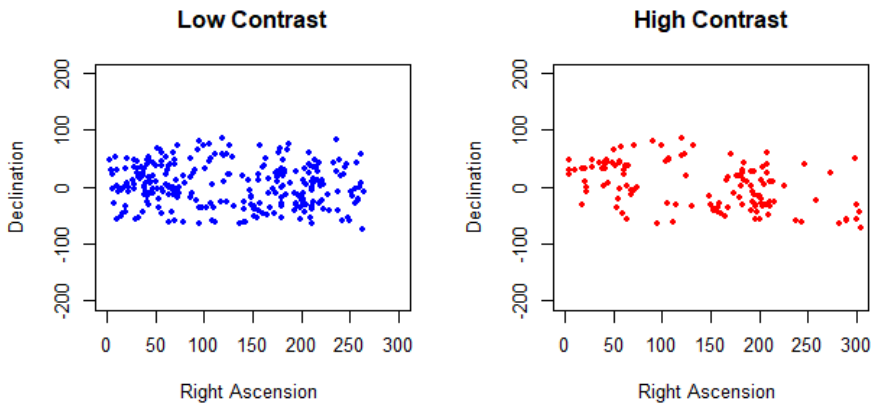


Figure 97 In this Figure we plotted all the groups of Low contrast and High contrast.

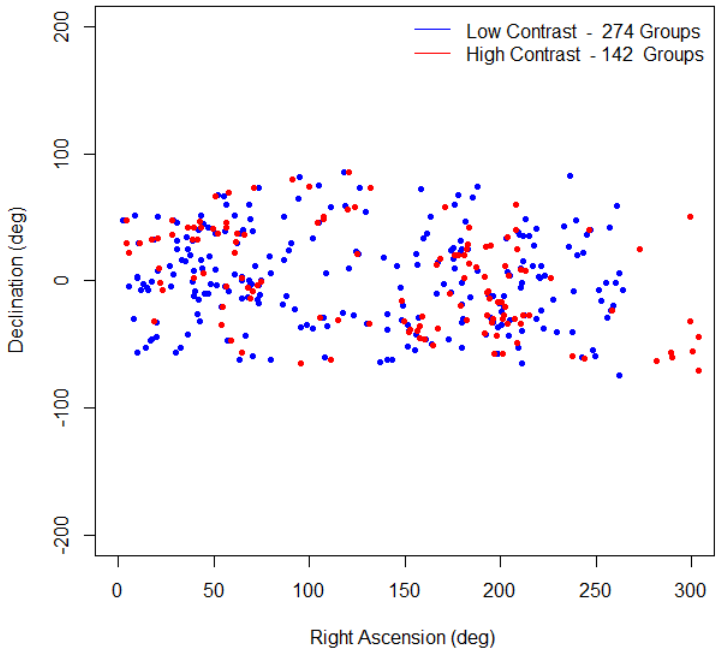


Figure 98 Here we put all the groups in one plot

Next we have plotted the Gaussian and Non-Gaussian groups of the two types contrast groups according to the metanalysis.

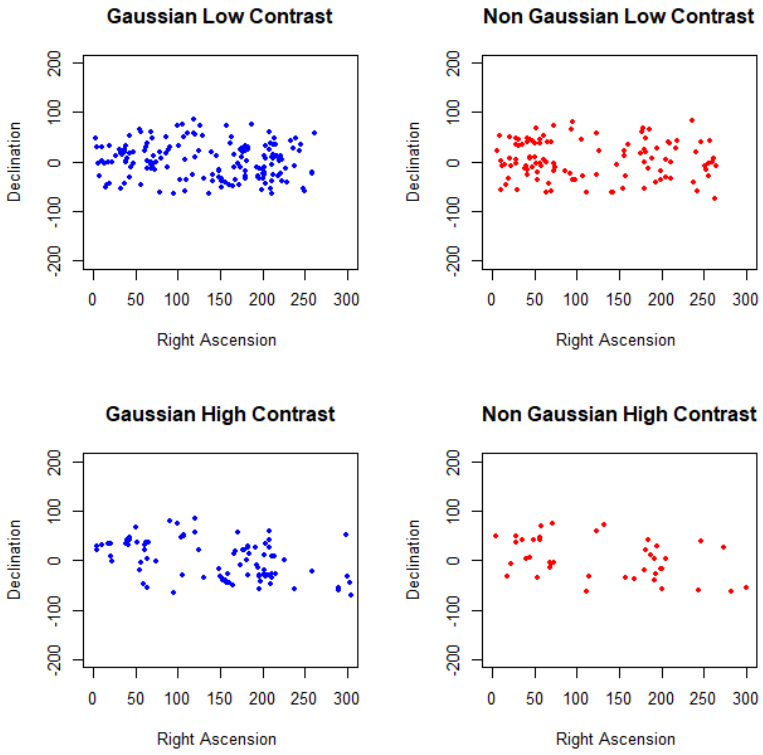


Figure 99

Or we can plot Non-Gaussian groups and Gaussian groups separated (Figure90) for Low and High contrast or all together (Figure91).

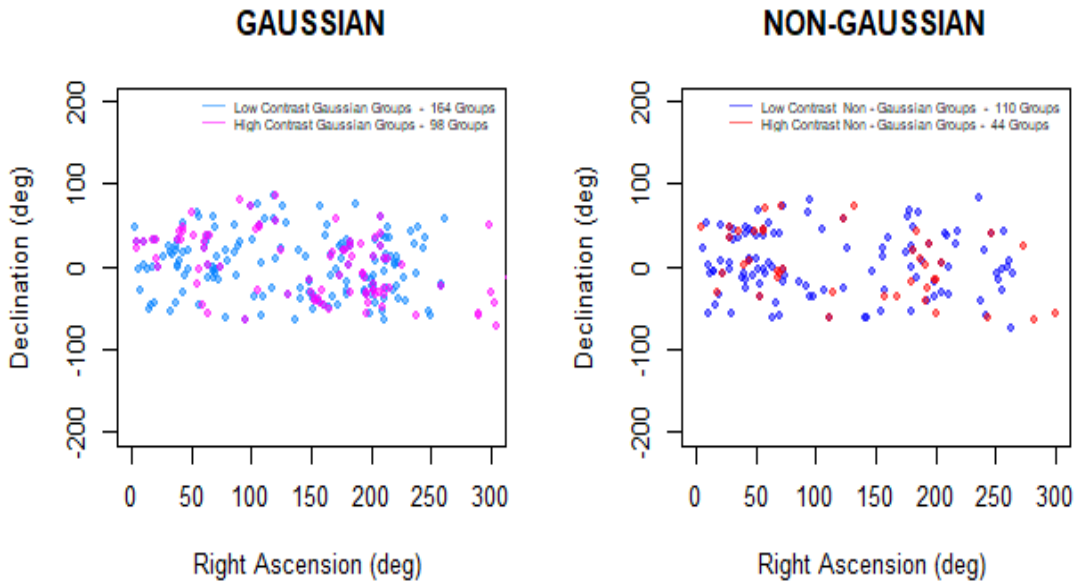


Figure 100

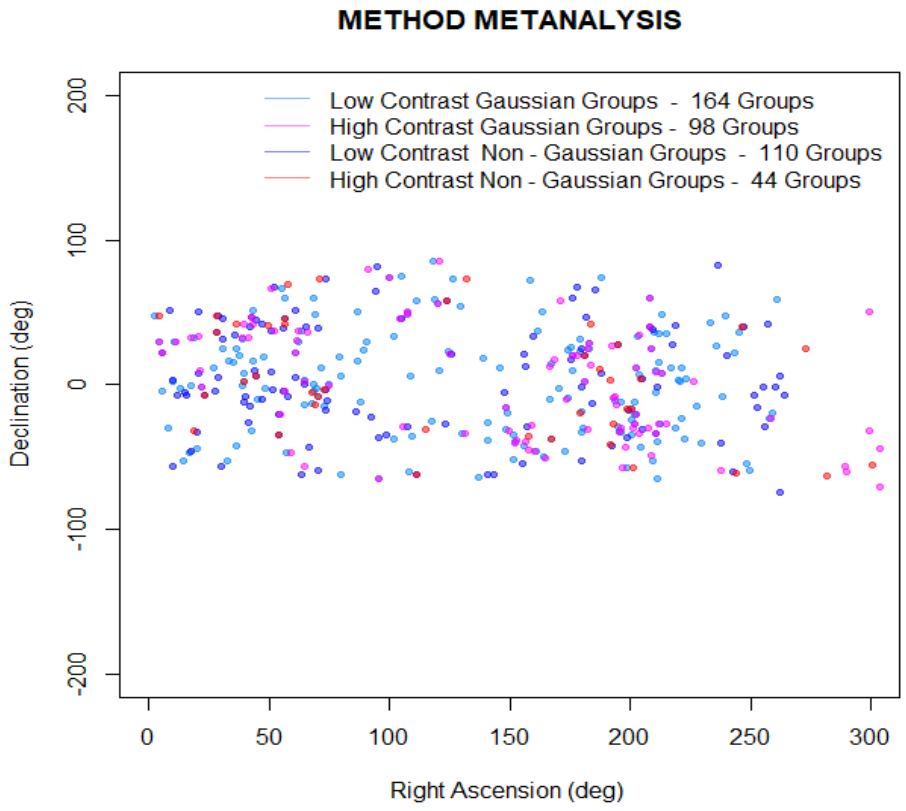


Figure 101

We want to test which groups are dynamically complex, using the radius of the groups and the number of members in each group. We represent six (6) histograms showing that Non-Gaussian groups are more dynamically complex than the Gaussian.

