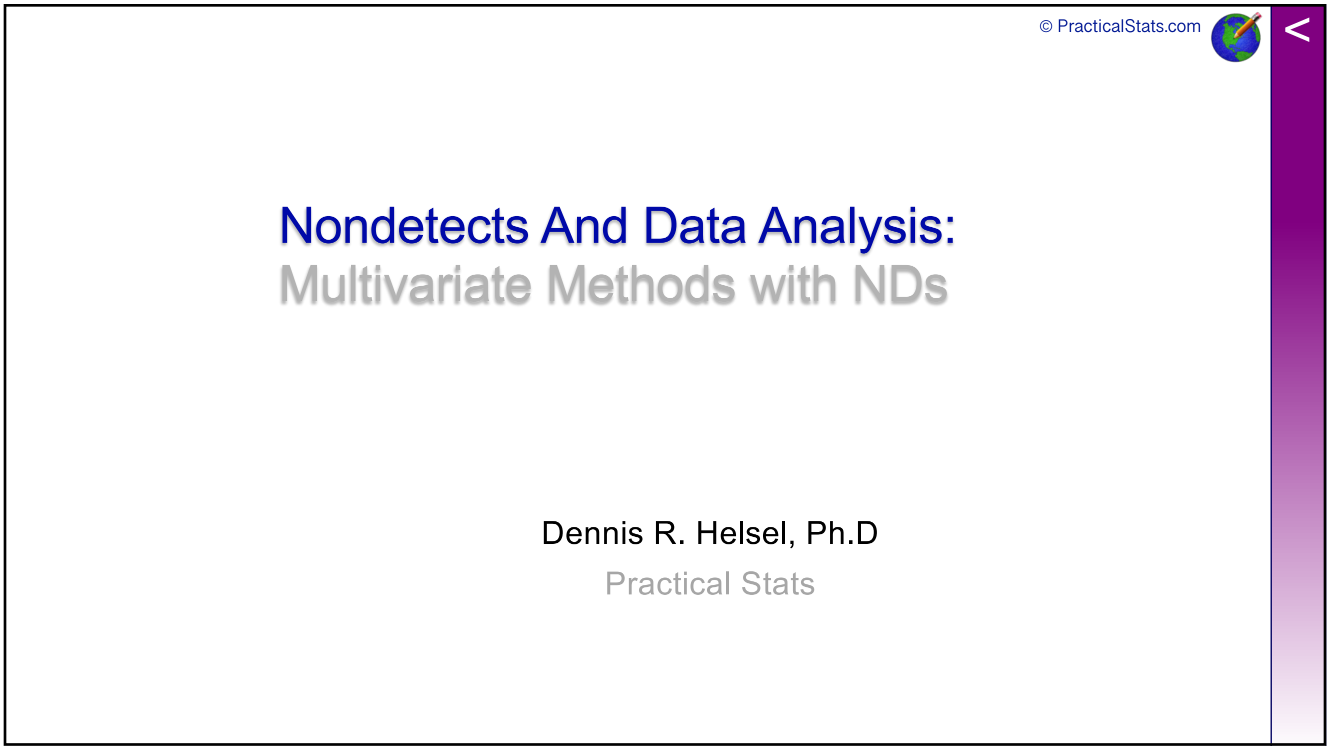
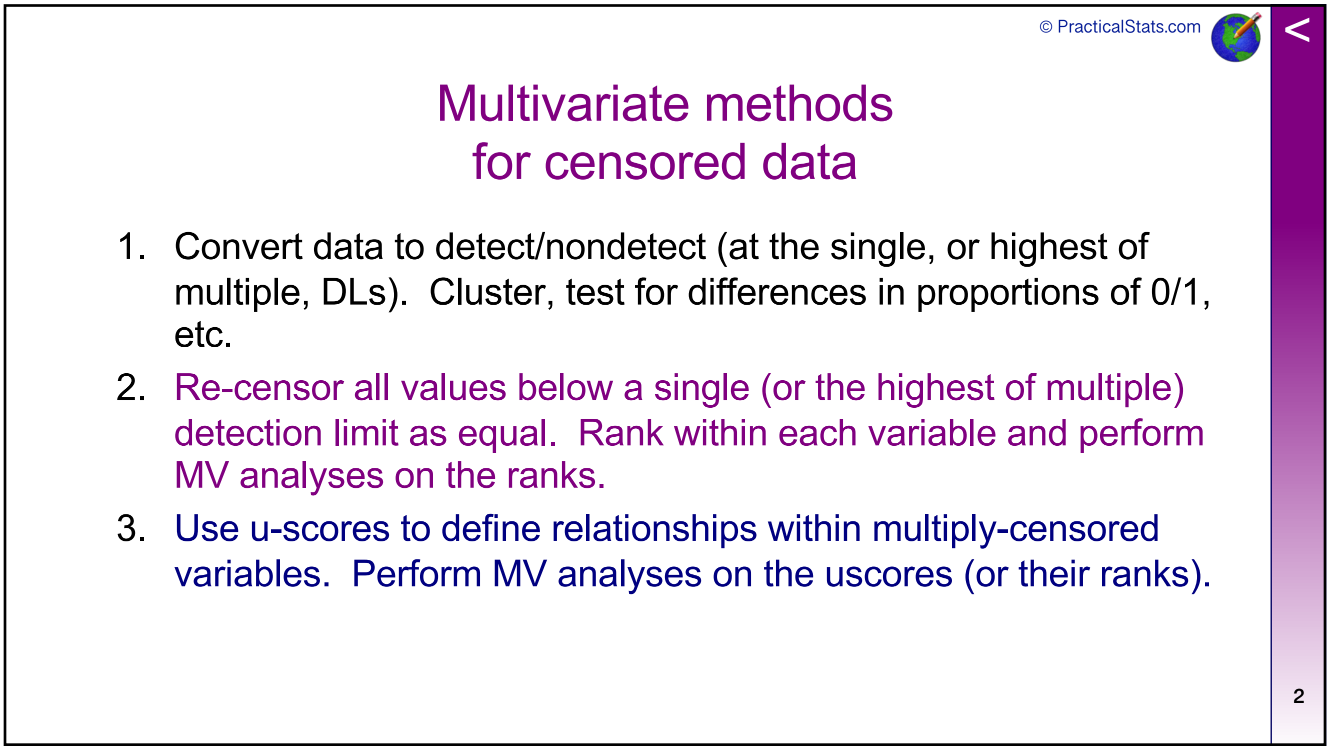


