Nonparametric Statistics

S. M. Bendre
Dept of Statistics, NEHU
Shillong

July 22, 2010 IIAP, Kavalur

- Robust procedures
- Tests for Normality
- Nonparametric Density Estimation
- Nonparametric Rank Tests
- Nonparametric Tests for Location
- Nonparametric Regression

- Experiment $\rightarrow$ Measurement, data collection
- Carry out relevant statistical inference on population quantities of interest, such as

  Graphical displays

  Inference on population parameters:

    mean, median etc

    spread of the population, variance (standard deviation), interquartile range (IQR)

  Population distribution:

    Shape of probability density function (pdf), cumulative distribution function (cdf)

NGC 4382 luminosity data

Measure of luminosity (n=59)

- What is the mean luminosity of the population?
- What is the probability distribution of luminosity? (Traditional model: Normal distribution of luminosity)

Sample mean = 26.905
Confidence interval for the mean using normal distribution

$$\bar{X} \pm 2\frac{S}{\sqrt{n}} \quad = \quad 26.905 \pm 00.524$$

Data:

$\underline{26.215}$, 26.506, 26.542, 26.551, 26.553, $\cdots$

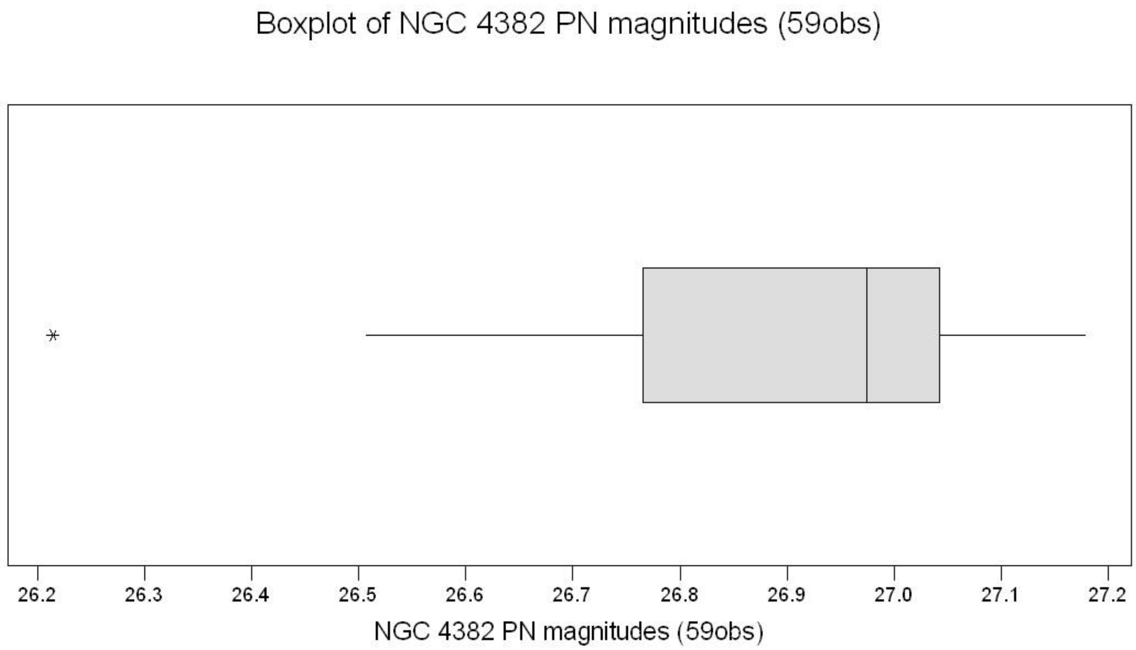


Figure 1: Boxplot of NGC 4382 Luminosity

One outlying observation in the data

Outliers can have arbitrarily large impact on the sample mean, sample standard deviation, and sample variance.

A single outlier can increase the width of the t-confidence interval and inflate the margin of error for the sample mean. Inference can be adversely affected.

It is bad for a small portion of the data to dictate the results of a statistical analysis.

Robust Procedures:

We like to have an estimator and a test statistic that is not overly sensitive to small portions of the data. (Structural robustness)

Robust Estimators

Estimators that are not overly affected by presence of a small proportion of outliers in the data.

For instance, the sample mean is not robust against the presence of even one outlier in the data. The sample median is robust against presence of outliers.

Variable	N	Mean	SE Mean	StDev	Minimum	Median
NGC 4382:no	58	26.917	0.0237	0.181	26.506	26.974
NGC 4382	59	26.905	0.0262	0.201	26.215	26.974

Population probability distribution

- What is the probability distribution of luminosity?

  Traditional model: Normal distribution of luminosity

The construction of confidence interval for population parameters depends on the assumption of population probability distribution.

The standard interval computed in most statistical packages assumes the model distribution is normal.

If this assumption is wrong, the resulting confidence coefficient can vary significantly.

The construction of a 95% confidence interval for the population variance is very sensitive to the shape of the underlying model distribution.

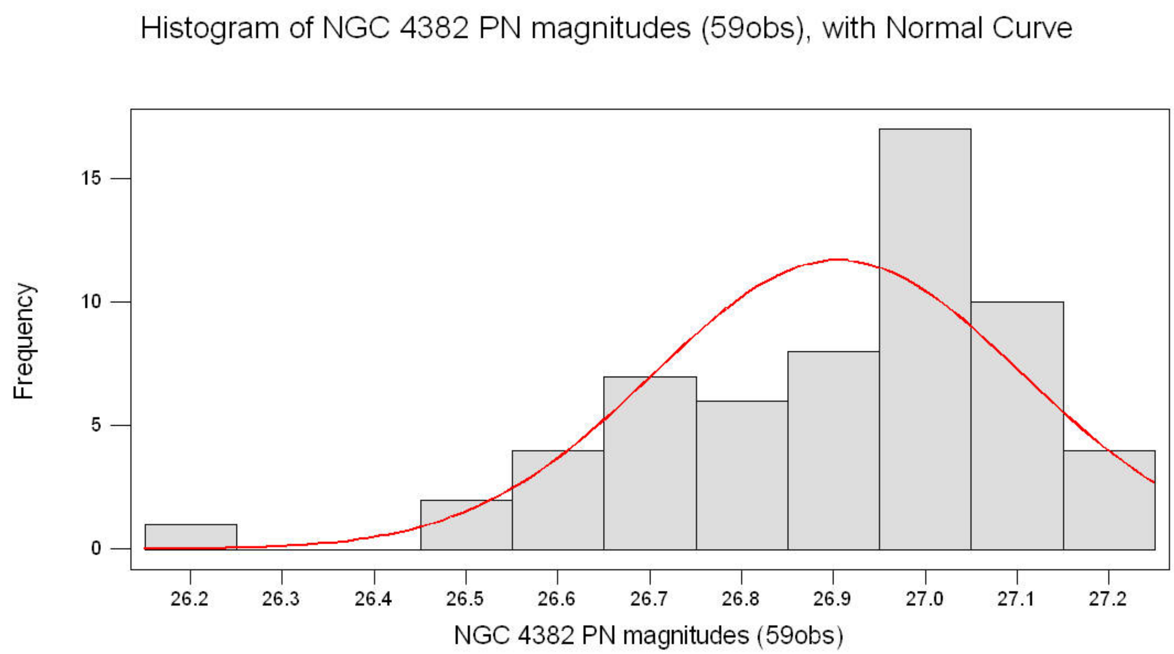


Figure 2: Histogram of NGC 4382 Luminosity

Q: Can the distribution be assumed to be normal?