- D' Agostino Test

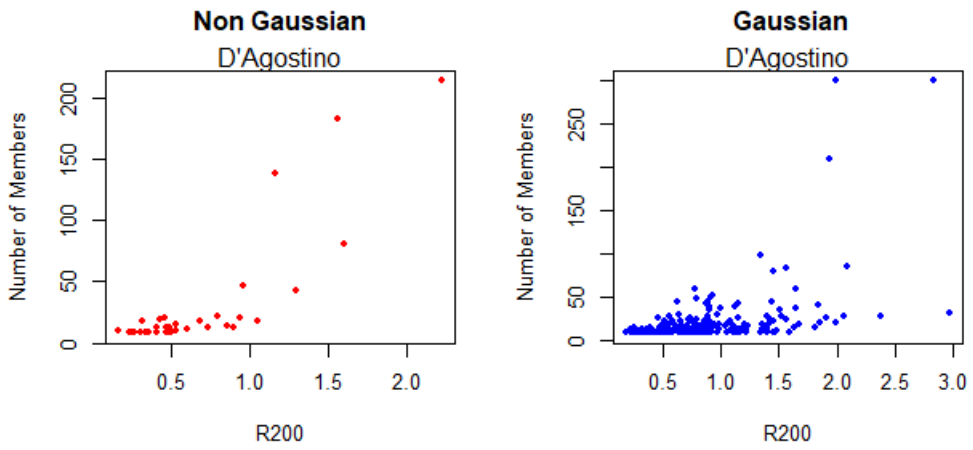


Figure 102

- Shapiro-Wilk Test

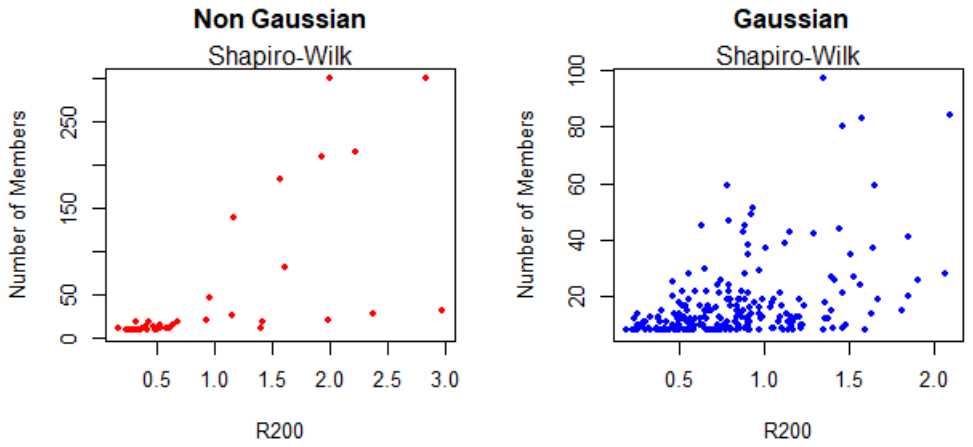


Figure 103

- Anderson-Darling Test

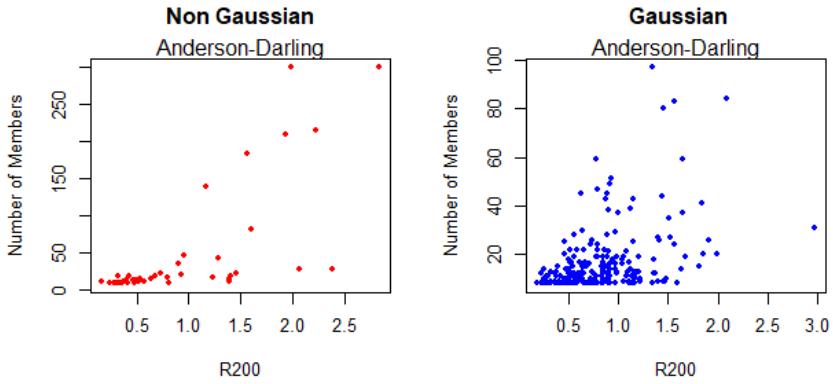


Figure 104

- Lilliefors Test

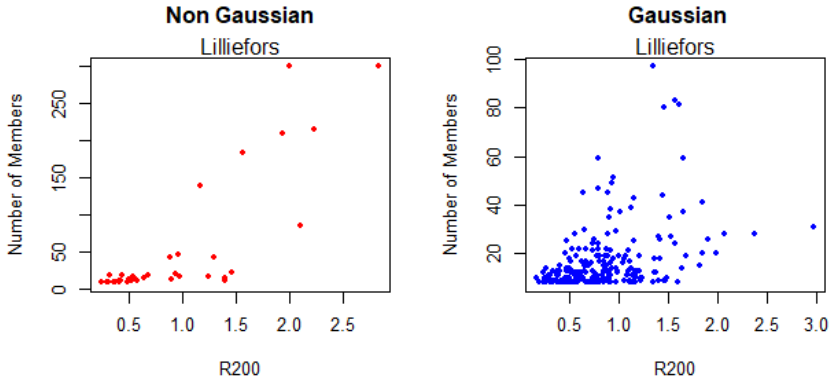


Figure 105

- Pearson's Test

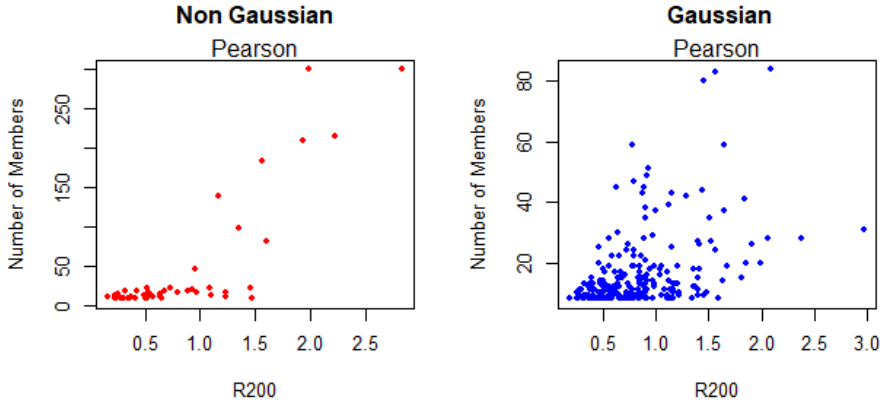


Figure 106

- Cramer-von Mises Test

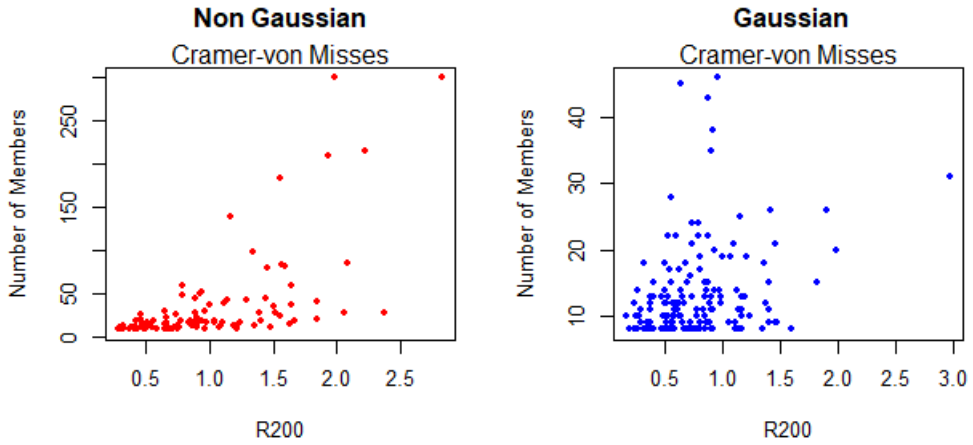


Figure 107

4 Rayleigh Distribution in Non-Gaussian Groups

The statistical analysis we did in the sample 2MASS was quite helpful understand the distribution of velocity dispersion of the low and high contrast groups. We conduct that the velocity distribution of the Low contrast groups follows a Rayleigh distribution. Also we found that the Gaussian and Non-Gaussian groups follows approximately the Rayleigh distribution. Below, we represent three (3) histograms showing the values of velocity dispersion of all the low contrast groups, the Gaussian groups and the Non-Gaussian groups.

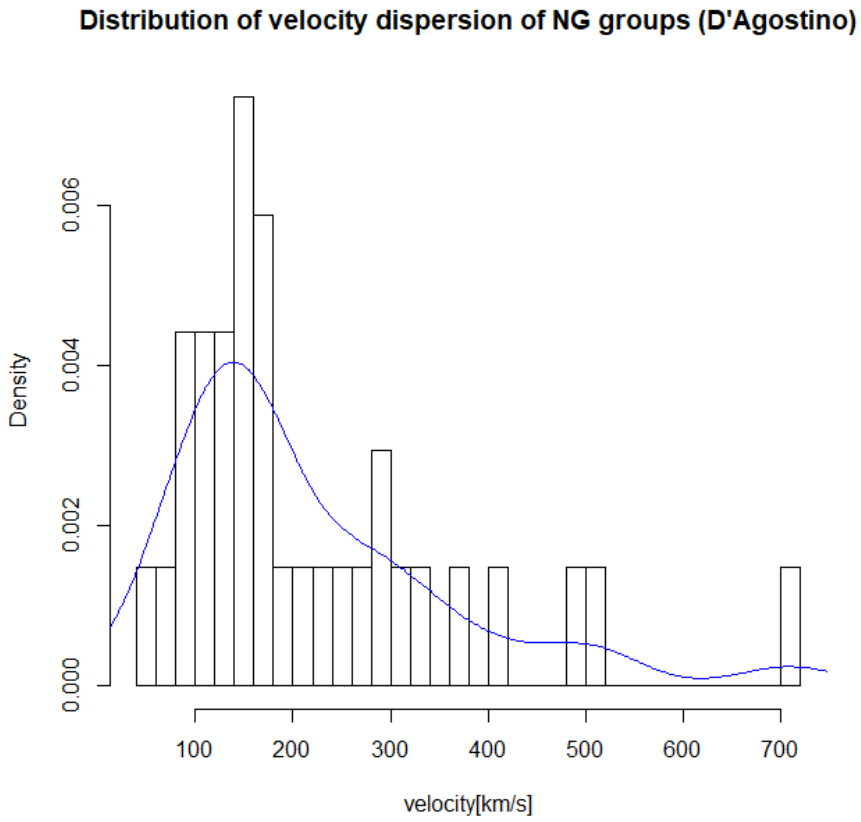


Figure 109 Values of the velocity dispersion of the Non-Gaussian groups. The method we use is the D'Agostino normality test

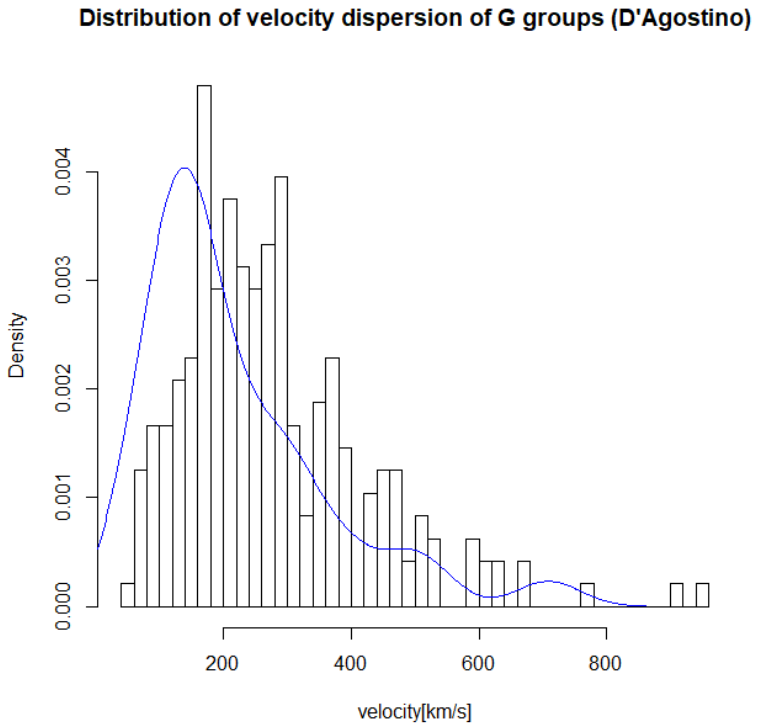


Figure 110 Values of the velocity dispersion of the Non-Gaussian groups. The method we use is the D'Agostino normality test

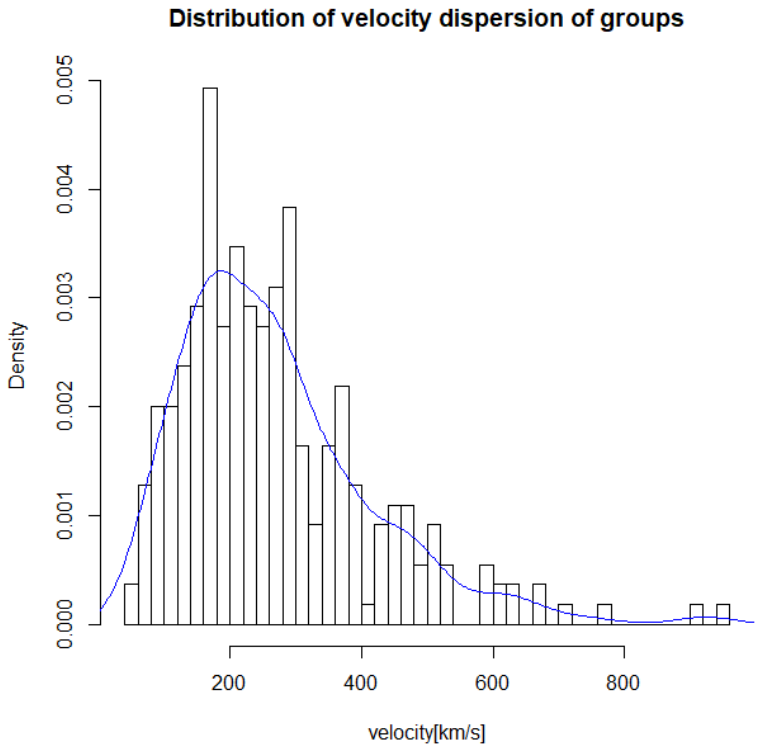


Figure 111 Values of the velocity dispersion of all the groups.

Now we will learn about the Rayleigh distribution.

In probability theory and statistics, the **Rayleigh distribution** is a continuous probability distribution for nonnegative-valued random variables. It is essentially a chi distribution with two degrees of freedom.

Definition

The probability density function of the Rayleigh distribution is:

$$f(x;\sigma) = \frac{x}{\sigma^2} e^{\frac{-x^2}{2\sigma^2}}, \qquad x \geq 0$$

Where σ is the scale parameter of the distribution. The cumulative distribution is :

$$F(x;\sigma) = 1 - e^{\frac{-x^2}{2\sigma^2}}, \qquad x \geq 0$$

Generating Random Variables

Given a random variable U draw from the Uniform distribution in the interval $(0,1)$, then the variable has a Rayleigh distribution with parameter σ. This is obtained by applying the inverse transform sampling-method.

$$X = \sigma\sqrt{-\ln(U)}$$

Parameter Estimation

Given a sample of N independent and identically distributed Rayleigh random variables x_i with parameter σ.

$$\hat{\sigma} \approx \sqrt{\frac{1}{2N}\sum_{i=1}^{N} x_i^2}$$

$\hat{\sigma}$ is the maximum likelihood estimate unbiased.

Here we represent histograms showing the probability density function and the cumulative distribution function of Rayleigh distribution.

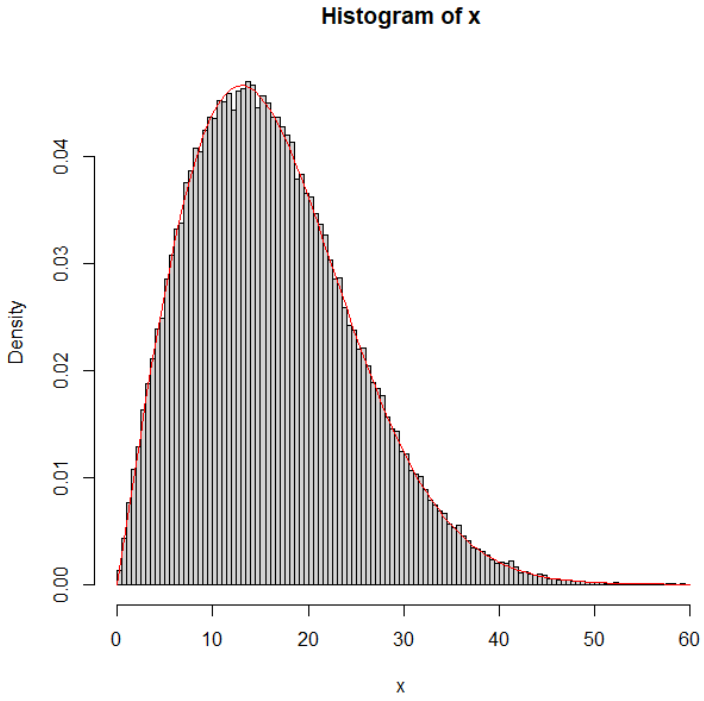


Figure 112 probability density function of a sample comes from a Rayleigh distribution

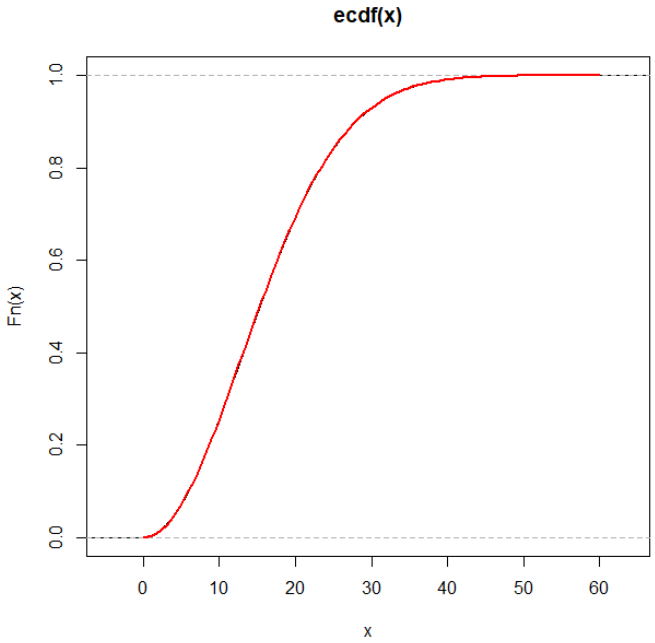


Figure 113 cumulative distribution function of a sample comes from a Rayleigh distribution

Back to our analysis we represent some simulations in order to find the empirical parameter of Rayleigh for the Gaussian groups of Low contrast sample.

