Thurs 1-31-2019
Probability Distribution (continued)

1. Robust statistics and the boxplot
2. Assessment of normality
3. Data transformation
4. Trial runs of geosa2….lightning talk

Assignment a2: due Tues, Feb 5

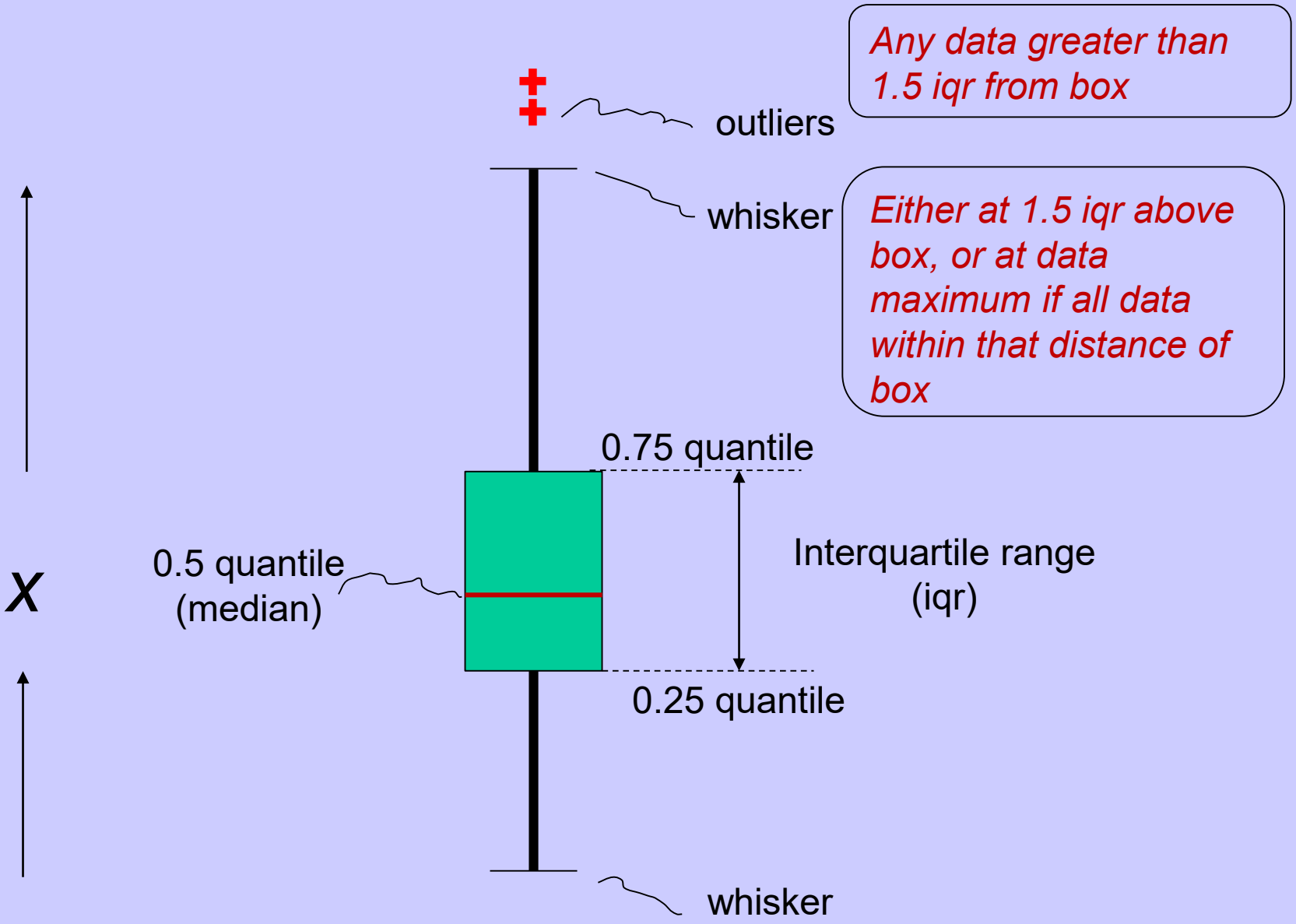
-read notes_2.pdf

-laptops to class each Tuesday

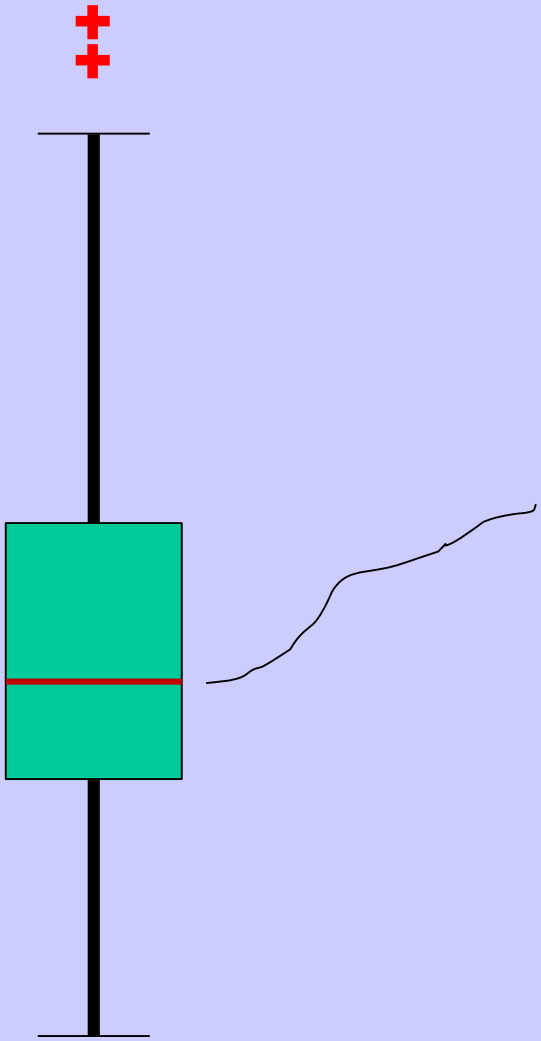
Robust statistics

- The basic descriptive statistics – mean , variance, skew – can be sensitive to changes in just a few of the observations
- Robust statistics, in contrast, are little affected by changes in a few observations
- Robust statistics for location, spread and symmetry can be defined from the ranked data values as represented in the ecdf
- A useful graphic for describing these robust statistics is the boxplot

Boxplot


Any data greater than 1.5 iqr from box
outliers
whisker
Either at 1.5 iqr above box, or at data maximum if all data within that distance of box
0.75 quantile
Interquartile range (iqr)
X
0.5 quantile (median)
0.25 quantile
whisker