We need

- either an accurate model specification, or
- a sampling distribution for the estimator and the test statistic that is not sensitive to changes or misspecifications in the model or population distribution. (distributional robustness)

This type of robustness provides stable p-values for testing and stable confidence coefficients for confidence intervals.

Quantile-Quantile (Q-Q) Plot
(Probability-Probability Plot)

Graphical techniques to verify whether the assumption of normality of distribution is valid,

- Q-Q plot: compare the sample ordered observations with the corresponding population quantiles
- P-P plot: compare the empirical cumulative probabilities at observations with the cumulative probabilities of the normal distribution

If the assumption of normality is satisfactory, the points are expected to show an approximate straight line.

Alternatively, one can carry out formal goodness of fit test procedures such as Kolmogorov-Smornov test to verify whether the data fit some proposed model.

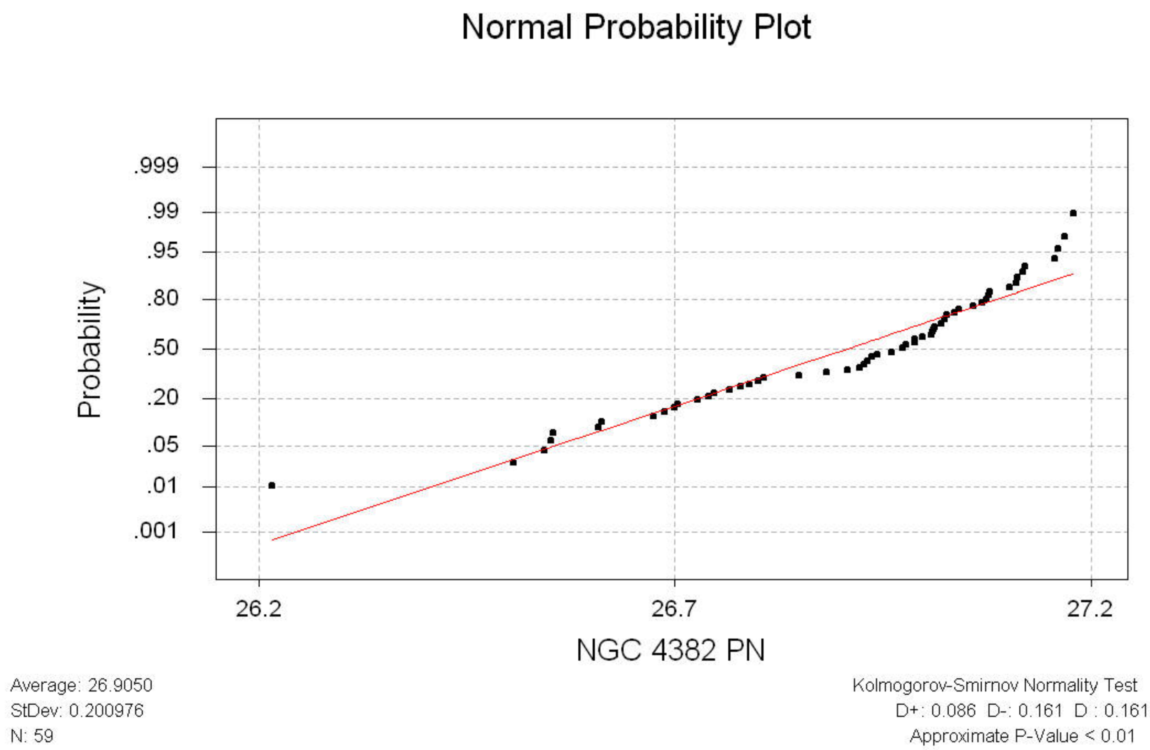


Figure 3: Normal probability plot of NGC 4382 Luminosity

Kolmogorov-Smirnov Test

Procedure to verify whether the sampled population follows some specified distribution.
Suppose we observe $X_1, ..., X_n$ i.i.d. from a continuous distribution function $F(x)$.
To test the hypothesis

$H_0 : F(x) = F_0(x) \forall x$, against $H_1 : F(x) \neq F_0(x)$ for some x,

where F_0 is a distribution which is completely specified before we collect the data.
Let $\widehat{F}_n(x)$ be the empirical cumulative distribution function (CDF) defined by

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

The one sample *Kolmogorov-Smirnov* (K-S) statistic is

$$M = \max_x \left| \widehat{F}_n(x) - F_0(x) \right|$$

- A large value of M supports $F(x) \neq F_0(x)$ and we reject the null hypothesis if M is too large or p-value too small.

- The exact null distribution of M is the same for all F_0, but different for different n. Table of critical values are given for different n in many books.
- The statistic can also be used for constructing confidence band for the distribution which helps in identifying departures from the assumed distribution F_0.
- K-S procedure can be used to reject normality, but not necessarily to accept normality. ‘The inconvenient truth is that it may accept many possible models, some of which can be very disruptive to the t-test and sample means’.

- One situation in which K-S is misused is in testing for normality. For K-S to be applied, the distribution F_0 must be completely specified before we collect the data. In testing for normality, we have to choose the mean and the variance based on the data. This means that we have chosen a normal distribution which is a closer to the data than the true F so that M is too small. We must adjust the critical value to adjust for this as we do in χ^2 goodness of fit tests. Lilliefors has investigated the adjustment of p-values necessary to have a correct test for this situation and shown that the test is more powerful than the χ^2 goodness of fit test for normality.

- Two sample K-S test to verify quality of two population distributions which compares the two empirical distribution functions

$$M = \max_x \left| \widehat{F}_n(x) - \widehat{G}_n(x) \right|$$

Nonparametric Density Estimation

To estimate the density function f based on the random sample $X_1, X_2, \cdots, X_n$ from a probability density function f with unknown functional form.
Histogram : the oldest and widely used *nonparametric density estimator*
Not a satisfactory estimator.

Kernel Density Estimation

The *kernel density estimator* of $f(x)$ at x_o is given by

$$\hat{f}_n(x_o) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_o - X_i}{h}\right)$$

where $K(.)$ is the *kernel function* satisfying the conditions

- $\int_{-\infty}^{\infty} K(x)dx = 1$
- K(.) is symmetric around 0, giving $\int_{-\infty}^{\infty} xK(x)dx = 0$
- $\int_{-\infty}^{\infty} x^2K(x)dx = \sigma^2(K) > 0$

Note that the estimate of f at point x is a weighted function of observations in the h-neighborhood of x with weights depending on the kernel function $K(.)$.
Some kernel functions are

- Uniform kernel: $K(u) = \frac{1}{2}I[|u| \leq 1]$
- Triangle kernel: $K(u) = (1 - |u|)I[|u| \leq 1]$
- Epanechnikov kernel: $K(u) = \frac{3}{4}(1 - u^2)I[|u| \leq 1]$
- Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$

The kernel density estimator satisfies the property

$$\int_{-\infty}^{\infty} \hat{f}_n(x)dx = 1$$

and on the whole gives a better estimate of the underlined density.