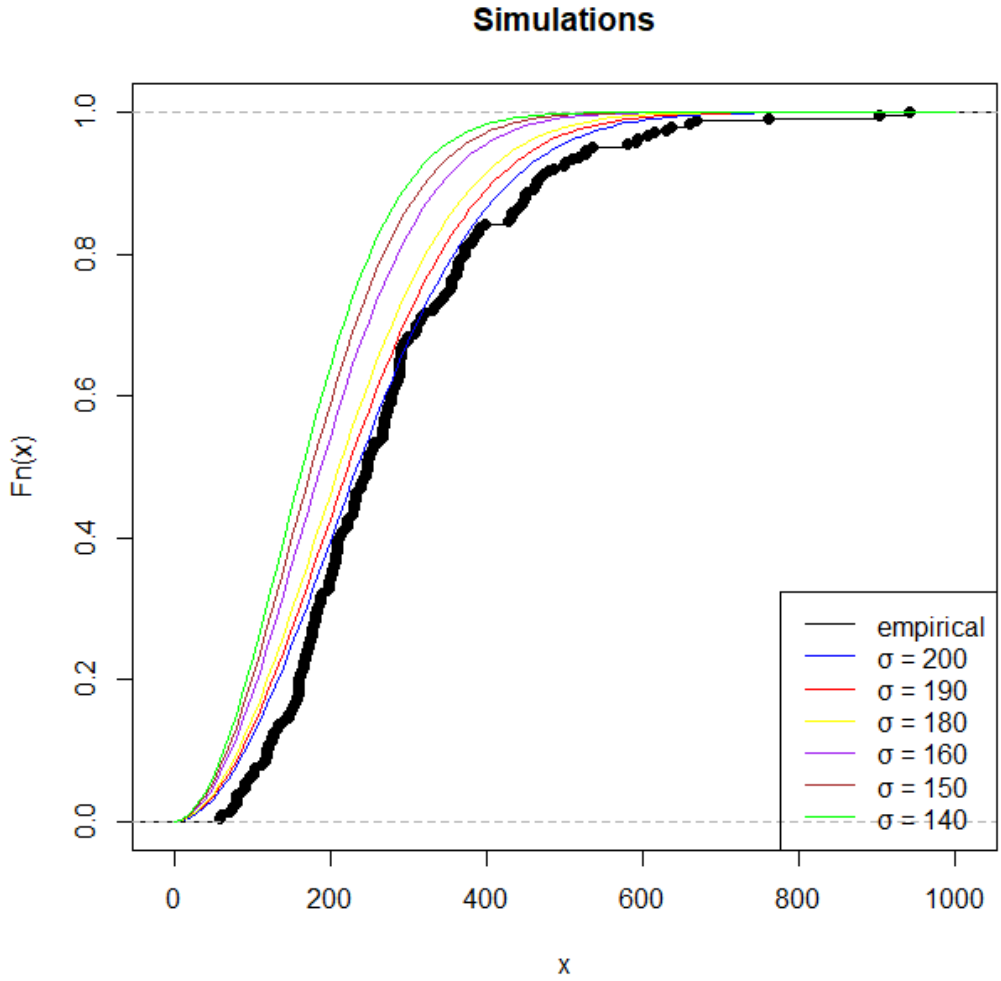


Figure 114

We understand that for σ=200 the empirical and the theoretical curve is close, not too close.

Now we will find the $\hat{\sigma}$ of the data.

We found that $\hat{\sigma} = 221.016$

We were close enough, because we found that the value 200 is close to the empirical approximately. But we find better empirical parameter from the MLE estimator.

Also we found that the velocity of Non-Gaussian have higher velocity dispersions than Gaussian groups. Also the curve of Non-Gaussian is sharper than the Gaussian.

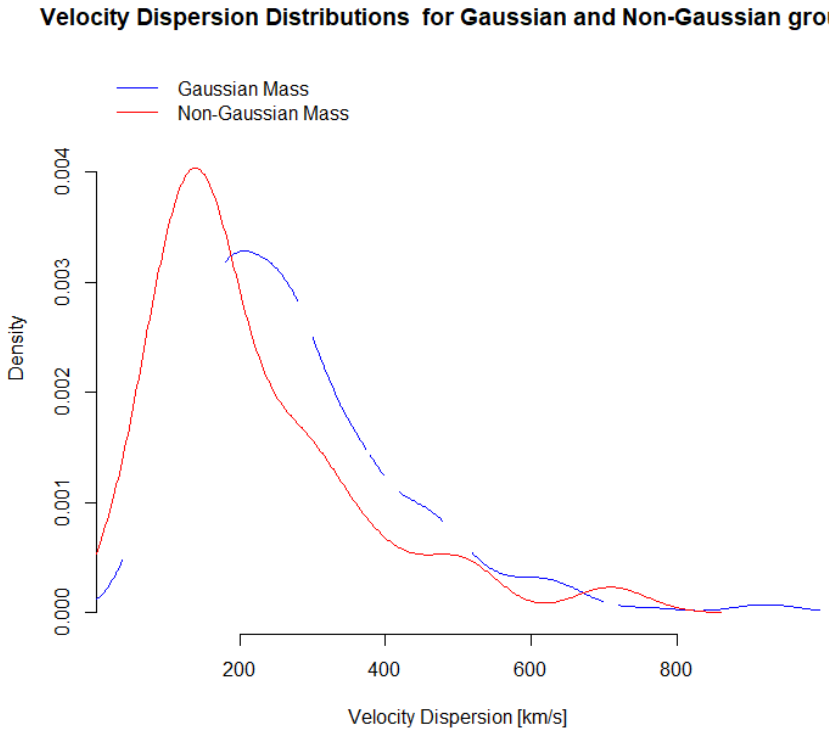


Figure 115 We see that the Non-Gaussian (red) curve is more sharper than Gaussian (blue) curve

We demonstrate also the Upper and Lower Area for various values of $\hat{\sigma}$.

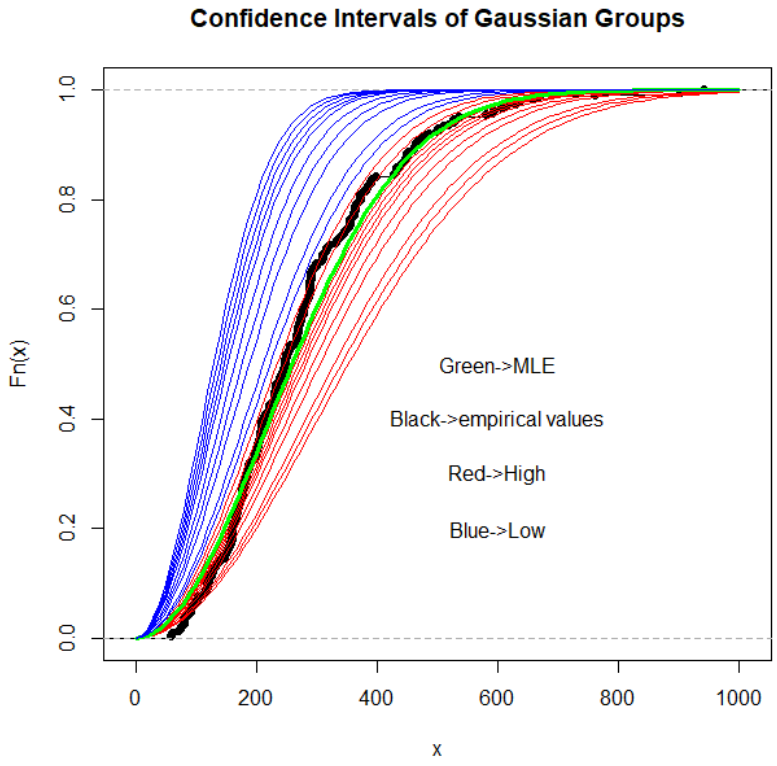


Figure 116

Next we have done the same work in order to find the empirical parameter of Rayleigh for the Non-Gaussian groups and all the groups of Low contrast sample. We first found the $\hat{\sigma}_{non-gaus} = \mathbf{184.2705}$ and $\hat{\sigma}_{all} = \mathbf{216.7951}$. We demonstrate and Upper and Lower Area for various values of $\hat{\sigma}$.

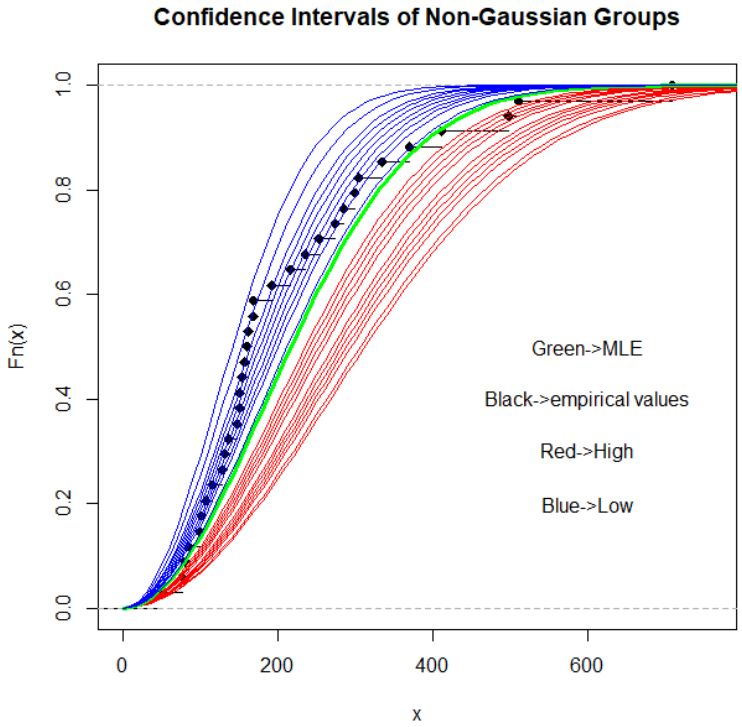


Figure 117 Confidence Intervals of Non-Gaussian Groups

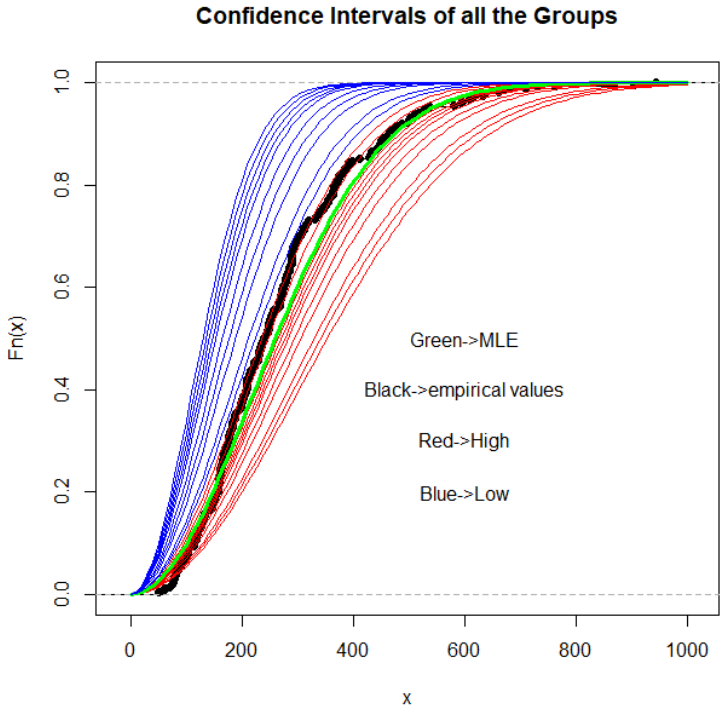


Figure 118 Confidence Intervals of all the Groups

Also we can observe the Law of Hubble in our data.

Firstly, we represent two (2) plots for Gaussian and Non-Gaussian groups.

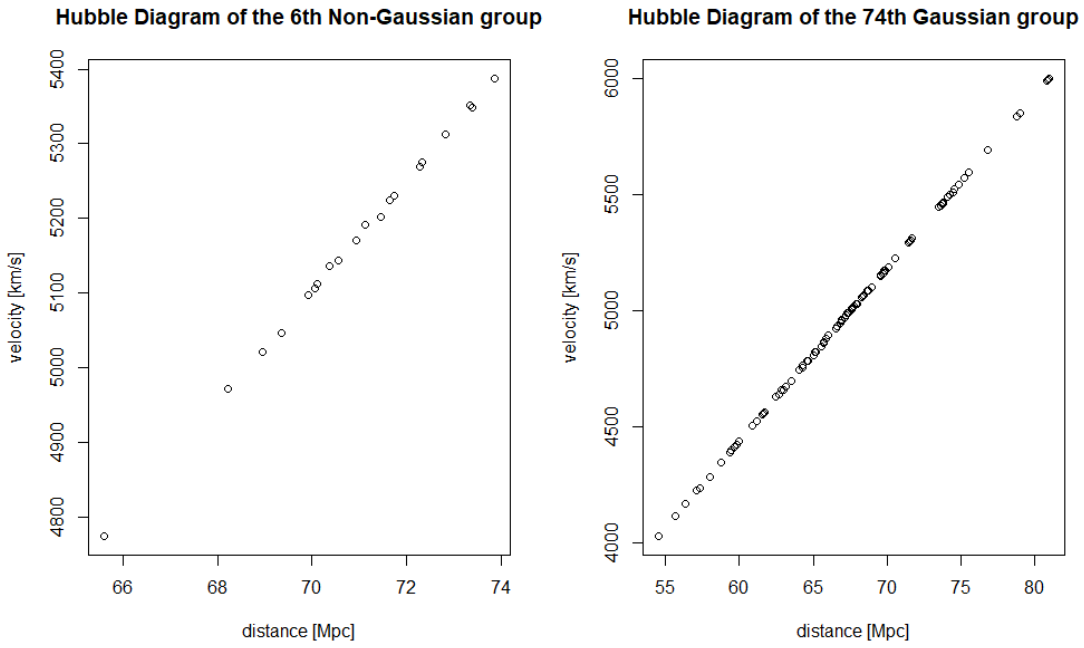


Figure 119 We observe that all the galaxies of the two groups are distributed in a line .

Also, we have plotted the Low and High contrast galaxies.

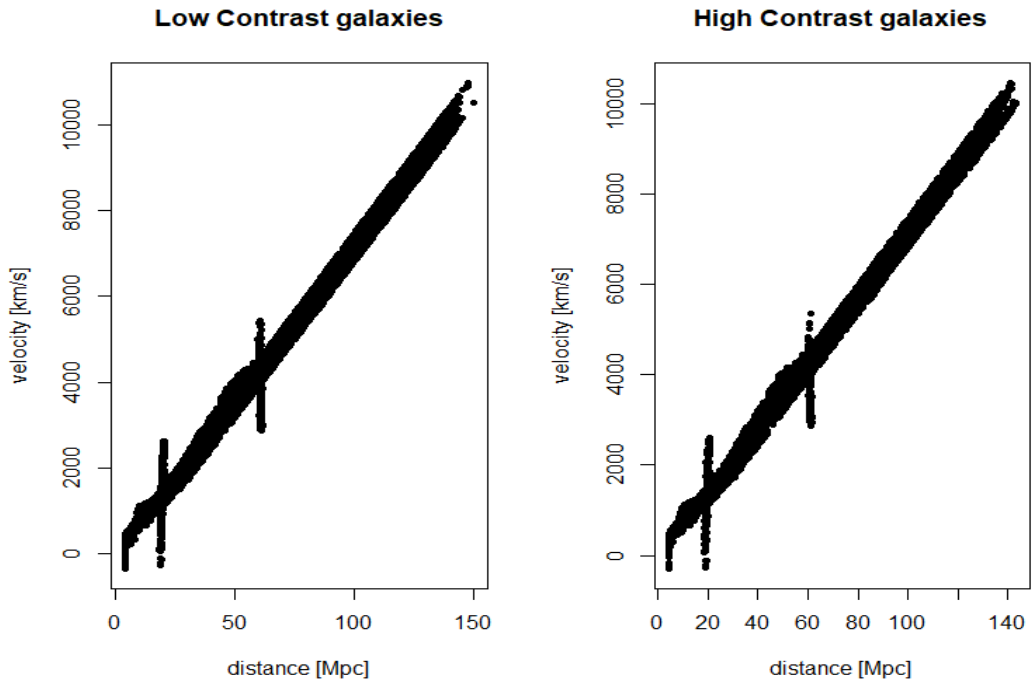


Figure 120 It is obvious that Hubble's Law is verified

On the other side, we have plotted again the coordinates of Non-Gaussian, Gaussian Low Contrast and High Contrast groups using contours. The space diagrams are shown below.