Nondetects And Data Analysis:
Multivariate Methods with NDs

Dennis R. Helsel, Ph.D

Practical Stats



Multivariate methods for censored data

1. Convert data to detect/nondetect (at the single, or highest of multiple, DLs). Cluster, test for differences in proportions of 0/1, etc.
2. Re-censor all values below a single (or the highest of multiple) detection limit as equal. Rank within each variable and perform MV analyses on the ranks.
3. Use u-scores to define relationships within multiply-censored variables. Perform MV analyses on the uscores (or their ranks).

Types of multivariate methods

Inter-dependence (all variables equivalent)

- NMDS
- Clustering

Dependence (response and explanatory variables)

- ANOSIM
- Seriation (nonparametric correlation / trend test)

Distance or Similarity Matrix

The basis underneath multivariate methods

Quantifies the similarity or distance between rows, or between columns

Triangular shape of unique cells, with #cells = r(r-1)/2 or c(c-1)/2 where r = # of rows and c = # of columns

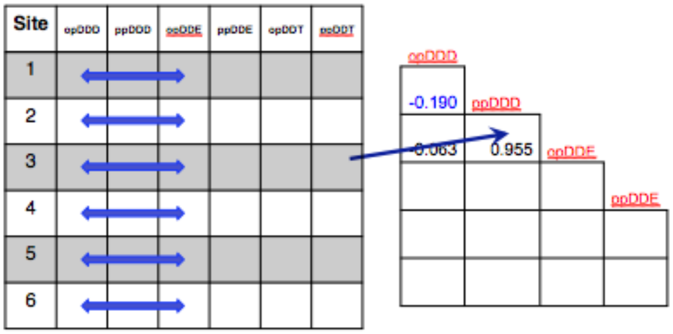
Similarities measured by something like correlation coefficients

Distance measures include Euclidean (straight-line) and Bray-Curtis distances between two points in multivariate space

R -mode

Which variables / columns have similar patterns?

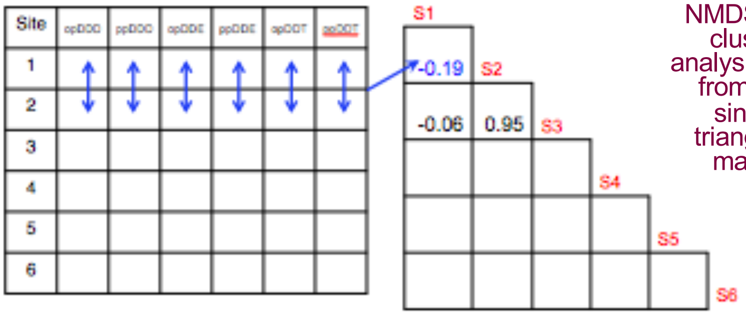
Shown below: Correlation between the 1st and 3rd columns (two chemical compounds)