Bit of calculations show that

$$E(\hat{f}(x_o)) = f(x_o) + \frac{1}{2}h^2 f''(x_o)\sigma_K^2 + \cdots$$

and

$$Var(\hat{f}(x_o)) = \frac{1}{nh} f(x_o) \int K^2(t)dt$$

So we want the bandwidth $h \to 0$ and $nh \to \infty$ for a satisfactory estimator

Some of the properties of the density estimator are

- Increasing the bandwidth h is equivalent to increasing the amount of smoothing in the estimate. Very large $h(\to \infty)$ will give an oversmooth estimate and $h \to 0$ will lead to a needlepoint estimate giving a noisy representation of the data.

- The choice of the kernel function is not very crucial. The choice of the bandwidth, however, is crucial and the optimal bandwidth can be chosen by minimizing integrated mean square error. The choice of bandwidth is extensively discussed in the literature.

  For instance, with Gaussian kernel, the optimal (MISE) bandwidth is

  $$h_{\text{opt}} = 1.06\sigma n^{-\frac{1}{5}}$$

  where σ is the population standard deviation, which is estimated from the data by $\hat{\sigma} = min\{S, 0.75IQR\}$.

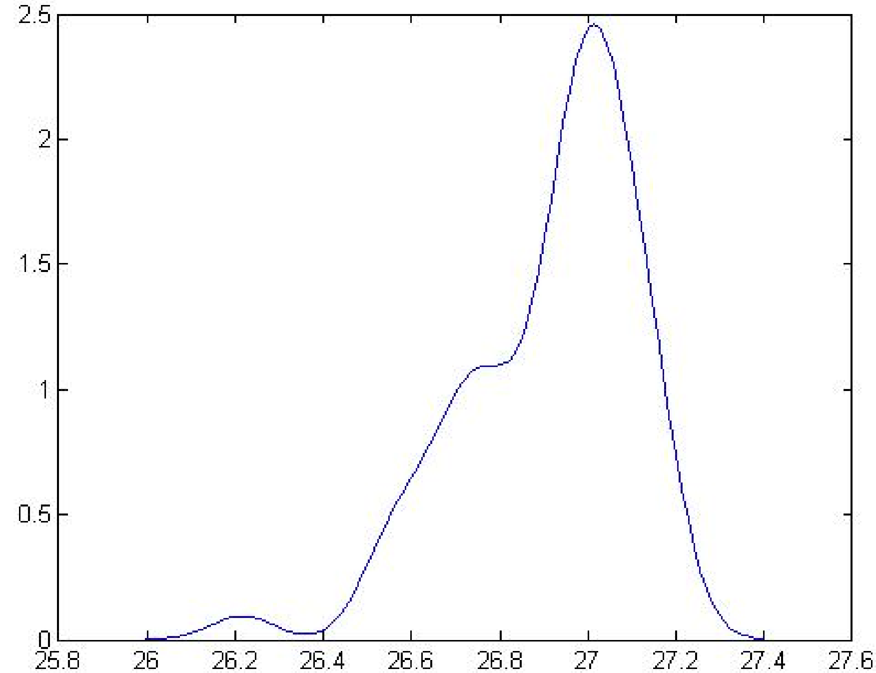


Figure 4: Kernel density estimator of NGC 4382 Luminosity

Nonparametric Tests Procedures

Procedures based n the ranks of observations, which are free from the underlined population distribution (distributional robust)

Single Sample Procedures

We introduce the concept of *location parameter* first.
A population is said to be located at μ_0 if the population median is μ_0.
Suppose $X_1, \cdots, X_n$ is a sample from the population. We say that $X_1, \cdots, X_n$ is located at μ if $X_1 - \mu, \cdots, X_n - \mu$ is located at 0.
Thus any statistic

$$S(\mu) = S(X_1 - \mu, \cdots, X_n - \mu)$$

is useful for the location analysis if $E[S(\mu_0)] = 0$ when the population is located at μ_0. This simple fact leads to some test procedures to test the hypothesis of population locations.

Sign Test

One of the oldest nonparametric procedures where the data are converted to a series of plus and minus signs.
Let $S(\mu)$ be the *sign statistic* defined by

$$\begin{aligned} S(\mu) &= \textstyle\sum_{i=1}^{n} sign(X_i - \mu) \\ &= \#[X_i > \mu] - \#[X_i < \mu] \\ &= S^+(\mu) - S^-(\mu) \\ &= 2S^+(\mu) - n \end{aligned}$$

To find a $\hat{\mu}$ such that $S(\hat{\mu}) = 0$, we get $\hat{\mu} = \text{median}(X_i)$.
Thus if μ_0 is the median of the population, we expect $E[S(\mu_0)] = 0$.
Suppose we wish to test the hypothesis that the population median is μ_0 giving

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu \neq \mu_0.$$

Based on $S(\mu_0)$, the proposed decision rule is:

$$\text{Reject } H_0 \text{ if } |S(\mu_0)| = |2S^+(\mu_0) - n| \geq c$$

where c is chosen such that

$$P_{\mu_0}[|2S^+(\mu_0) - n| \geq c] = \alpha.$$

It is easy to see that under $H_0 : \mu = \mu_0$, the distribution of $S^+(\mu_0)$ is Binomial$\left(n, \dfrac{1}{2}\right)$ irrespective of the underlined distribution of X_i's and hence c can be chosen appropriately. Hence, we reject H_0 if

$$S^+(\mu_0) \leq k \quad \text{or} \quad S^+(\mu_0) \geq n - k$$

where

$$P_{\mu_0}[S^+(\mu_0) \leq k] = \frac{\alpha}{2}.$$

This fact can be used to construct a confidence interval for the population median μ.
Consider

$$P_d[k < S^+(d) < n - k] = 1 - \alpha$$

and find the smallest d such that [the number of $X_i > d$] $< n - k$. Suppose we get

$$\begin{aligned} d = X_{(k)} \ &: \ \#[X_i > X(k)] = n - k \\ d_{min} = X_{(k+1)} \ &: \ \#[X_i > X(k+1)] = n - k - 1. \end{aligned}$$

On the same lines, we find $d_{max} = X_{(n-k)}$. Then a $(1 - \alpha)100\%$ distribution-free confidence interval for μ is given by $[X_{(k+1)},\ X_{(n-k)}]$
Since the median is a robust measure of location, the sign test is also robust and insensitive to the outliers and hence the confidence interval is robust too.

Wilcoxon Signed Rank test

Unlike sign test, utilizes the signs as well as the ranks of the differences between the observed values and the hypothesized median.
Suppose $X_1, \cdots, X_n$ is a random sample from an unknown population with median μ. We assume that the population is symmetric around μ. The hypothesis to be tested is $\mu = \mu_0$ against the alternative that $\mu \neq \mu_0$.
We define $Y_i = X_i - \mu_0$ and rank the absolute values of $|Y_i|$. Let R_i be the rank of the absolute value of Y_i corresponding to the i^{th} observation, $i = 1, \cdots, n$. The signed rank of an observation is the rank of the observation times the sign of the corresponding Y_i.
Let

$$S_i = \begin{cases} 1 & \text{if } (X_i - \mu_0) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

and

$$WS = \sum_{i=1}^{n} S_i R_i.$$

WS is called the *Wilcoxon signed rank statistic.*

A large or a small value of WS indicates a departure from the null hypothesis and we reject the null hypothesis if WS is too large or too small.
The critical values of the Wilcoxon Signed Rank test statistic are tabulated for various sample sizes. The tables of exact distribution of WS based on permutations is given in Higgins(2004).