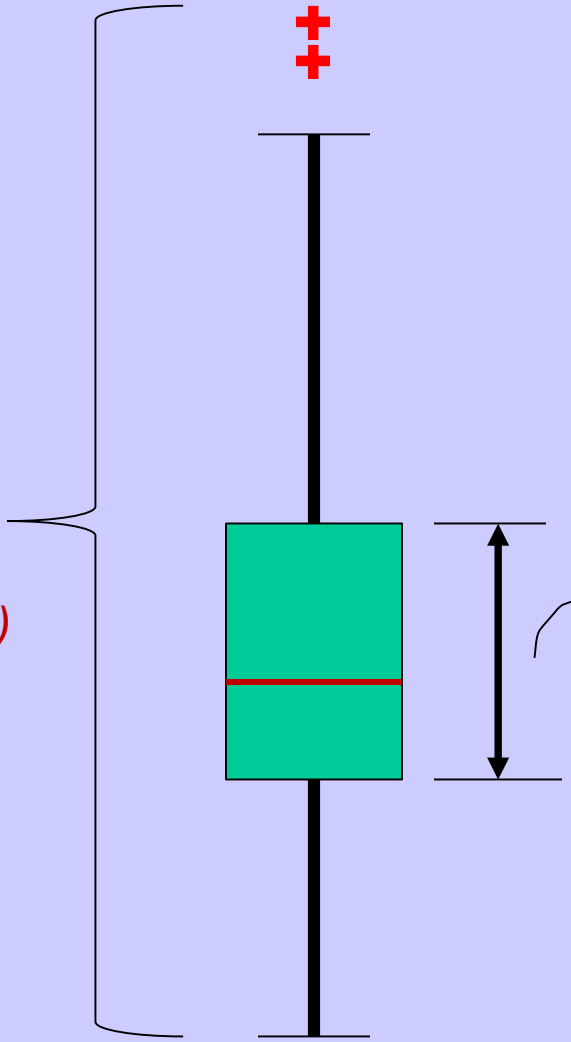
Boxplot -- location



Median

- *Robust statistic of location*
- *Insensitive to changes in other data as long as half of the observations remain higher and half lower*

Boxplot -- spread



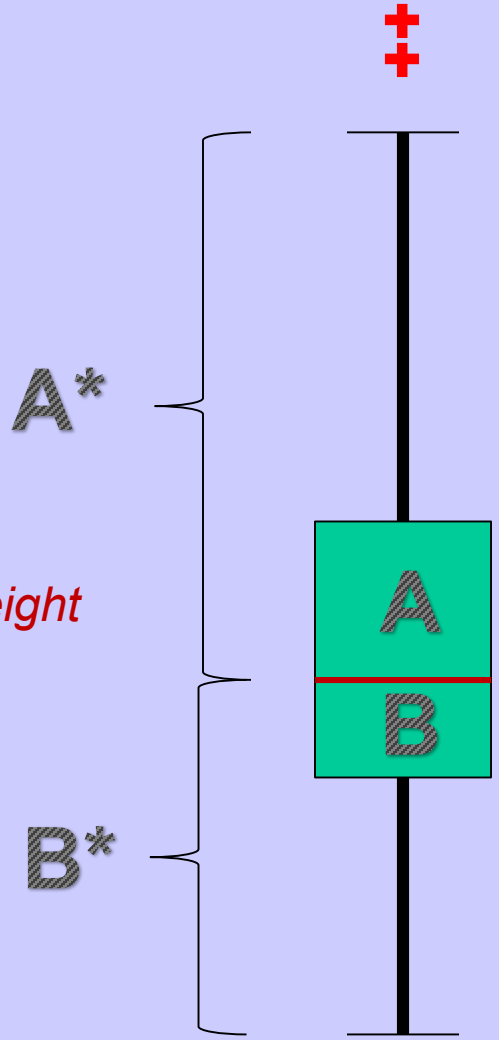
quantile

$\text{iqr} = q_{0.75} - q_{0.25}$

Range
(extremely non-robust)
$x_{\max} - x_{\min}$

Interquartile range
A robust measure of spread ("middle half" of the data)

Boxplot -- symmetry



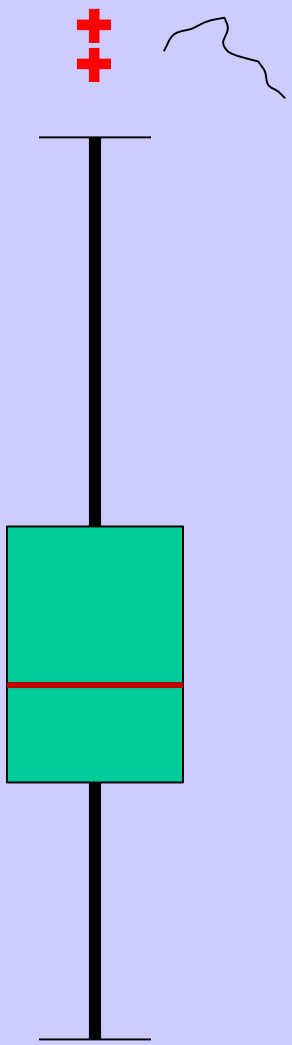
Skew in tails: Difference in height of A and B**