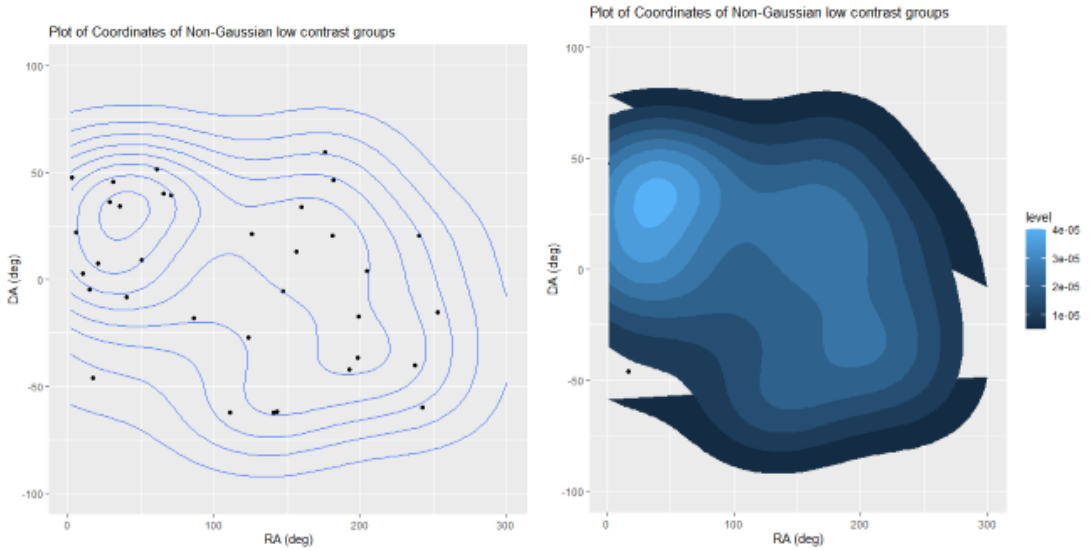


Figure 121 Plots of Coordinates of Non-Gaussian low contrast groups

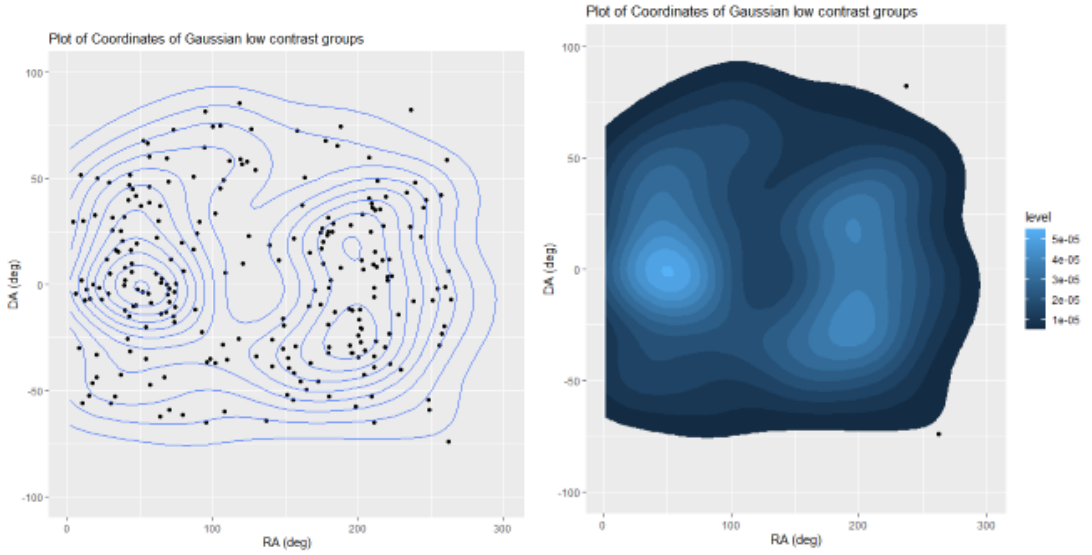


Figure 122 Plots of Coordinates of Gaussian low contrast groups

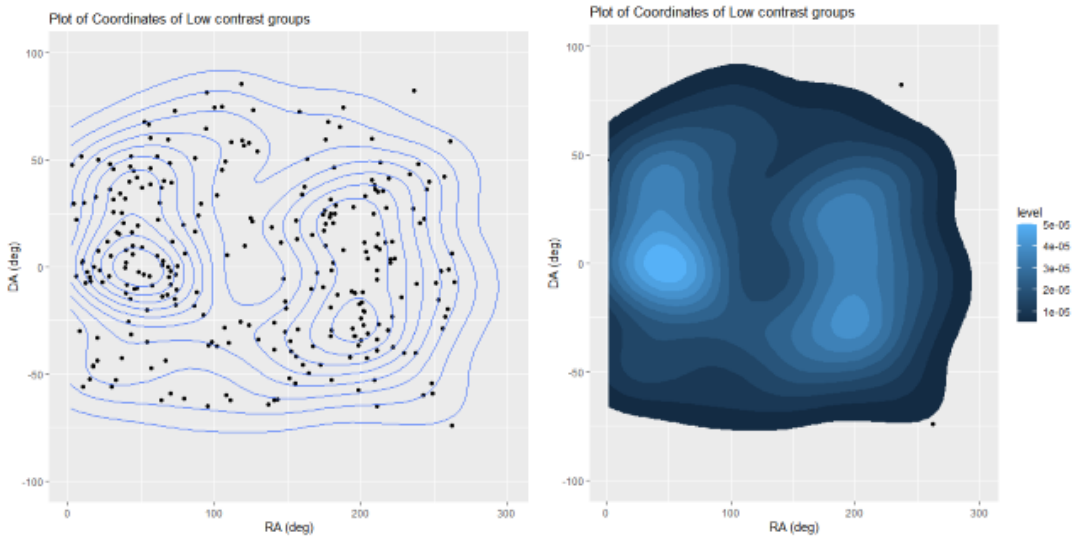


Figure 123 Plots of Coordinates of Low contrast groups

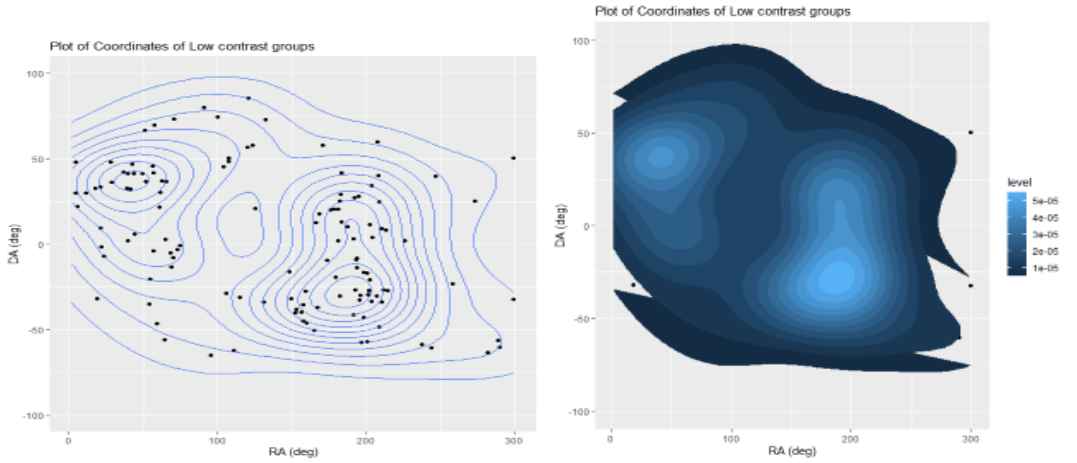


Figure 124 Plots of Coordinates of High contrast groups

We conclude that our results reinforce the idea of NG systems as complex structures. The Gaussian galaxies are extensively concentrated in the phase space diagram. The Non-Gaussian galaxies are less concentrated with less tight density contour level. These features suggest that Non-Gaussian systems are distinct, and dynamically younger than Gaussian groups, which agrees well with the results of Ribeiro, Lopes & Trevisan (2010). This scenario is also illustrated in Figure 125, where we show the stacked G and NG groups, and the stacked modes. Galaxies in these composite groups have distances and line-of-sight velocities with respect to the centers normalized by R200 and σ, respectively. We can approve the above with the help of the A.L.B. Ribeiro P.A.A. Lopes and M. Trevisan survey.

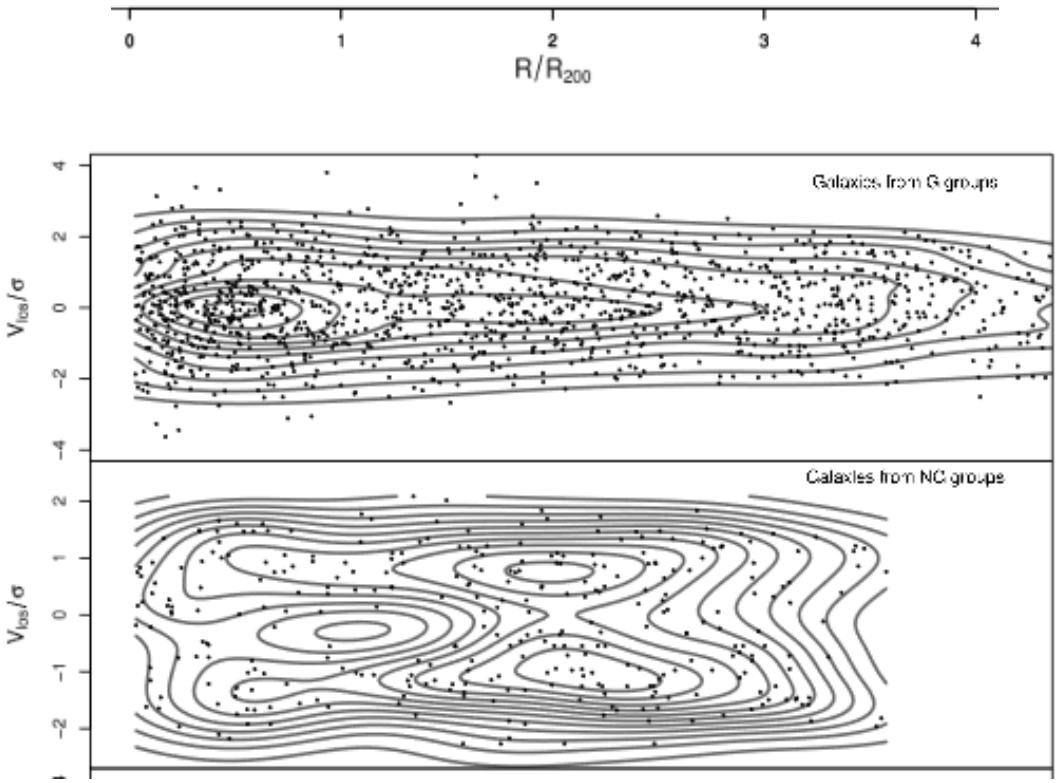


Figure 125 Phase space diagrams for a typical Gaussian (upper box), non-Gaussian (middle box) systems. Density contours indicate galaxy concentration across the diagrams. Distances to the center of the systems are normalized by R_{200}. Radial velocities are subtracted from the median velocity and divided by the velocity dispersion of the groups.

5 Univariate distributions of K-band magnitudes

Here we compare the univariate distributions of K-band magnitudes of globular clusters in the Milky Way Galaxy (MWG) and Andromeda Galaxy (M 31). The distribution summaries quickly show a large offset between the distributions: the MWG values are intrinsic (absolute) magnitudes while the M 31 values are observed (apparent) magnitudes. The difference between them is called the distance modulus of M 31. We calculate one estimator of the distance modulus using the Hodges–Lehmann (HL) estimator for shift. In base R, the HL estimator is built into the Wilcoxon rank sum test function. Note how we assign the result of a calculation to a new variable (e.g. DMmn for the distance modulus based on the mean) and then display the result on the R console by stating the variable.

We apply the Wilcox normality test.

```
wilcox.test(K2, K1, conf.int=T)

Wilcoxon rank sum test with continuity correction

data:  K2 and K1

W = 29160, p-value < 2.2e-16

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

 24.52993 25.25604

sample estimates:

difference in location

        24.90017
```

We can now offset the M 31 magnitudes by the estimated DM to place them on an absolute magnitude scale, and inquire into the similarity or differences in the shapes of K globular cluster luminosity functions of the Milky Way and M 31 samples. R has the function ecdf to construct a univariate empirical cumulative distribution function. A plot of confidence bands around the e.d.f. based on the Kolmogorov–Smirnov statistic is available in the CRAN package sfsmisc as function ecdf.ksCI. The R function quantile produces sample quantiles with options of nine algorithms for interpolation when small samples are considered. There is little mathematical guidance regarding the best method. The function qq makes a quantile-quantile (Q-Q) plot for two samples. A common variant is the one-sample plot made with qqnorm which compares a dataset to the normal distribution. A straight line in a Q-Q plot indicates agreement between the distributions. The first figure comparing the e.d.f.'s shows that the M 31 distribution is substantially narrower than the MWG distribution. (Inconveniently, the abscissa scale has bright clusters on the left; the axis direction can usually be reversed using the plot.window function but this option does not work with step function plots.) Of course, this finding may not reflect the underlying distribution; for example, the M 31

measurements may be truncated at faint magnitudes by observational limitations. For our purposes here, we will continue with the comparison of the datasets without scientific interpretation. The second figure shows Q-Q plots comparing the MWG sample with the M 31 sample and both empirical distributions with a normal distribution. We see that the two galaxy samples are systematically different; the MWG cluster distribution is quite close to a Gaussian while the M 31 distribution is deficient in the lower tail.