Q-mode

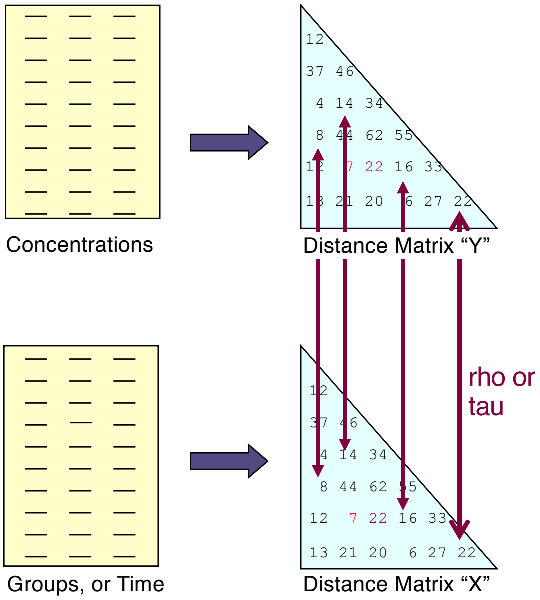
Which sites / dates / observations / rows have similar patterns?

Shown below: similarity between the 1st and 2nd rows (sites S1 and S2)



NMDS and cluster analysis work from this single triangular matrix

Relating two sets of data: Mantel test



ANOSIM, Seriation:

Compute Spearman’s rho or Kendall's tau between elements of the two triangular Distance matrices. One from the Y (concentration) variables, the other X from a group assignment or continuous variable (Time, etc.)

You should NOT use Pearson’s r correlation for this test -- there’s no reason the (X,Y) pairs of elements should have a linear shape with normally distributed residuals, as required by Pearson’s r.

Distance Measures for Censored Data

Distance Measure	characteristic
A. Simple matching binomial	above/below RL
B. Euclidean distance ordinal	ranks
C. Euclidean distance Wilcoxon	uscores

A. Methods for Binary Data

- Compute the distance measure
- Compute 4 multivariate procedures
  1. NMDS ordination plot
  2. ANOSIM -- MV Kruskal-Wallis test for group differences
  3. Cluster Analysis
  4. Trend Analysis

Simple matching coeff (similarity measure) between rows j and k

$$S_{jk} = {(a+d)}/{(a+b+c+d)}$$

where $a_{jk} = (1,1)$, $d_{jk} = (0,0)$ Matches

$b_{jk} = (0,1)$ $c_{jk} = (1,0)$ Mismatches

$1 = \text{value} < \text{RL}$, $0 = \text{value} \geq \text{RL}$

Dissimilarity (distance) measure $D_{jk} = 1 - S_{jk}$

Example data: FishDDT.xls

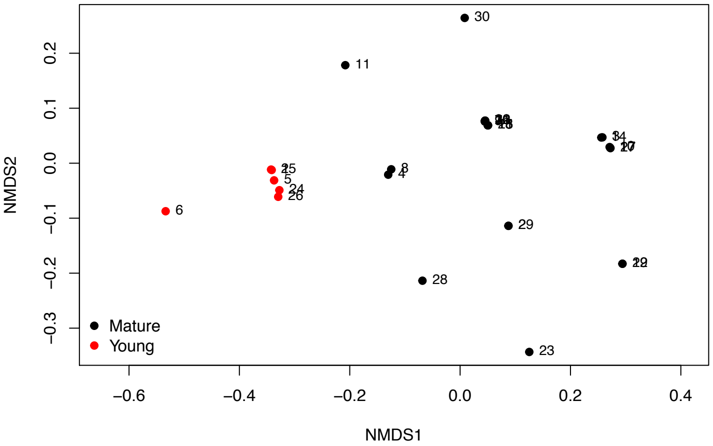
concentrations — group — indicators (1 = nondetect)

```
> head(FishDDT)
  opDDD ppDDD opDDE ppDDE opDDT ppDDT Age      Date oD_LT1 pD_LT1 oE_LT1 pE_LT1 oT_LT1 pT_LT1
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1   1       1   1      14     1     1 Young  1996        1      1      1      0      1      1
2   1      42   8.4   130     1    31 Mature 1990.       1      0      0      0      1      0
3   5.3    38   1     250     1    11 Mature 1994.       0      0      1      0      1      0
4   1      12   1      57     1     1 Mature 2002.       1      0      1      0      1      1
5   1       1   1      16     1     1 Young  2000.       1      1      1      0      1      1
6   1       1   1       1     1     1 Young  2000.       1      1      1      1      1      1
```

NMDS on simple-matching distances

```
> DDTcen <- data.frame(oD_LT1, pD_LT1, oE_LT1, pE_LT1, oT_LT1, pT_LT1)
> binaryMDS (DDTcen, Age, title = "NMDS of DDT in fish")
```

- Create a data frame just with the indicators
- Distances between points on the 'map' are in the same rank order as distances (1-S_{jk}) between rows
- There is definitely a left-right differentiation between Mature and Young fish.
- The overlap of some sites is due to ties between pairs -- there's only 6 variables, each with only two possible values, 0 and 1.