Yule-Kendall Index
(Wilks, 1995, p. 27)

- *Difference in height of A and height of B*
- *Asymmetry in "middle" of distribution*
- *>0 is positive skew; <0 is negative skew*

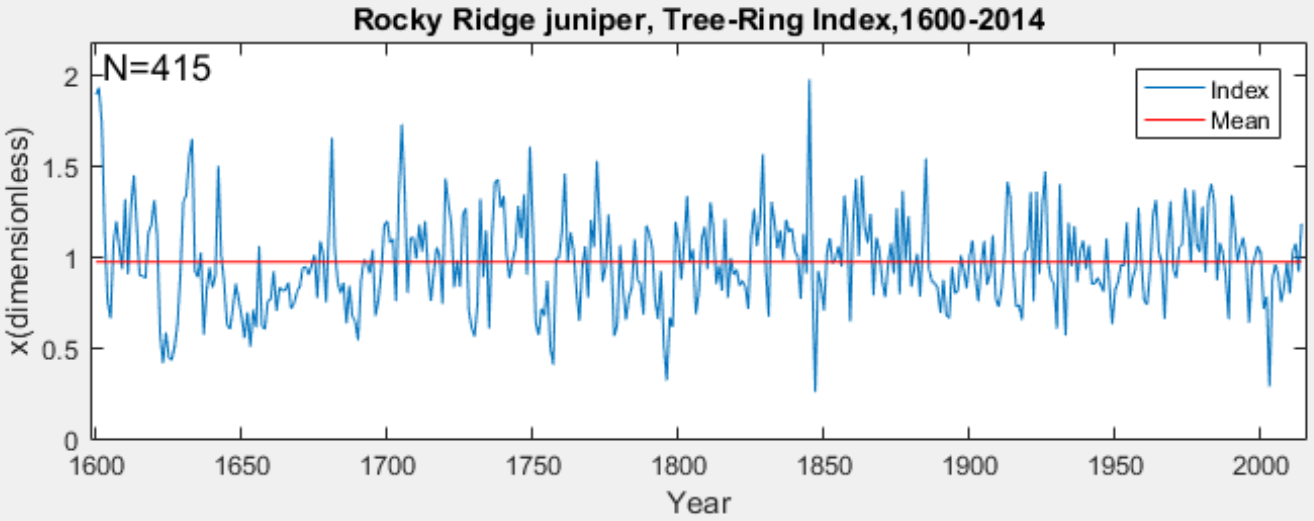
$$\gamma_{\mathrm{YK}} = \frac{\left(q_{0.75} - q_{0.50}\right) - \left(q_{0.50} - q_{0.25}\right)}{\mathrm{iqr}}$$

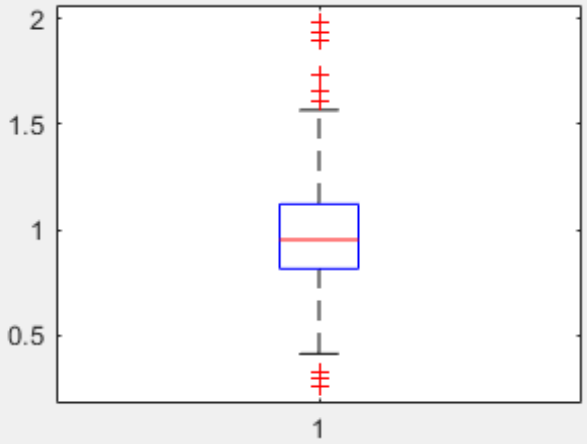
Boxplot -- outliers



- For a normal distribution, 0.7% of the population is outside whiskers (0.35% on each end)
- So, for a large sample, “outliers” are not necessarily a sign of non-normality
- For example,with N=1000, and normality, would expect 3.5 observations outside top and bottom whiskers.
- Moreover, the boxplot is not intended to test for normality; symmetry does not imply normality