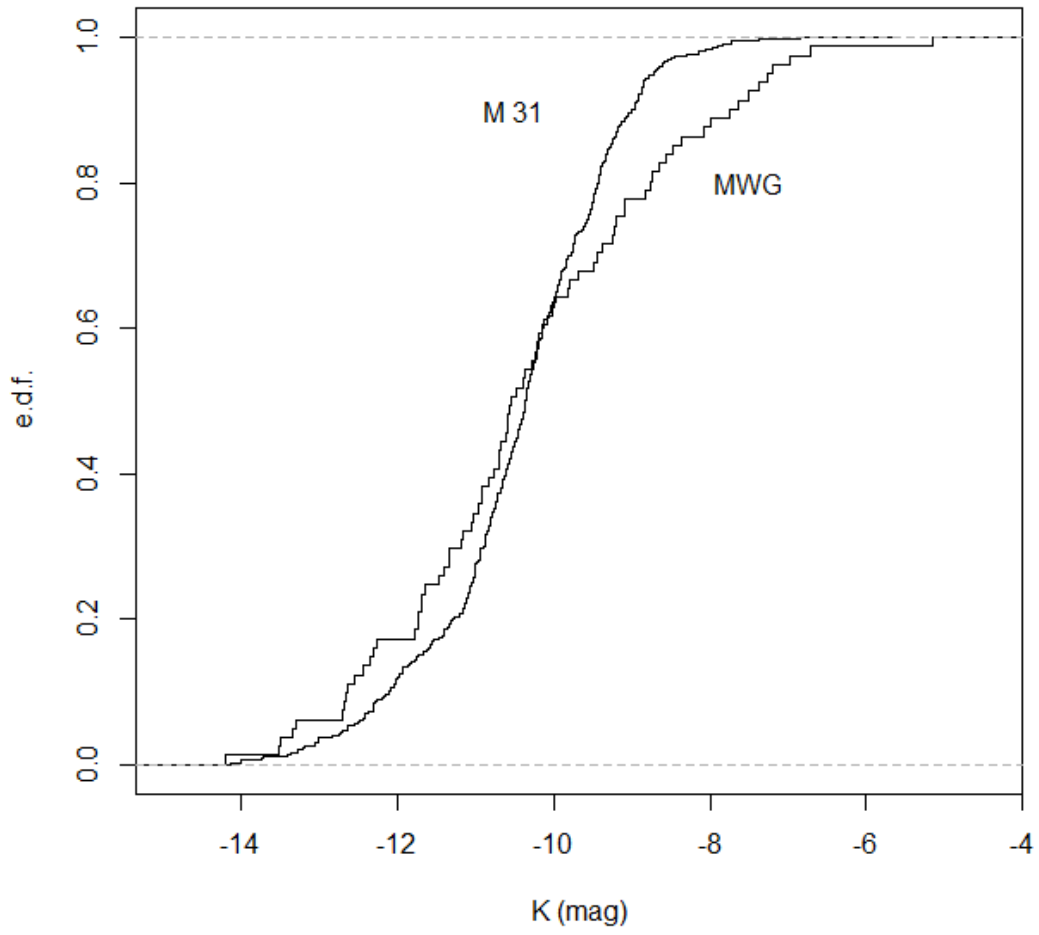


Figure 126 Empirical distribution functions for the K-band magnitudes of globular clusters in the Milky Way Galaxy and the Andromeda Galaxy (M31). The MWG is near to Normal curve, while the M31 is deficient in lower tail.

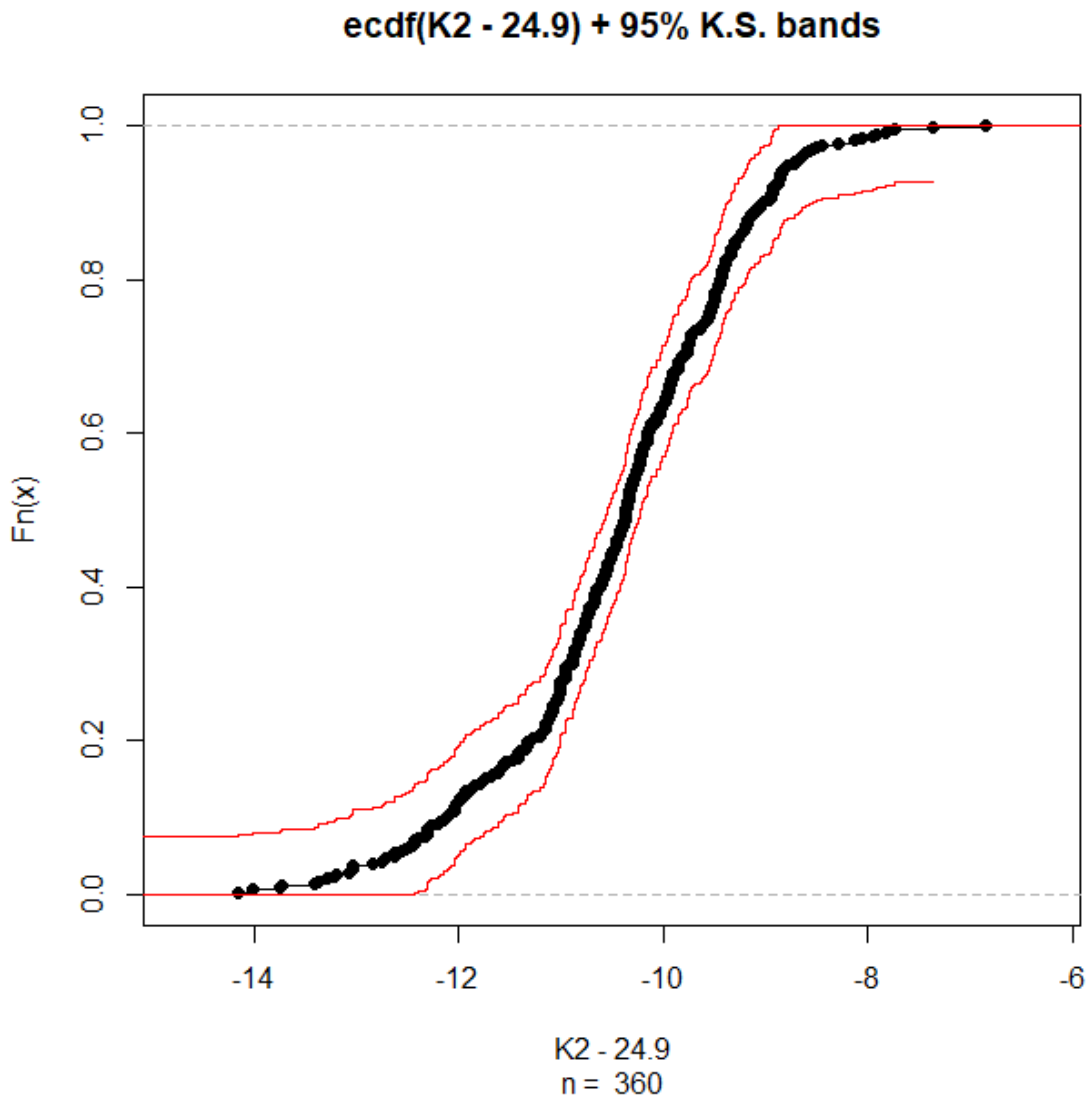


Figure 127 Confidence bands around the e.d.f. based on the Kolmogorov–Smirnov statistic

Now we want to test the normality of our Milky Way Galaxy and the Andromeda galaxy. The Figure99 shows that our galaxy is close to normal than the Andromeda. We apply the following tests:

- Cramer-von Mises
- Anderson-Darling

- Cramer-von Mises normality test

data: K1 (MWG)

W = 0.052798, p-value = 0.4662

- Cramer-von Mises normality test

data: K2 (M31)

W = 0.28935, p-value = 0.000438

- Anderson-Darling normality test

data: K1 (MWG)

A = 0.30257, p-value = 0.5674

- Anderson-Darling normality test

data: K2 (M31)

A = 1.7939, p-value = 0.0001351

We understand that our galaxy is normal distributed with p-value = 0.4662 for the Cramer-von Mises and 0.5674 for the Anderson – Darling. The Andromeda is not normally distributed because p-value < 0.1, 0.05, 0.001.

We can understand this easily from the QQ – plots.

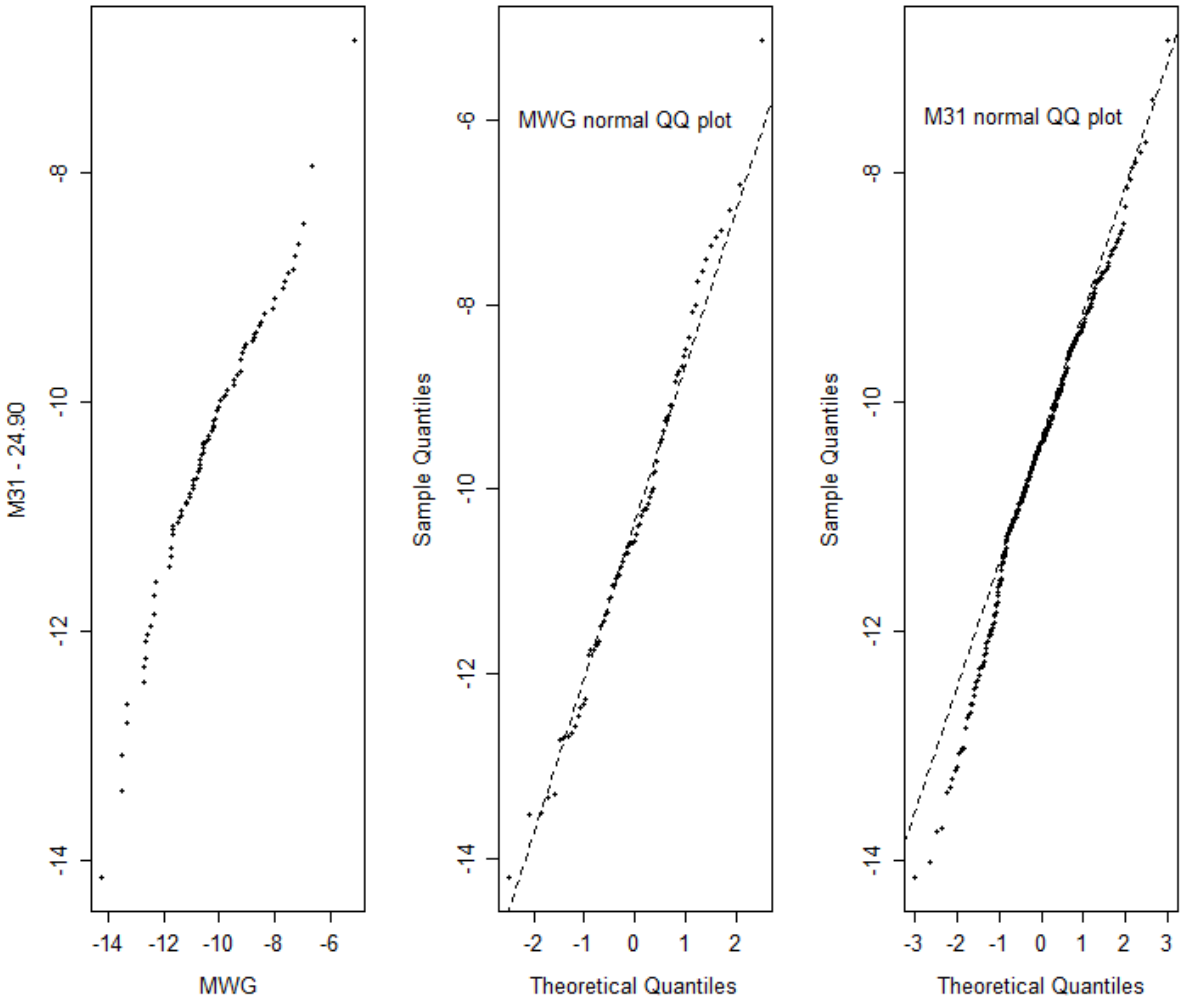


Figure 128 Quantile-quantile plots for comparing univariate empirical distributions. Left: Comparison of Milky Way and M31 globular cluster distributions. Middle and Right: Normal Q-Q plot comparing the Milky Way and M31 globular cluster distributions to a Gaussian.

6 COMBO-17 galaxy data-set

Dataset

Wolf et al. (2004) provide the first public catalog of a large dataset (63,501 objects) with brightness measurements in 17 bands in the visible band. (Note that the Sloan Digital Sky Survey provides a much larger dataset of 10^8 objects with measurements in 5 bands.) We provide here a subset of their catalog with 65 columns of information on 3,462 galaxies. These are objects in the Chandra Deep Field South field which Wolf and colleagues have classified as `Galaxies'. The column headings are formally described in their Table 3, and the columns we provide are summarized here with brief commentary:

Col 1: Nr, object number

Col 2-3: Total R (red band) magnitude and its error. This was the band at which the basic catalog was constructed. Magnitudes are inverted logarithmic measures of brightness. A galaxy with R=21 is 100-times brighter than one with R=26. The error is the standard deviation derived from detailed knowledge of the measurement process. This dataset is an excellent example of astronomical datasets where each variable is accompanied by heteroscedastic measurement errors of known variances.

Col 4-5: ApDRmag is the difference between the total and aperture magnitude in the R band. This is a rough measure of the size of the galaxy in the image where ApDRmag=0 corresponds to a point source. Negative values are not physically meaningful. mu_max is the central surface brightness of the object in the R band. The difference between Rmag and mu_max should also be an indicator of galaxy size.

Col 6-9: Mcz and MCzml are two redshift estimates. Mcz is the preferred value. e.Mcz is its estimated error, and chi2red is the reduced chi-squared value of the least-squares fit of the 17-band magnitudes to the best-fit template galaxy spectrum. Galaxies with large e.Mcz or chi2red might be omitted as unreliable.

Col 10-29: These give the absolute magnitudes (i.e. intrinsic luminosities) of the galaxy in 10 bands, with their measurement errors. They are based on the measured magnitudes and the redshifts, and represent the intrinsic luminosities of the galaxies; a galaxy with M=-15 is 100-times less luminous than one with M=-20. These magnitudes are not all independent of each other's, but they are important for representing intrinsic properties of the galaxies. Below is one of several redshift-stratified plots of the B-band absolute magnitude (abscissa) against the difference of magnitude (i.e. ratio of luminosities) between the 2800A ultraviolet and blue band, which is a sensitive indicator of star formation. A redshift-dependent bimodal distribution is seen.

Col 30-55: Observed brightnesses in 13 bands in sequence from 420 nm in the ultraviolet to 915 nm in the far red. These are given in linear variables with units of photon flux densities, photons/m^2/s/nm. Again, each measurement is accompanied by a measurement error which can be used to distinguish measurement from intrinsic dispersions in the distributions.

Col 56-65: Observed brightnesses in 5 traditional broad spectral bands, UBVRI. These are largely redundant with the 13 bands in the previous columns.

We illustrate unsupervised clustering algorithms using a two-dimensional color–magnitude diagram constructed from the COMBO-17 (Classifying Objects by Medium-Band Observations in 17 Filters) photometric survey of normal galaxies (Wolf et al. 2003). The R script below starts with which function to filter the dataset, keeping only low-redshift galaxies with $z < 0.3$ and removing a few points with bad data values. Most of the original 65 variables are ignored, and we keep only the galaxy absolute magnitude in the blue band, MB, and the ultraviolet-to-blue color index, M280−MB. The resulting color–magnitude diagram (Figure 112, left panel) shows the well-known concentrations of luminous red galaxies around (MB,M280 −MB) (−16,−0.2) and fainter blue galaxies around (−13,−0.9).We represent the color magnitude diagram for low redshift COMBO -17 galaxies

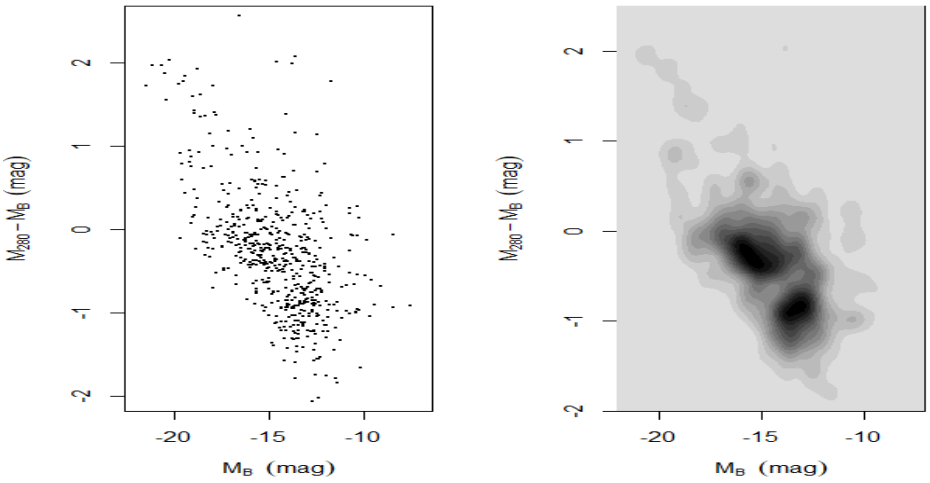


Figure 130 A magnitude diagram of nearby, normal galaxies from theCOMBO-17 survey. The left panel shows individual galaxies, and the right panel shows a kernel smooth estimator highlighting the red sequence (lower right) and blue sequence (upper left) concentrations. The gray scale is linear in surface density of points in the plot.