Normal approximation

It can be shown that for large sample, the null distribution of WS is approximately normal with mean μ and variance σ^2 where

$$\mu = \frac{n\,(n+1)}{4}, \quad \sigma^2 = \frac{n\,(n+1)\,(2n+1)}{24}$$

and the Normal cut-off points can be used for large values of n.

Two Sample Procedures

Luminosity measures on NGC4494 and NGC4382

Q: Do the two differ in luminosity?

Luminosity measurements (data)

NGC 4494 (m = 101)
26.146, 26.167, 26.173, $\cdots$, 26.632, 26.641, 26.643

NGC 4382 (n = 59)
26.215, 26.506, 26.542, $\cdots$, 27.161, 27.169, 27.179

Statistical Model:
Two normal populations with possibly different means but with the same variance.
Translation:

$$H_0 : \mu_{4494} = \mu_{4382} \quad \text{vs} \quad H_1 : \mu_{4494} \neq \mu_{4382}$$

A two sample t-test to compare two means - normality?

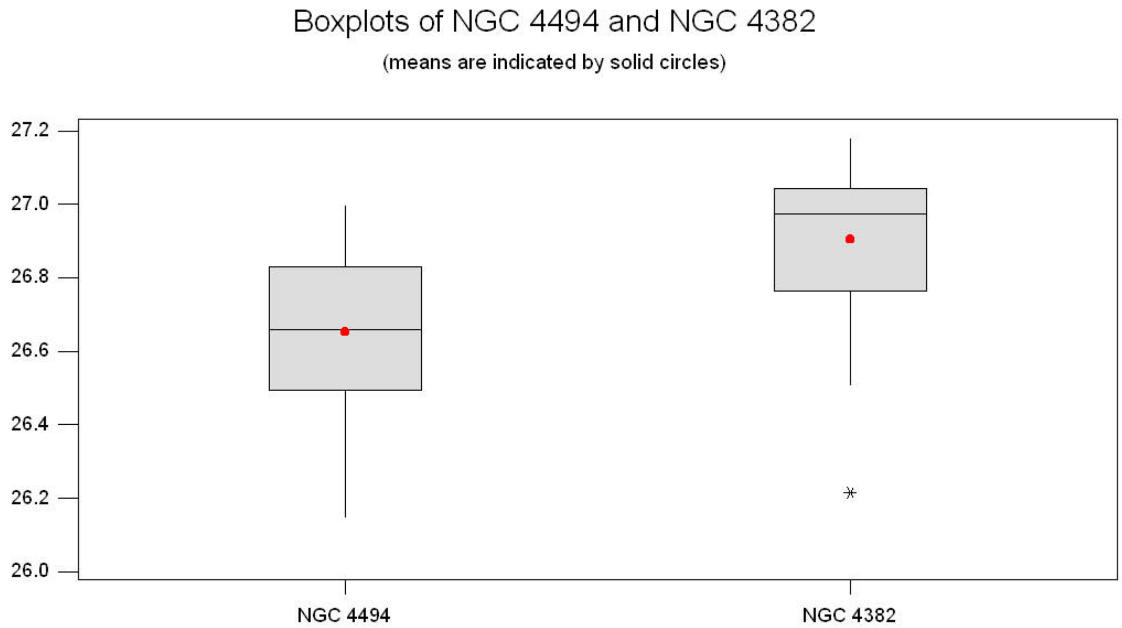


Figure 5: Boxplots of NGC 4494 and NGC 4382 Luminosity

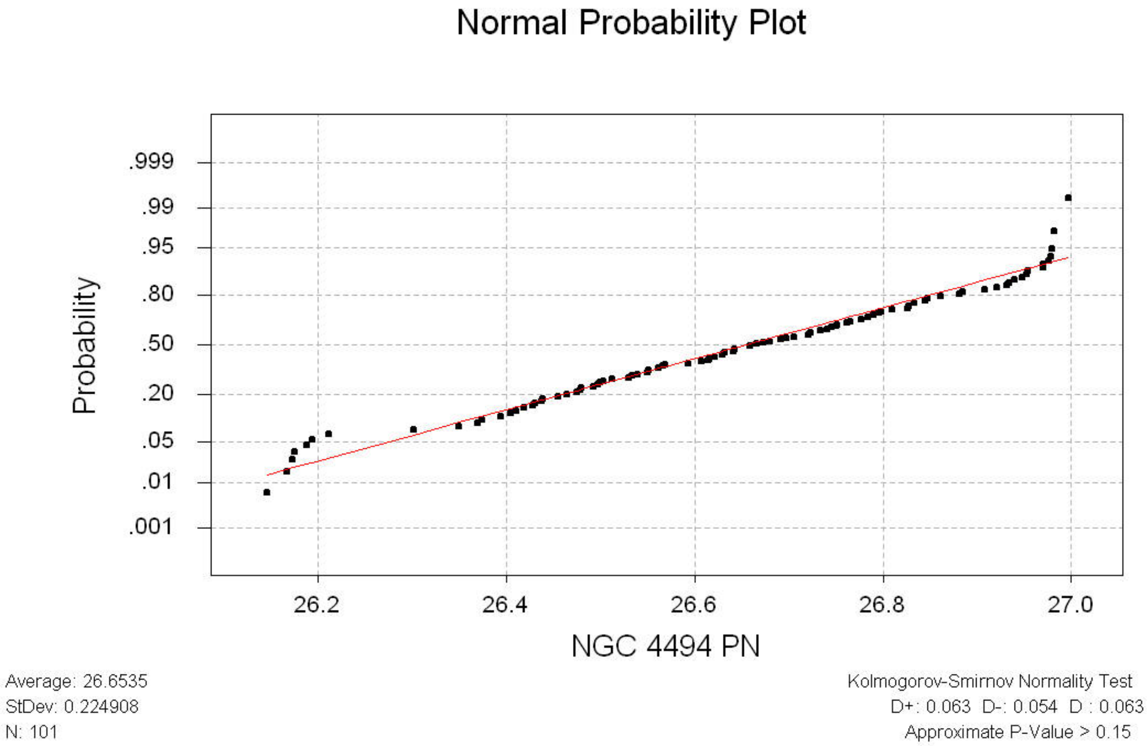


Figure 6: Normal probability plot of NGC 4494 Luminosity

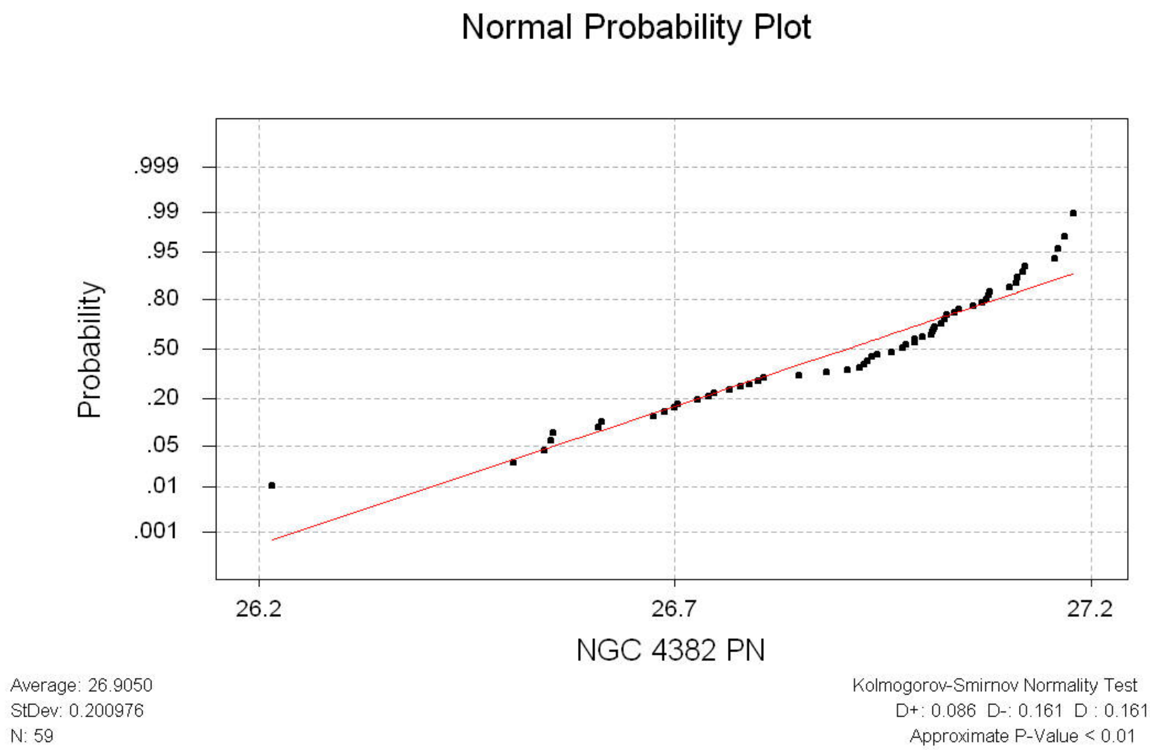


Figure 7: Normal probability plot of NGC 4382 Luminosity

- Outlier in NGC 4382.

  Mean and variance are not robust against outliers

- Assumption of normality not valid for NGC 4382.

  Distribution of two sample t-test is not robust. Problem with p-value of the test.

- The two sample t-test is sensitive to the assumption of equal variances.

Alternative test which is robust - two sample nonparametric tests

Two independent random samples

$X_1, ..., X_n$ from distribution function $F(x)$,
and
$Y_1, ..., Y_n$ from distribution $G(y)$
Both F and G are continuous distributions.

Nonparametric procedures for making inference about the difference between the two location parameters of F and G here.

Assume:

$$G(y) = F(y + \delta)$$

where δ is the difference between the medians.

Hypothesis to be tested:

$$H_0 : \delta = 0 \text{ against } H_1 : \delta \neq 0.$$