Boxplot – example
A tree-ring chronology





mean = 1.0
median=0.953
} Location

s=0.258
iqr=0.306
} Spread

g=0.525
γ_{YK}=0.107
} Symmetry

Testing for Normality

Generally look for a good match in pdf's or cdf's (empirical with theoretical)

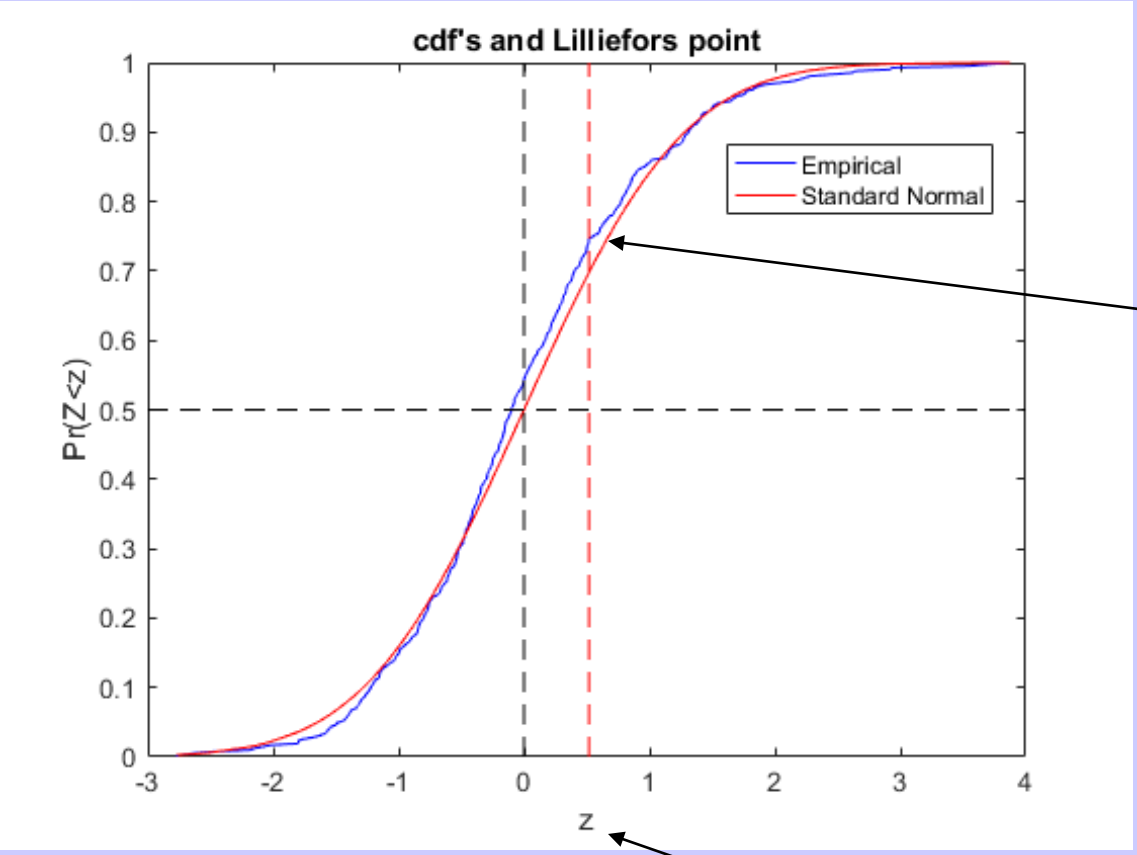
1. Graphical: Matlab functions histfit , normplot
2. Statistical
    - Various statistical tests have been proposed, including
        - Chi square test of histogram against normal histogram
        - Kolmogorov-Smirnov (KS) test: tests based on departure of cdf's, and requires full specification of the null model
        - *Lilliefors test: similar to KS test, but does not require specification of parameters of the null model