The blue galaxies are spirals and irregular galaxies that have experienced recent active star formation, while the red galaxies are mostly elliptical that have only older stars formed early in the Universe's history. Note that many galaxies have properties distributed around the major concentrations. .For the nonparametric procedures where we assume Euclidean distances between points in the 2-space.

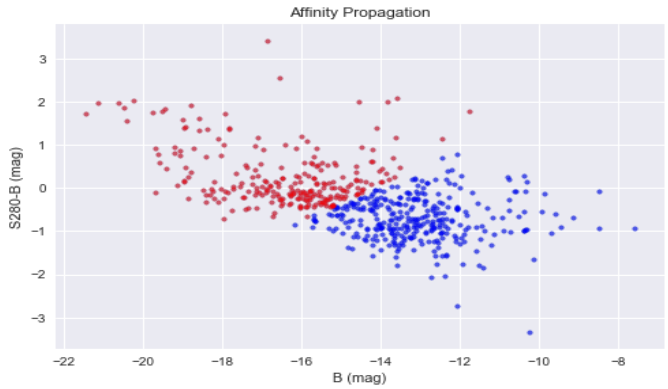


Figure 131 A color-magnitude diagram of nearby, normal galaxies from the COMBO-17 survey.

There is a package in R called "MASS". This data set considers physical information on velocities (km/second) for 82 galaxies reported by Roeder (1990). These are drawn from six well-separated conic sections of the Corona Borealis region. So, we will analyze the velocity distribution for Corona Borealis region

The constellation Corona Borealis, the Northern Crown, is visible in the northern hemisphere in the spring and summer. It can be seen at latitudes between 90 degrees and -50 degrees. It is a small constellation covering 179 square degrees of the sky. It ranks 73rd in size among the 88 constellations in the night sky. It is bordered by Boötes to the north and west, Serpens Caput to the south, and Hercules to the east.

Corona Borealis is one of the 48 constellations listed by the Greek astronomer Ptolemy in the second century. Its name means "northern crown" in Latin. This is an ancient constellation that has its roots in many cultures. It has been depicted as a circle of elders, an eagle's nest, and a bear's den. In Celtic mythology, it was known as Caer Arianrhod, the Castle of Arianrhod. It represented the home of the Lady Arianrhod. In Greek mythology, it represented the crown that Ariadne, the daughter of King Minos of Crete, wore at her wedding. Her ball of thread helped Theseus defeat the Minotaur and find his way out of the labyrinth. Theseus gave her the crown when they married. The crown was created by the supreme goldsmith of the gods, Hephaestus.

Pronunciation: (koh-ROH-nuh BOR-ee-AL-is)
Abbreviation: CrB **Genitive:** Coronae Borealis
Right Ascension: 16 hours **Declination:** 30 degrees
Area in Square Degrees: 179
Crosses Meridian: 9 PM, June 30
Visible Between Latitudes: 90 and -50 degrees

Corona Borealis is a relatively dim constellation with only three stars brighter than magnitude 3. The brightest is Alphecca, with a visual magnitude of 2.23. It is a binary star system located approximately 75 light years from Earth. The second brightest star is Nusakan with a magnitude of only 3.68. It is also a binary star system. Nusakan lies 114 light years from our solar system. Gamma Coronae Borealis is the third brightest star with a magnitude of only 3.84. It, too, is a binary star system. The two stars orbit each other every 93 years and are about as far apart as our Sun and the planet Neptune.

Corona Borealis contains no Messier objects but does contain a few notable deep-sky objects that can only be seen in large telescopes. The Corona Borealis Galaxy Cluster is a dense cluster of about 400 galaxies that spans an area of about one degree of the sky. This constellation is also the home of the Blaze Star, also known as T Coronae Borealis. It is a recurring nova that normally has a visual magnitude of about 10. During its outbursts, it can brighten to a magnitude of as high as 2.

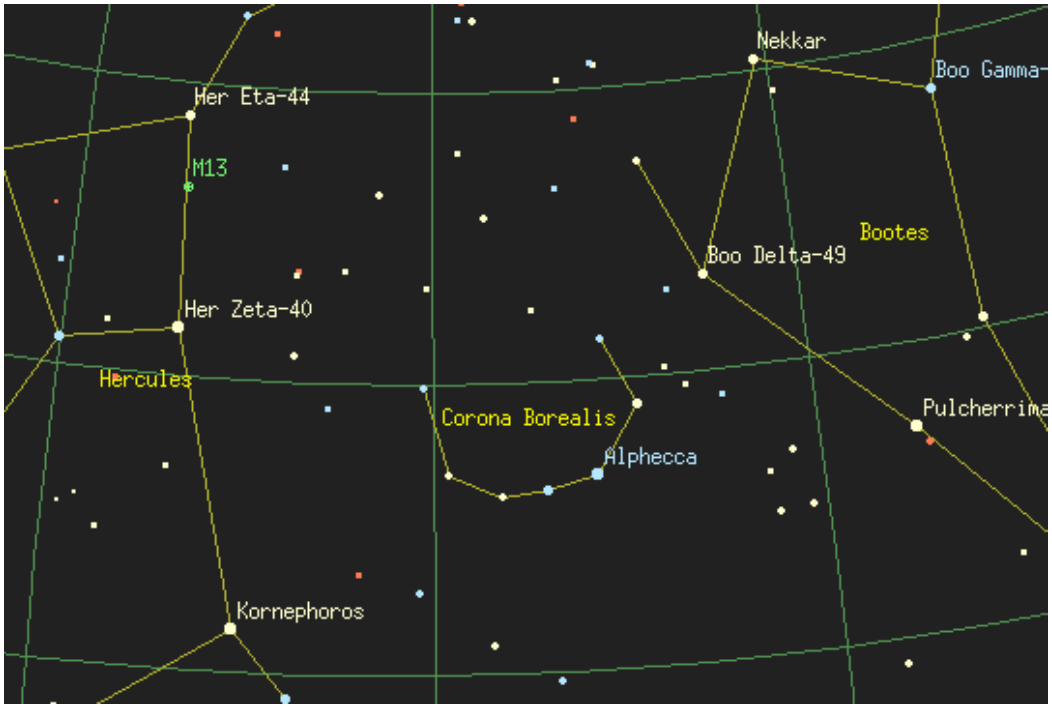


Figure 132 Corona Borealis region

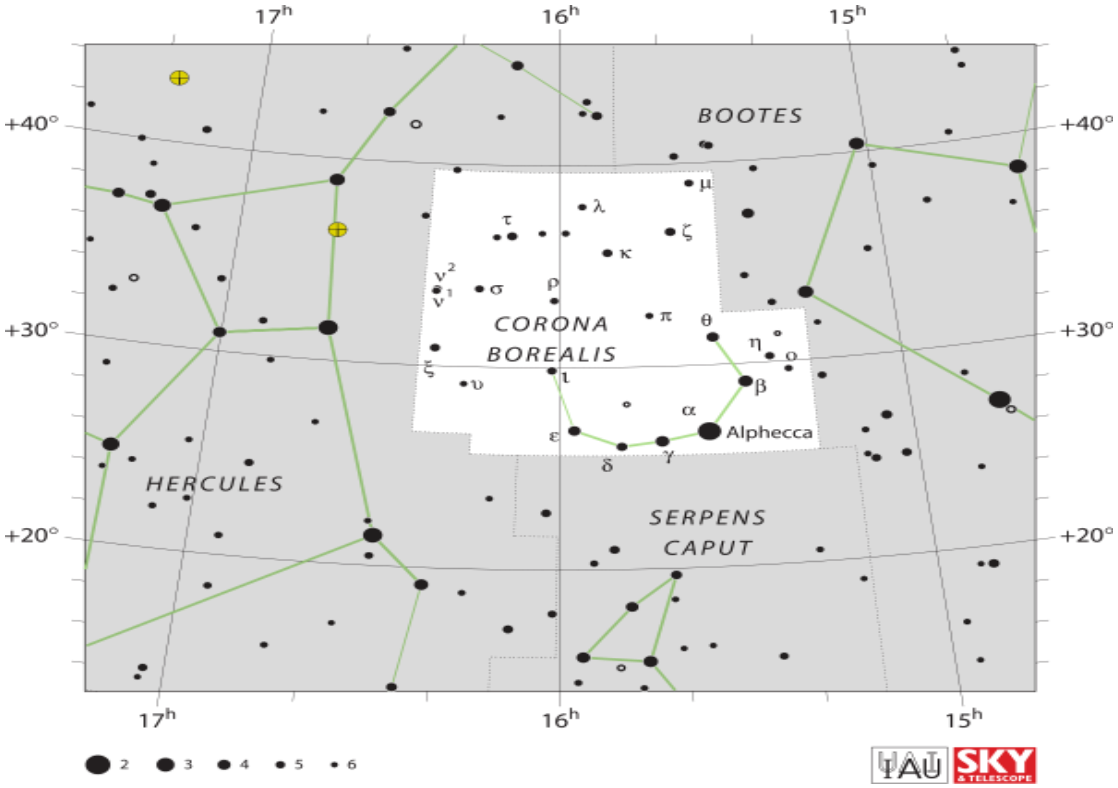


Figure 133 Corona Borealis region

We have plotted the velocities of the galaxies. We understand that there are many galaxies having the same velocities.

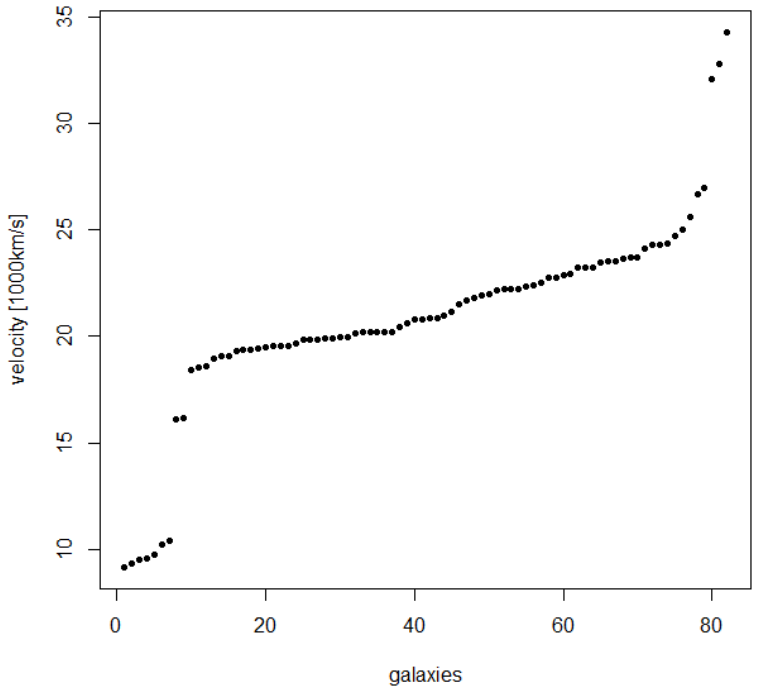


Figure 134

Next we represent the velocity distribution curve of these galaxies

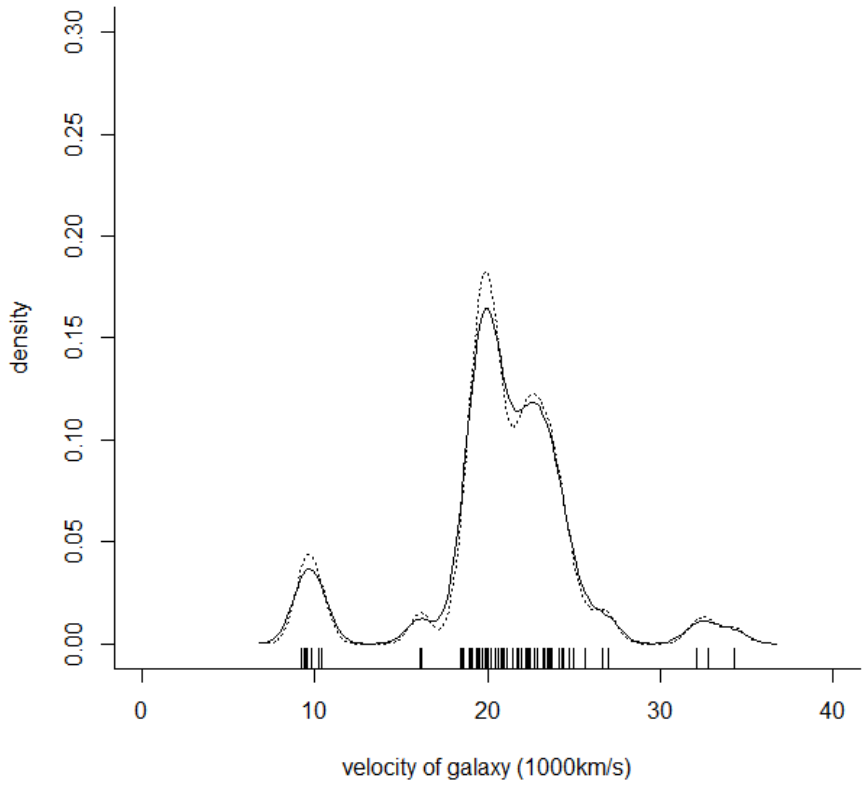


Figure 135

As we said above this curve is sharp in the middle but we have 4-5 peaks. So we hypothesize that Corona Borealis include many types of galaxies. But this super cluster is not normally distributed. For that reason we conduct the Anderson-Darling normality test. But the p_{value} of this test is 9.037e-10. So the super cluster is not normally distributed. These galaxies are not very bright in the night sky.