Wilcoxon rank sum statistic

Combine and jointly rank all the observations.
Let R_i and S_j be the ranks associated with X_i and Y_j.
Define

$$H = \sum_{i=1}^{n} R_i$$

Note that if $\delta > 0$, then the $X_i's$ should be greater than the $Y_j's$, hence the $R_i's$ should be large and hence H should be large. A similar motivation works when $\delta < 0$.
Thus we support the alternative hypothesis $H_0 : \delta \neq 0$ if H is too large or too small.
This test is called the *Wilcoxon rank-sum test*.
Tables of exact distribution of H are available in Higgins (2004).

Mann-Whitney test

Let

$$V_{ij} = X_i - Y_j,$$

We define

$$U = \#\left(V_{ij} > 0\right)$$

which is the *Mann-Whitney* statistic. The Mann-Whitney test rejects the null hypothesis $H_0 : \delta = 0$ if U is too large or too small.

It can be shown that there is a relationship between the Wilcoxon rank sum H and the Mann-Whitney U :

$$H = U + \frac{n\left(n+1\right)}{2}.$$

Hence the critical values and p-values for U can be determined from those for H.

Luminosity of NGC 4494 and NGC 4382

Mann-Whitney Test and CI:

```
 NGC 4494 PN magn, NGC 4382 PN magn


NGC 4494   N = 101     Median =      26.659
NGC 4382   N =  59     Median =      26.974

Point estimate for ETA1-ETA2 is      -0.253

95.0% CI for ETA1-ETA2 is (-0.328,-0.182)
W = 6284.5

Test is significant at 0.0000.
```

The data supports the alternative hypothesis that the two medians are not equal.

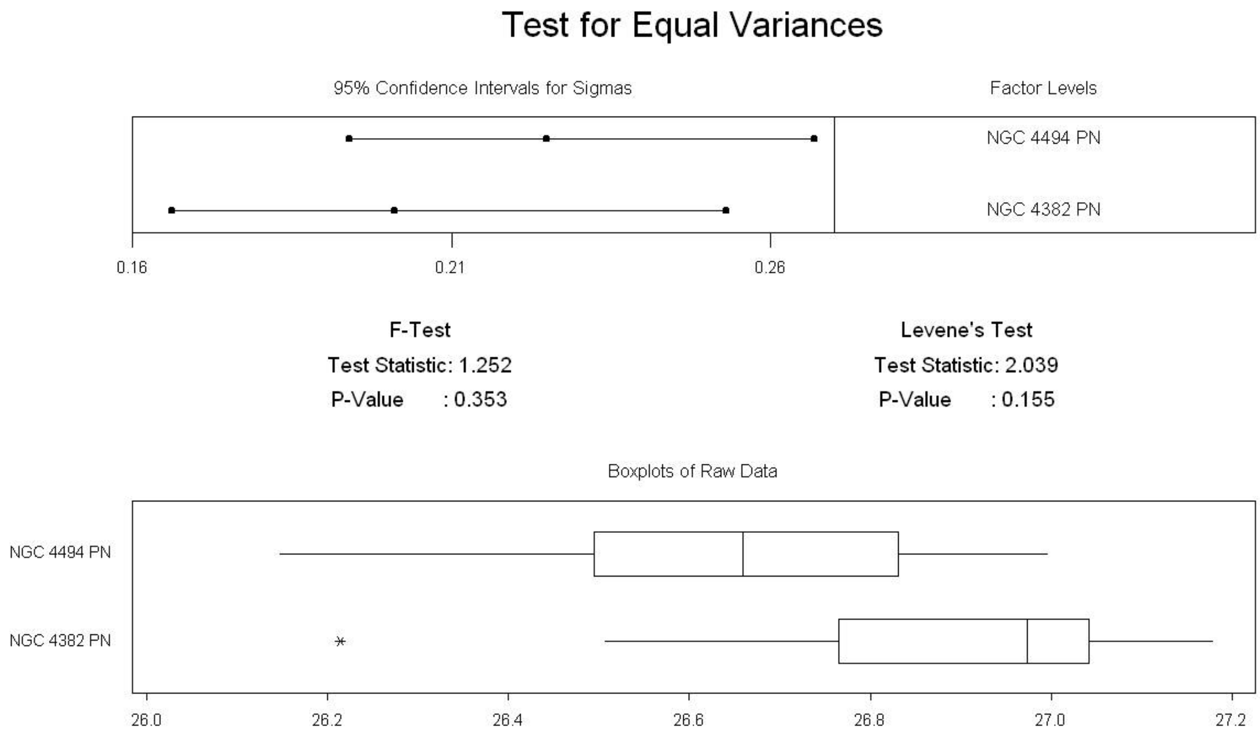
Two-Sample T-Test and CI:

```
NGC 4494 PN magn (101obs), NGC 4382 PN magn  test(59obs)


Two-sample T

             N      Mean     StDev   SE Mean
NGC 4494  101    26.654     0.225     0.022
NGC 4382   59    26.905     0.201     0.026

Difference = mu NGC 4494 (101obs) - mu NGC 4382  (59obs)
Estimate for difference:  -0.2515
95% CI for difference: (-0.3215, -0.1814)
T-Test of difference: T-Value = -7.09  P-Value = 0.000  DF = 158
Both use Pooled StDev = 0.216
```