Used in course

Lilliefors Test -- approach

1. Convert time series to “z scores” by subtracting mean and dividing by standard deviation
2. Plot cdf of the z-score time series along with cdf of the standard normal distribution evaluated at same points
3. Test statistic is the largest vertical departure in the two cdf’s
4. Test statistic is evaluated by reference to a lookup table (in Matlab) that is based on Monte Carlo sampling*

Lilliefors Test – cdf plots for sample data (Rocky Ridge Juniper tree-ring index)



Maximum departure

Standardized variable

Hypothesis testing: general steps and as applied in Lilliefors Test

Haan 2002

1. Formulate hypothesis to be tested: H0: the data are drawn from a normal distribution
2. Formulate and alternative hypothesis: H1: the data are drawn from a non-normal distribution.
3. Determine the test statistic: T: the maximum absolute difference between the ecdf of the sample and the theoretical cdf of a standard normal distribution.
4. Determine the distribution of the test statistic: This test statistic, T, has no theoretical distribution. A distribution is estimated by repeated drawing of samples (e.g., thousands of times) the same length as the time series of interest from a a standard normal distribution and building a table of how frequently various levels of T are reached by chance alone.
5. Determine the rejection region or critical region of the test statistic: Set the rejection region at the level of T achieved in less than 5% of the Monte Carlo samples.

Hypothesis testing: general steps and as applied in Lilliefors Test --- cont.

6. Collect the data needed to calculate the test statistic, and calculate the test statistic for that sample:This is the sampe time series; the statistic T is computed as indicated in previous slide
7. Determine if the computed statistic falls in the rejection region: Compare the computed value of T to the values in Matlab's lookup table. Does the computed T exceed the value in the table achieved by chance alone by fewer than 5% of the Monte Carlo simulations?

Hypothesis testing: general steps and as applied in Lilliefors Test --- cont.

In rejection area (T large):

- Reject H0 of normality
- Rejected at α=0.05, which means there is a 5% chance of of Type I error (rejected H0 when H0 was really true)
- Significance level of the rejection is 100(1- α), or 95%

Not in rejection area (T small)

- Accept H0 of normality
- Conclude T not significant at α=0.05, or significance level < 95%
- There is a chance of Type II error: accepting H0 of normality while the series is actually from a non-normal distribution
- The probability of a Type II error depends on the distribution the sample was actually drawn from, and will vary depending on the distribution assumed

Lilliefors test result for the tree-ring series of earlier example

>>[h, p]=lillietest(z,'Alpha',0.05)

h→1 ….reject H0 at α=0.05

p→0.0175 … would not reject at α=0.01

What if series fails normality test and want to apply in analysis requiring normality?

Data Transformation

1. Logarithmic transformation
2. Power transformation
   (e.g., square root)

Widely use in hydrology and dendrochronology

Logarithmic transformation

x_t : original series

$$y_t = \log_{10}(x_t) \quad \text{or} \quad y_t = \log_{10}(x_t - c)$$

Lower-bound Parameter (often set to 0)

x	y
100,000	→5
10,000	→4
1,000	→3
100	→2

De-emphasizes importance of large values to the variability of the data

Beware: 1) log10 of zero is minus infinity, and
2) log10 of a negative number is undefined

Power transformation

x_t : original series

$$y_t = \left(x_t - c\right)^b$$

Power

Lower-bound Parameter (often set to 0)

- Typically, *b*=1/2 or 1/3
- Like the log10 transform, power transformation, with those settings for b, de-emphasize high values, but the effect is less extreme than with the log10

Transformation -- example

demo02a.m in matlab