Object	Designation	Name / Meaning	Object Type	V Mag
1	Alphecca	"Broken Ring"	Binary Star System	2.23
2	Nusakan	"The Two Series"	Binary Star System	3.68
3	Gamma Coronae Borealis	N/A	Binary Star System	3.84
4	Epsilon Coronae Borealis	N/A	Multiple Star System	4.13
5	Theta Coronae Borealis	N/A	Binary Star System	4.13
6	Delta Coronae Borealis	N/A	Yellow Giant Star	4.63
7	Iota Coronae Borealis	N/A	Binary Star System	4.98

Figure 136 Corona Borealis map

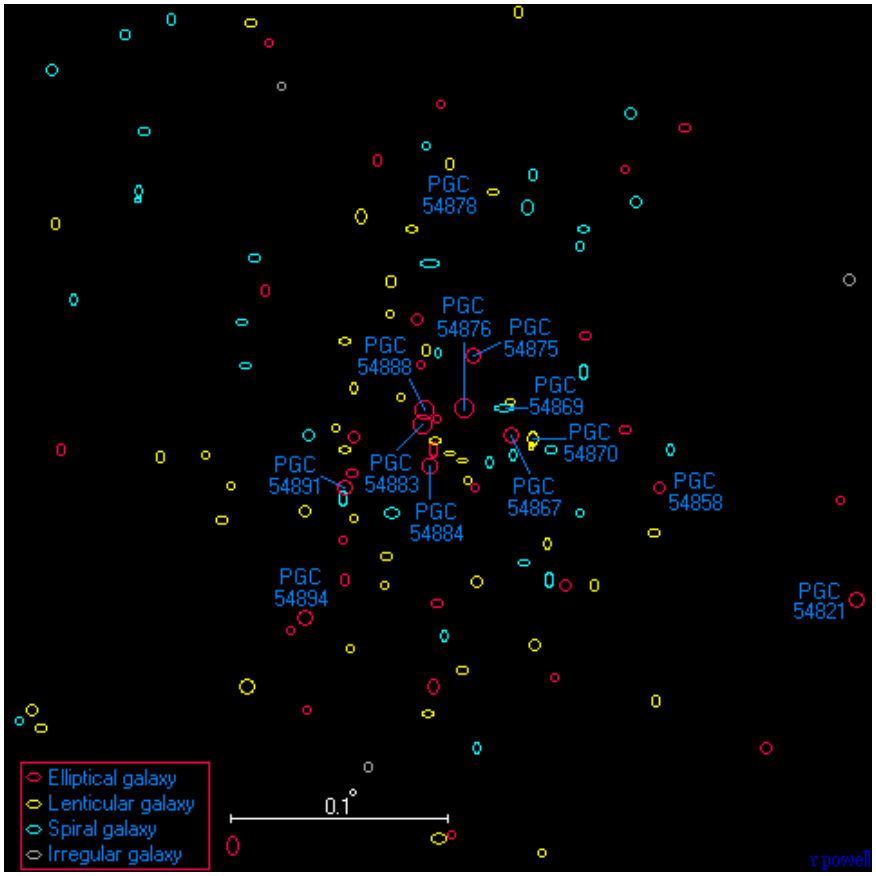


Figure 137 This is a map of the A2065 cluster. This map shows 119 of the brightest galaxies in this cluster. A lot of the galaxies in this cluster have not been accurately classified, so some of the classifications used in this map are probably wrong, although it is clearly a cluster which contains a wide variety of different galaxy types.

We observe many elliptical galaxies. For that reason the curve of the velocity distribution is sharp in the middle. Corona Borealis also include spirals, lenticular and irregular galaxies. For that reason we have four (4) peaks in the Figure102. We conclude that most of the galaxies are in early type. So, we conclude that Non-Gaussian groups that include many elliptical galaxies may be in early type.

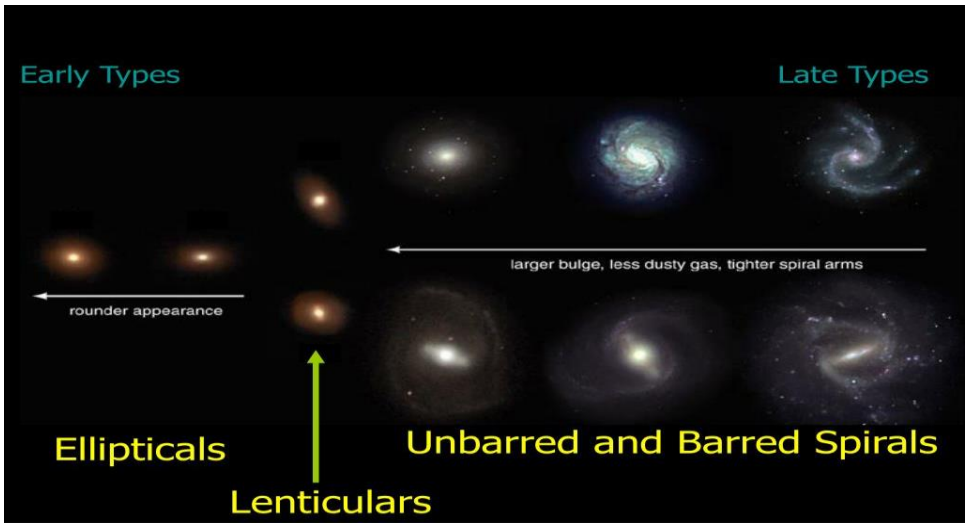


Figure 138 The evolution of galaxies

Here we study the Morphology of galaxies. The data used Garcia, A. M. for his investigation helped to do this study. We present a whole sky catalog of nearby groups of galaxies taken from the Lyon-Meudon Extragalactic Database. From the 78000 objects in the database, we extracted a sample of 6392 galaxies, complete up to the limiting apparent magnitude Bo = 14.0. Moreover, in order to consider solely the galaxies of the local universe, all the selected galaxies have a known recession velocity smaller than 5500 km/s. Two methods were used in group construction: a Huchra-Geller-derived percolation method and a Tully-derived hierarchical method. Each method gave us one catalog. These were then compared and synthesized to obtain a single catalog containing the most reliable groups. There are 485 groups of a least three members in the final catalog. The sample used to study the galaxy group membership is described. It contains 23490 galaxies with weighted mean distance moduli derived from both radial velocity through the Hubble law (Ho= 75 km /s*Mpc) and distance criteria (Maximum velocity rotation and Luminosity Index). This sample is complete up to the limiting apparent magnitude B = 14.0.

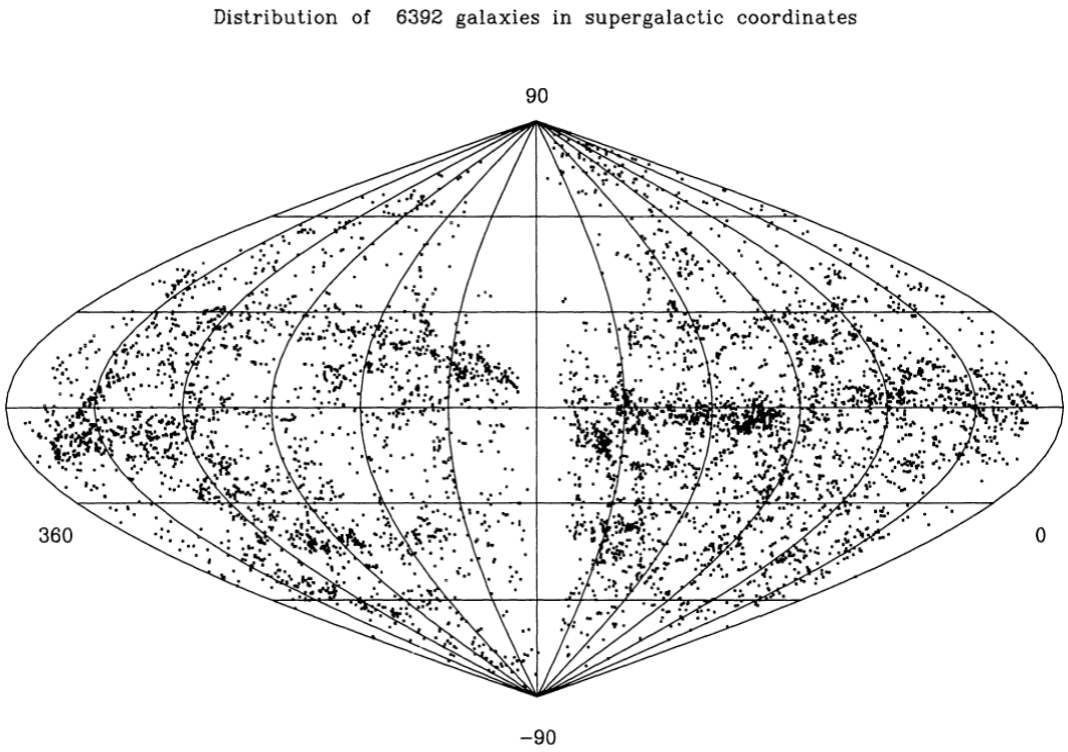


Figure 139

We applied Anderson-Darling and Cramer-von Mises normality tests for the group galaxies. We found that only the 285 is Non-Gaussian. There is an explanation for this. Below, we represent the table of number and percentage of each morphological type of galaxy.

Morphology of Galaxy	Number	Percentage (%)
Elliptical	322	8.187135%
Spiral	2369	60.23392%
Lenticular	640	16.27257%
Irregular	318	8.085431%

So, we observe that Spiral Galaxies hold the 60.23392% of the sample. So, if a group galaxy consists mainly of spiral galaxies then it is normally distributed. (In this sample we set the significance level to be 0.1 or 10%). But if a group has many Irregular, Lenticular Galaxies, then the normality exists or not?

On the other side we found a Non-Gaussian group (285 group galaxy). We analyzed this group and we compare the velocities with the velocities of a Gaussian group. In Figure102 we represent the velocity distribution of the two (2) groups.

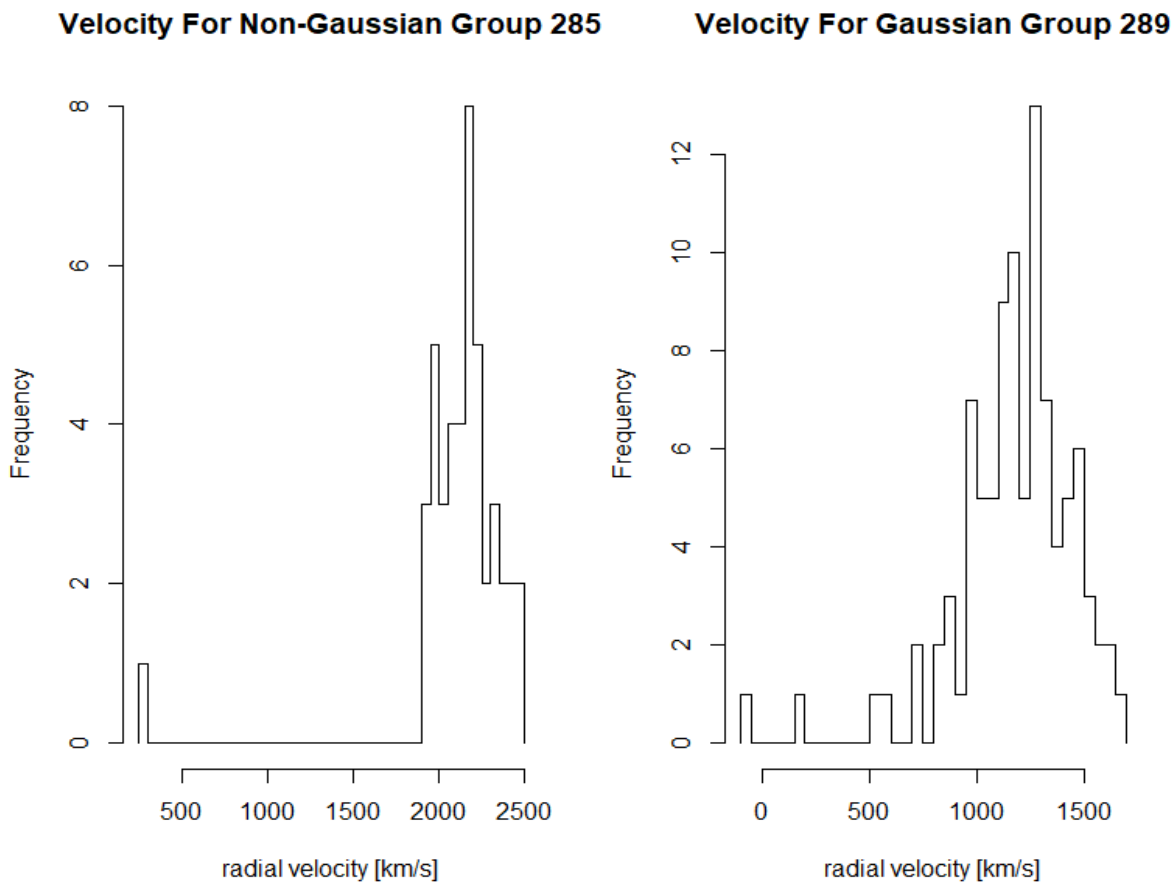


Figure 140 The difference is obvious

Also we have plotted the Elliptical and Spiral Galaxies of the data in order to compare them.

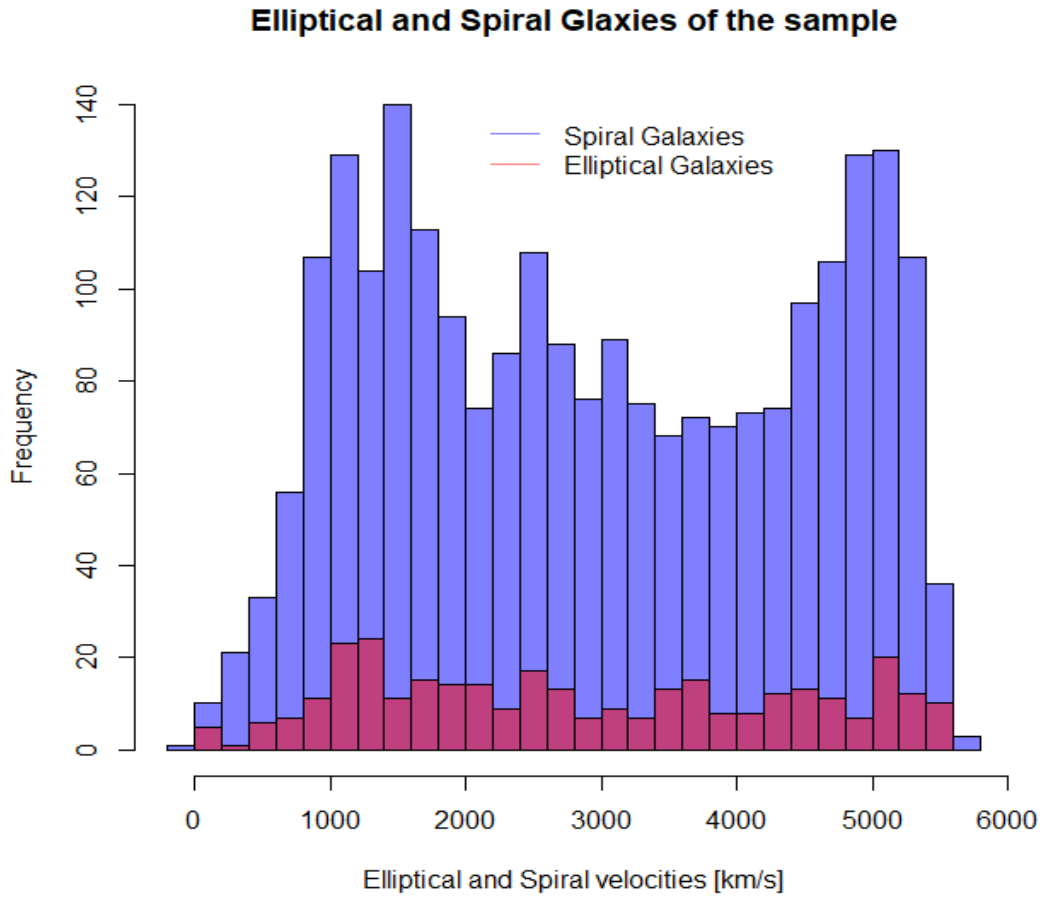


Figure 141 We observe that the data contains many galaxies in late type

Also, we observe that from the above Figure the velocities of each type of galaxy follow a Uniform distribution. Notice that it has been proved that the **velocity dispersion** of the groups follows a Rayleigh distribution and the velocities of **each** group comes from a normal distribution or not. Here, we will examine the Uniformity of the velocities of **each type** of galaxy for all the sample (not of each group).