Test for Equal Variances
95% Confidence Intervals for Sigmas
Factor Levels
NGC 4494 PN
NGC 4382 PN
0.16
0.21
0.26
F-Test
Test Statistic: 1.252
P-Value : 0.353
Levene's Test
Test Statistic: 2.039
P-Value : 0.155
Boxplots of Raw Data
NGC 4494 PN
NGC 4382 PN
26.0
26.2
26.4
26.6
26.8
27.0
27.2

The Mann-Whitney test is less sensitive to the assumption of equal scale parameters.

The null distribution is nonparametric. It does not depend on the common underlying model distribution.

It depends on the permutation principle: Under the null hypothesis, all (m+n)! permutations of the data are equally likely. This can be used to estimate the p-value of the test: sample the permutations, compute and store the MW statistics, then find the proportion greater than the observed MW.

Paired data

Analogous to the paired t-test in parametric inference, we can propose a nonparametric test of hypothesis that the median of the population of differences between pairs of observations is zero.

Suppose we observe a sequence of i.i.d. paired observations

$(X_1, Y_1), ..., (X_n, Y_n)$. Let μ_D be the median of the population of differences between the pairs. The goal is to draw inference about μ_D. Let

$$D_i = X_i - Y_i$$

The distribution of D_i is symmetric about μ_D. Therefore, we may used the procedures discussed earlier for the one-sample model, based on the observations D_i.

k-Sample Procedure

Mann-Witney-Wilcoxon procedure generalized to compare k samples.

Nonparametric analogue of the parametric one-way analysis of variance procedure.

k independent random samples of sizes $n_i, i = 1, \cdots, k$

$X_{ij}, j = 1, \cdots, n_i;\ i = 1, \cdots, k.$
Let the underlined location parameters be denoted by $\mu_i, i = 1, \cdots, k.$

$H_0 : \mu_i$ are all equal against $H_1 : \mu_i \neq \mu_{i*}$ for some $i \neq i*$

Procedure: combine the k samples and rank them.
Let
R_{ij} = the rank associated with X_{ij}
and
$\overline{R}_{i.}$ = the average of the ranks in the i^{th} sample.

If the null hypothesis is true, the distribution of ranks over different samples will be random and no sample will get a concentration of large or small ranks.
Thus under the null hypothesis, the average of ranks in each sample will be close to the average of ranks for under the null hypothesis.
The Kruskal-Wallis test statistic is given by

$$KW = \frac{12}{N(N+1)} \sum n_i \left(\overline{R}_{i.} - \frac{N+1}{2} \right)^2$$

If the null hypothesis is not true, the test statistic KW is expected to be large and hence we reject the null hypothesis of equal locations for large values of KW.
The tables of exact critical values are available i the literature. We generally use a χ^2 distribution with $k-1$ degrees of freedom as an approximate sampling distribution for the statistic.

A Strategy:

- Use robust statistical methods whenever possible.
- If you must use traditional methods (sample means, t and F tests) then carry out a parallel analysis using robust methods and compare the results. Start to worry if they differ substantially.
- Always explore your data with graphical displays. Attach probability error statements whenever possible

Nonparametric Regression

Suppose we have n observations $(Y_1, X_1), \cdots, (Y_n, X_n)$ on (Y, X) where Y is the response variable and X is the predictor variable and the aim is to model Y as a function of X.

Linear Regression

- Most widely used statistical procedure
- $E[Y|X = x]$ is assumed to be a linear function of X, specified by

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \cdots, n,$$

  and the errors ϵ_i are taken to be uncorrelated with zero mean and variance σ^2.
- When not appropriate, fitting a linear regression model to a nonlinear relationship results in a totally misleading and unreliable inference.

A more general alternative is **Nonparametric regression** when the functional form of $E[Y|X = x]$ can not be assumed.

In particular, the model considered is

$$Y_i = m(X_i) + \epsilon_i$$

where the regression curve $m(x)$ is the conditional expectation $m(x) = E[Y|X = x]$ with $E[\epsilon|X = x] = 0$ and $\mathrm{Var}[\epsilon|X = x] = \sigma^2(x)$.

The model removes the parametric restrictions on $m(x)$ and allows the data to dictate the alternative structure of $m(x)$ by using the data based estimate of $m(x)$.

The available estimation procedures estimate the regression curve using the information available in the neighborhood and are called *smoothing techniques*.

Different smoothing techniques lead to different nonparametric regression estimators.

Kernel Estimator

We have

$$\begin{aligned} m(x) &= E[Y|X=x] \\ &= \int y \frac{f(x,y)}{f(x)} dy \end{aligned}$$

where $f(x)$ and $f(x,y)$ are the marginal density of X and the joint density of X and Y respectively. On substituting the univariate and bivariate kernel density estimates of the two densities and noting the properties of kernel function $K(.)$ specified in Section 3, we get

$$\begin{aligned} \hat{m}_{NW}(x) &= \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} \\ &\equiv \sum_{i=1}^{n} W_{hi}(x) Y_i \end{aligned}$$

which is a weighted average of the response variables in a fixed neighborhood around x.

The weights are given by

$$W_{hi}(x) = (nh)^{-1} \frac{K\left(\frac{x - X_i}{h}\right)}{\hat{f}(x)}.$$

$\hat{m}_{NW}(x)$ is called the *Nadaraya-Watson kernel estimator*. Note that

- The weights depend on the kernel function $K(.)$, the bandwidth h and the whole sample $\{X_i, i = 1, \cdots, n\}$ through the kernel density estimate $\hat{f}(x)$.
- For the uniform kernel, the estimate of $m(x) = E[Y|X = x]$ is the average of Y_j's corresponding to the X_j's in the h-neighborhood of x.
- Observations Y_i obtain more weight in those areas where the corresponding X_i are sparse.
- When the denominator is zero, the numerator is also equal to zero and the estimate is set to be zero.

- Analogous to kernel density estimation, the bandwidth h determines the level of smoothness of the estimate and is called the smoothing parameter. Decreasing bandwidth leads to a less smooth estimate. In particular, for $h \to 0$ the estimate $\hat{m}(X_i)$ converges to Y_i and for $h \to \infty$ the estimate converges to $\bar{Y}$. The criteria of bandwidth selection and guidelines for selecting the optimal bandwidth are available in the literature.

In case the predictors $X_i,\ i = 1, \cdots, n$ are not random, alternative estimators such as *Gasser-Müller kernel estimator* are more appropriate.

It can be shown that the Nadaraya-Watson kernel estimator is the solution of the weighted least squares estimator obtained on minimizing

$$\sum_{i=1}^{n}(Y_i - \beta_0)^2 K\left(\frac{x - X_i}{h}\right)$$

over β_0. This corresponds to locally approximating $m(x)$ with a constant while giving higher weights to the Y_j's corresponding to the X_j's in the h-neighborhood of x.
This concept is further generalized to fitting higher order polynomials 'locally', i.e. in the neighborhood of x. In particular, we consider minimizing

$$\sum_{i=1}^{n}[Y_i - \beta_0 - \beta_1(x - X_i) - \beta_2(x - X_i)^2 - \cdots - \beta_p(x - X_i)^p]^2 K\left(\frac{x - X_i}{h}\right)$$

over $\beta_0, \beta_1, \cdots, \beta_p$. The resulting estimator is called the *local polynomial regression estimator* and the appropriate choice of the degree of polynomial p can be made based on the data.

k-Nearest Neighbor Estimator

A weighted average of response variables in the neighborhood of x.

Unlike the kernel estimator with a fixed h-neighborhood, considers a varying neighborhood around x defined by the k X_j's which are closest to x.

In particular, for every x, define the set of indexes

$J_x = \{i \ : \ X_i$ is one of the k nearest observations to $x\}$

and construct the weight sequence $\{W_{ki}(x), \ i = 1, \cdots, n\}$ given by

$$W_{ki}(x) = \begin{cases} \dfrac{n}{k} & \text{if } \ i \in J_x \\ 0 & \text{otherwise} \end{cases}$$

Then the $k - NN$ estimator of $m(x)$ is defined as

$$\hat{m}_k(x) = n^{-1} \sum_{i=1}^{n} W_{ki}(x) Y_i.$$

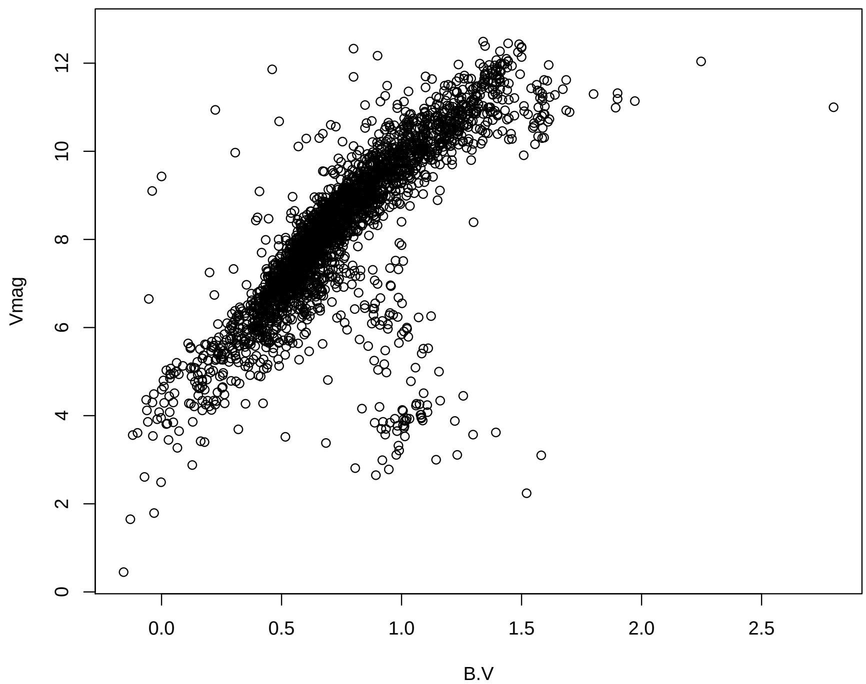
k is the smoothing parameter here as it controls the degree of smoothness of the estimated curve. For $k = n$ the neighborhood covers the entire sample for each x, giving $\hat{m}_{ni}(x) = \bar{Y}$. On the other hand, $k = 1$ gives a step function which is equal to Y_i for $x = X_i$ and jumps in the middle between two adjacent values of X. Variations of the $k - NN$ estimator using different weight sequences are also proposed in the literature.

LOWESS Estimator

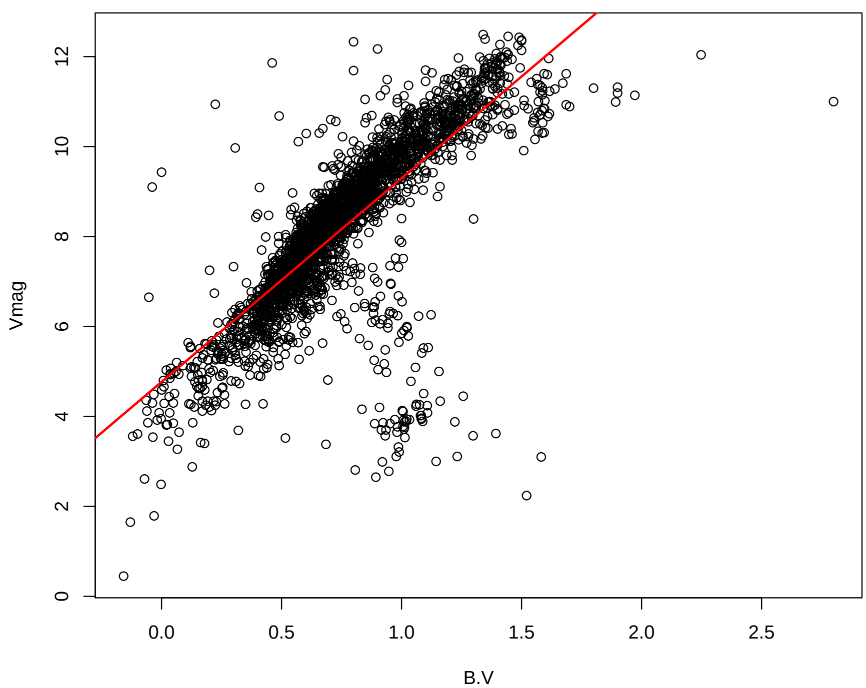
LOWESS stands for a LOcally WEighted Scatter plot Smoother which combines the two smoothing techniques discussed above and is more flexible and robust.

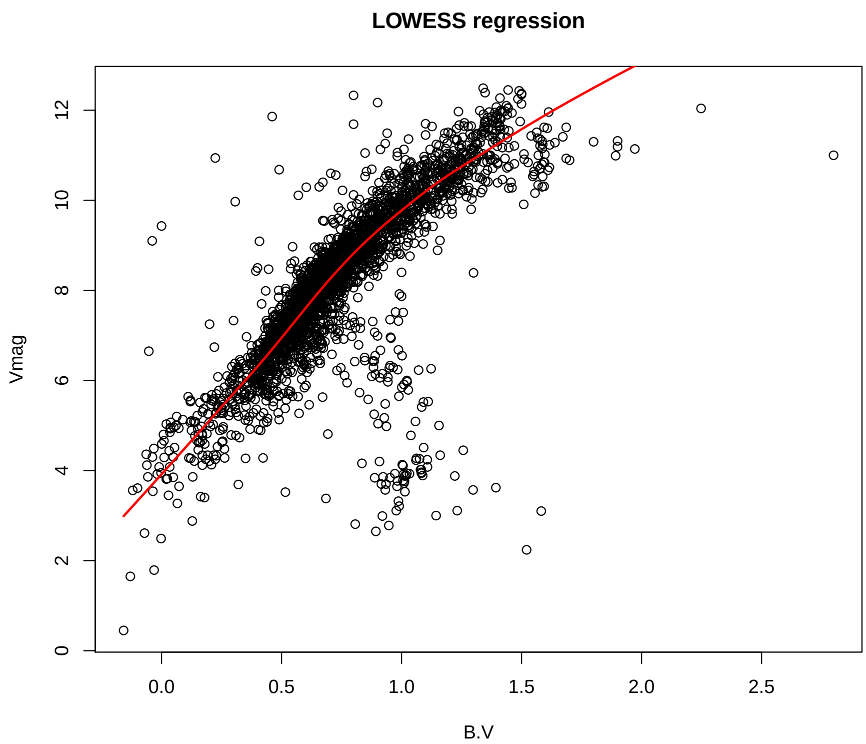
- initially selects varying bandwidth based on the nearest neighbors
- iteratively uses the polynomial weighted least squares fit in each neighborhood.