We perform Rayleigh and Wilcoxon tests for uniformity for the three (3) types of galaxies. We do not study the Irregular galaxies.

Rayleigh Test for Uniformity

Classical Test in S^1

The polar coordinates of S^1 yield $U_i = (cos\Theta_i, sin\Theta_i)'$ with random angles $\Theta_i \in [0,2\pi), i = 1,2, \dots, n$. In the sequel, we inspect several well-known uniformity tests for circular data. We work with the Rayleigh test.

The Rayleigh test is based on a simple fact:

When $U \sim Uniform$, then $E[U] = 0$ or, equivalently, $||E[U]||^2 = 0$. This provides a simple way of testing H_0 by using Rayleigh's statistic

$$R_n := 2n\left|\left|\bar{U}\right|\right| = \frac{2}{n}[\left(\sum_{i=1}^{n} cos\Theta_i\right)^2 + (\sum_{i=1}^{n} sin\Theta_i)^2]$$

where $\bar{U} = \frac{\sum_{i=1}^{n} U}{n}$. The asymptotic distribution of R_n under H_0 is a χ_2^2, a chi-squared distribution with 2 degrees of freedom. The Rayleigh test is the most powerful invariant test against the Cramer-von Mises. Note, however, that $E[U] = 0$ does not imply uniformity, and as a consequence non-unimodal alternatives with $E[U] \approx 0$ are likely not detectable with the Rayleigh test. So, we found the following:

Elliptical	Spiral	Lenticular
Rayleigh Test of Uniformity	Rayleigh Test of Uniformity	Rayleigh Test of Uniformity
General Unimodal Alternative	General Unimodal Alternative	General Unimodal Alternative
Test Statistic: 0.0263	Test Statistic: 0.0148	Test Statistic: 0.0456
P-value: 0.8006	P-value: 0.5934	P-value: 0.2642

Wilcoxon Test for Uniformity

We perform Monte Carlo Simulations with 30000 iterations for the Wilcoxon test for uniformity. We generate three uniform distributions because we have three types of galaxies. The length of the Uniform depends on the length of the galaxies (Elliptical, Spiral and Lenticular). The minimum value will be the minimum velocity value of each type of the galaxies and the maximum value will be the maximum velocity value of each type of the galaxies. We found the following:

We perform Monte Carlo Simulations (30000 iterations)	
	Percentage (%) of rejection with significance level α=0.05
Galaxy Type	Wilcoxon Test
Elliptical	2.543333%
Spiral	96.11333%
Lenticular	3.04%

We observe a difference between the two tests. The velocities of Elliptical and Lenticular galaxy of the sample come from a Uniform distribution and the velocities of Spiral do not come from a Uniform distribution. ($\alpha = 0.05$). This might happens, because the spiral galaxies contain the 60% of the sample.

Furthermore, we found that the velocities of the Elliptical Galaxies and Lenticular are higher than the velocities of Spiral Galaxies (We explain it below). We took the median and we found that:

Galaxy Type	Median Velocities
Elliptical	2669.5
Spiral	2859.9
Lenticular	2921.5
Irregular	1250.0

Even though the Median of Spiral Galaxies is bigger than the Median of the Elliptical, the velocities of Elliptical are bigger than Spiral, because we know that Elliptical contain ~ 8.18% of the sample. We observe also that the Lenticular Galaxies (~16.27257%) have large velocities. So, we are going to put the Elliptical and Lenticular Galaxies in the same group. We conclude that Elliptical and Lenticular Galaxies have larger velocities than Spiral Galaxies. The Irregular galaxies have the lowest velocities.

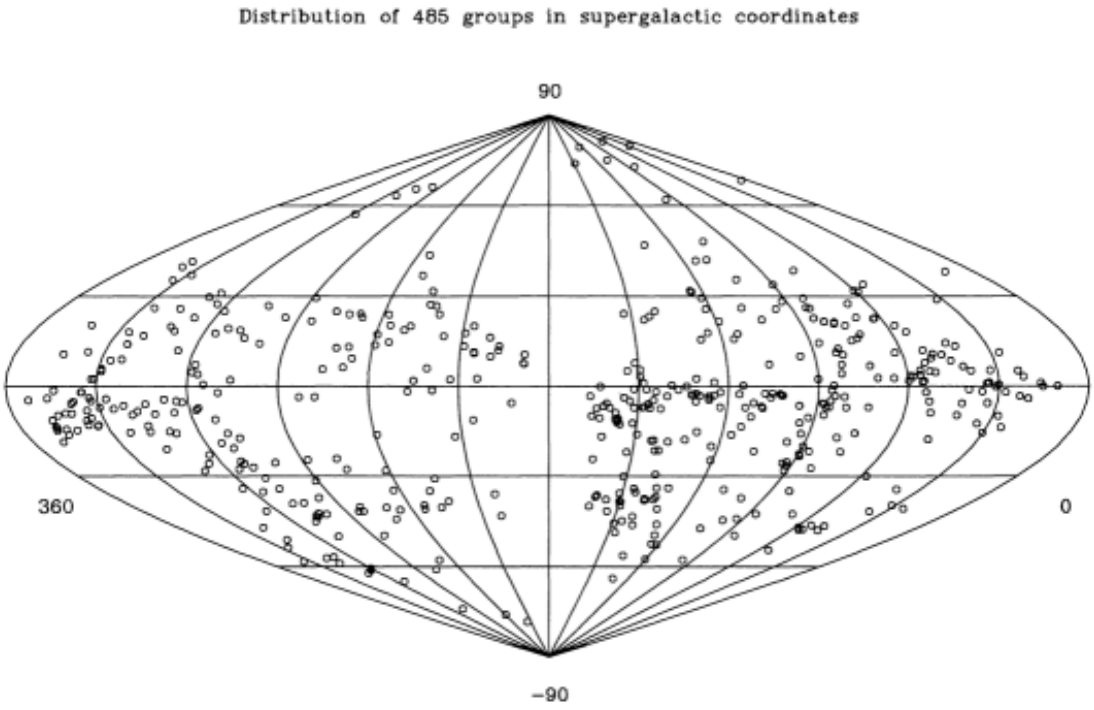


Figure 142

9. SUMMARY

Analysis of galaxy group dynamics requires the use of tools that are reliable and powerful for small sample sizes. We have applied three goodness-of-fit tests, the Pearson's χ^2, Lilliefors, Shapiro-Wilk, D'Agostino, A–D tests, to a subset of the 2MASS groups in order to determine which test can best classify galaxy group dynamics. Based on our initial application of the aforementioned tests and on the results of our Monte Carlo simulations and power studies, we conclude that the A–D test is the most reliable statistic to distinguish between relaxed (Gaussian) and dynamically complex (non-Gaussian) groups. The results of our Monte Carlo simulations for the Gapper Estimator and rms dispersion calculations indicate that for our sample size, the Gapper and rms estimator algorithm are almost the same. On the other hand for CNOC 2 (Hou et.al.) groups the Gapper is a more accurate estimate of the true velocity dispersion, which is in agreement with Beers et al. (1990).

For the following result we took the results of Hou.et.al.

We then apply the A–D test to all 2MASS groups with $n_{members} > 7$, using the mean velocity of the group members as the estimated μ and the intrinsic velocity dispersion as the estimated σ. The groups are then classified as being in either a relaxed (Gaussian) or complex (non-Gaussian) dynamical system, based on their computed significance levels. (here $\alpha = 0.1$) The results of our analysis indicate that 42 of the 274, or~15.32847%, of the sample of 2MASS groups were classified as non-Gaussian and 232 of the 274, or 84.67153% of the sample of 2MASS groups were classified as Gaussian.

To investigate our claim that classified non-Gaussian groups are indeed dynamically more complex than Gaussian ones, we look at the VDPs of five CNOC2 groups with $n_{members} > 20$. Analysis of the resulting profiles indicates that:

1. The profiles of the two Gaussian groups (110 and 308) show declining velocity dispersion with radius

2. Two non-Gaussian groups (138 and 346) have rising profiles, a possible signature of mergers (Menci & Fusco-Femiano1996)

3. The profile of Group 110 flattens toward larger radii, a trend observed by Girardi et al. (1996) in galaxy clusters. The VDPs of the Gaussian and non-Gaussian groups are distinct, supporting our claim that the classified non-Gaussian groups are dynamically different from the Gaussian systems.

We also find that (with an error $= 10\%, 5\%, 1\%$)

- Velocity dispersion follows a Rayleigh distribution. We did simulations and compared the empirical with the theoretical Rayleigh distribution. We found that the e.d.f. distribution converge to theoretical.
- Non-Gaussian galaxies have higher velocities, masses, larger radii and contain more galaxies than Gaussian clusters from the Gaussian.

- We show that considering only faint galaxies in the outskirts, those in NG groups are older and more metal rich than the ones in G groups

From the sample of the Center of Astrostatistics, we found that our Galaxy (MWG) is normal and the Andromeda (M31) is not normal. We studied their K-Band Magnitude and did Q-Q plots and normality tests, such as Anderson-Darling and Cramer-von Mises. ($a = 0.1$)

From the sample of COMBO-17 of the Center of Astrostatistics, we plotted the magnitudes of the galaxies. The blue galaxies are spirals and irregular galaxies that have experienced recent active star formation, while the red galaxies are mostly elliptical that have only older stars formed early in the Universe's history. Note that many galaxies have properties distributed around the major concentrations. While spiral galaxies are bright, elliptical galaxies are dim. Spiral galaxies are hotbeds of star formation, but elliptical galaxies aren't nearly as prolific because they contain less gas and dust, which means fewer new (and brighter) stars are born. The existing stars inside an elliptical galaxy tend to be older, giving off more red light than younger stars.

Also, the R package "MASS" considers physical information on velocities (km/second) for 82 galaxies reported by Roeder (1990). We plotted the velocities of these galaxies and found that the curve of the velocity distribution is sharp in the middle. Most of the galaxies of Corona Borealis are in early type. So, we conclude that:

- Non-Gaussian groups that include many elliptical galaxies are in early type
- A sharp curve of velocity distribution corresponds to early type galaxies.

In the end, we use sky catalog of nearby groups of galaxies taken from the Lyon-Meudon Extragalactic Database (Garcia et.al. 1993). We found that Elliptical and Lenticular Galaxies have larger velocities than Spiral Galaxies. Spiral galaxies are the majority of the sample. We found that all the galaxy groups are Gaussian except of 285^{th} group. So, we observed. For the particular sample we understood that:

- If a group of galaxies (cluster, supercluster) contains a large number of Spiral galaxies, we can classify it as a Gaussian group (not in general).
- If a group of galaxies (cluster, supercluster) contains a large number of Elliptical galaxies, it is difficult to hypothesize Gaussianity.
- Elliptical and Lenticular galaxies of the sample come from a Uniform and Spiral do not come from a Uniform. (Rayleigh, Wilcoxon uniformity tests with $a = 0.05$)

Notice

We studied group of galaxies not galaxies individually. It is obvious that, there are many Non-Gaussian Spiral Galaxies (M31) and many Gaussian Elliptical Galaxies.

We thank the CNOC 2, N.A.S.A., VizieR, and the Center of Astrostatistics for their data sets.

10 REFERENCES

https://iopscience.iop.org/article/10.1088/0004-637X/702/2/1199/pdf

http://vizier.u-strasbg.fr/viz-bin/VizieR-4

http://vizier.u-strasbg.fr/viz-bin/VizieR-4

https://ui.adsabs.harvard.edu/abs/2014NewA...27...41S/abstract

https://ui.adsabs.harvard.edu/abs/1993A%26AS..100...47G/abstract

http://arxiv.org/abs/1103.0426v1

http://user.astro.columbia.edu/~gbryan/C2002_2010/c2002_classification.pdf

Wikipedia

"Το Σύμπαν που αγάπησα-Εισαγωγή στην Αστροφυσική" Μ. Δανέζη και Ε. Θεοδοσίου, Εκδόσεις Δίαυλος

https://astrostatistics.psu.edu/datasets/index.html

https://iopscience.iop.org/article/10.1088/0004-637X/702/2/1199/pdf

Modern Statistical Methods for Astronomy:

https://www.academia.edu/29436935/Modern_Statistical_Methods_for_Astronomy.pdf

Azzalini, A., & Capitanio, A. 1999,J. R. Stat. Soc. B , 61, 579

Carlberg, R. G., Yee, H. K. C., Morris, S. L., Lin, H., Hall, P. B., Patton, D. R.,Sawicki, M., & Shepherd, C. W. 2001,ApJ,552, 427

D'Agostino, R., & Stephens, M. 1986, Goodness-of-Fit Techniques (New York:Dekker)

https://www.lsw.uni-heidelberg.de/users/mcamenzi/Day_2_Hubble_2009.pdf

R Documentation

https://user2011.r-project.org/TalkSlides/Contributed/18Aug_1400_KaleidIIIa_3-Sanderson.pdf

http://www.atlasoftheuniverse.com/superc/cbo.html

https://stats.stackexchange.com/questions/74057/assessing-the-power-of-a-normality-test-in-r

https://books.google.gr/books?id=EAJwDQAAQBAJ&pg=PA296&lpg=PA296&dq=how+to+simulate+ad+test+in+r&source=bl&ots=VoSVkl73dE&sig=ACfU3U3XVUb_zkAexxwVkAnYsxvm_uhngw&hl=el&sa=X&ved=2ahUKEwje1c6rnIHpAhUQzKQKHRLDBtwQ6AEwCXoECAoQAQ#v=onepage&q=how%20to%20simulate%20ad%20test%20in%20r&f=false