The polynomials considered are either linear or quadratic and the weights given to the response variables corresponding to X_i's in the neighborhood of x are determined by the choice of the kernel function.
LOWESS can not be expressed in a closed form and estimating it is a computer-intensive technique. A more general technique called LOESS is also available in the literature.

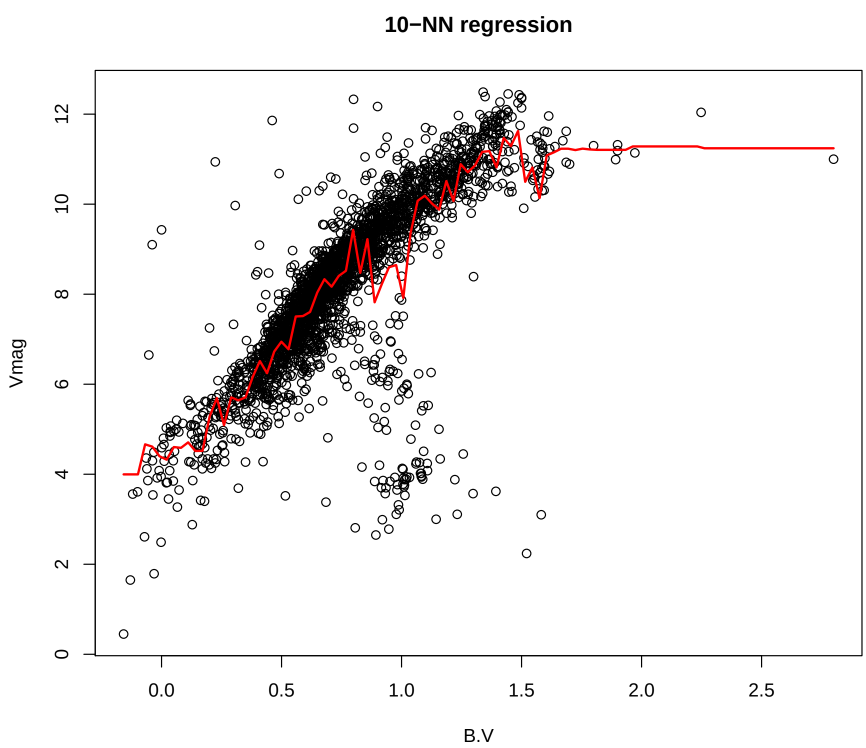


Linear regression

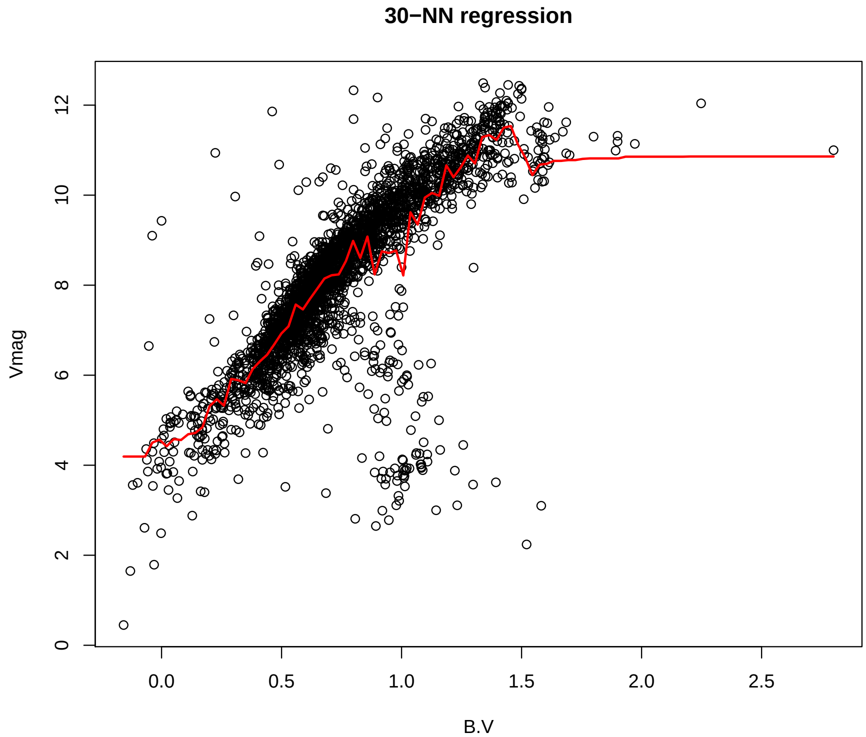




LOWESS regression
Vmag
0
2
4
6
8
10
12
0.0
0.5
1.0
1.5
2.0
2.5
B.V



10−NN regression
Vmag
0
2
4
6
8
10
12
0.0
0.5
1.0
1.5
2.0
2.5
B.V



30−NN regression
Vmag
B.V
0
2
4
6
8
10
12
0.0
0.5
1.0
1.5
2.0
2.5

References

1. Arnold, Steven (1990), *Mathematical Statistics* (Chapter 17). Prentice Hall, Englewood Cliffs, N. J

2. Beers, Flynn, and Gebhardt (1990), Measures of Location and Scale for Velocities in Cluster of Galaxies-A Robust Approach. *Astron Jr*, 100, 32-46.

3. Härdle, W. (1990) *Applied Nonparametric Regression.* Cambridge Univ Press, Cambridge.

4. Hettmansperger, T and McKean, J (1998), *Robust nonparametric Statistical Methods.* Arnold, London.

5. Higgins, James (2004), *Introduction to Modern Nonparametric Statistics.* Duxbury Press.

6. Hollander and Wolfe, (1999), *Nonparametric Statistical Methods* John Wiley, N.Y.

7. Johnson, Morrell, and Schick (1992), Two-Sample Nonparametric Estimation and Confidence Intervals Under Truncation, *Biometrics*, 48, 1043-1056.

8. Summer School in Statistics for Astronomers(2005). Lecture Notes by Steven Arnold Website: http://astrostatistics.psu.edu/

9. Summer School in Statistics for Astronomers(2008). Lecture Notes by Tom Hettmansperger. Nonparametrics.zip. Website: http://astrostatistics.psu.edu/

10. Summer School in Statistics for Astronomers(2010). Lecture Notes by Tom Hettmansperger. Nonparametrics.zip. Website: http://astrostatistics.psu.edu/

11. Website: http://www.stat.wmich.edu/slab/RGLM/

***** Part of these notes borrow heavily from the material in 8 and 9