ANOSIM: testing group differences in patterns of 0/1s

test statistic $$R = \frac{(\overline{rank}_b - \overline{rank}_w)}{n(n-1)/4}$$

where $\overline{\text{rank}}$ = average of ranks of distances

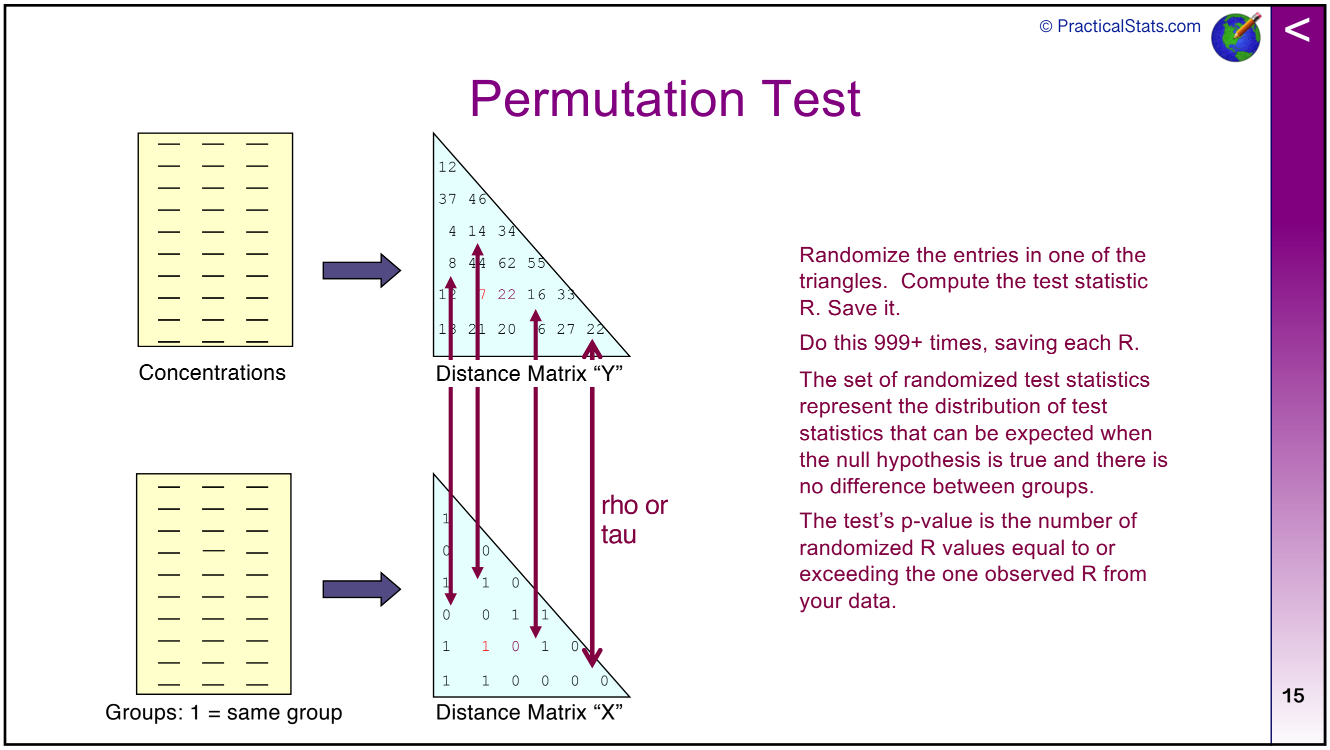
b = between different groups

w = within same group

ANOSIM: testing group differences in patterns of 0/1s

```
> DDTdissim <- binaryDiss (DDTcen)     # dissimilarity matrix.  Doesn’t include the
                                         grouping variable
> ano.ddt <- anosim(DDTdissim, Age)    # requires dissimilarity matrix as input
> ano.ddt
anosim(x = DDTdissim, grouping = Age)
Dissimilarity: simplematch
ANOSIM statistic R: 0.6876
      Significance: 0.001
Number of permutations: 999
```

The ANOSIM permutation test compares the observed test statistic R to 999 R permutations after randomization of the group assignment. Permutations represent the null hypothesis of no difference between groups.



© PracticalStats.com

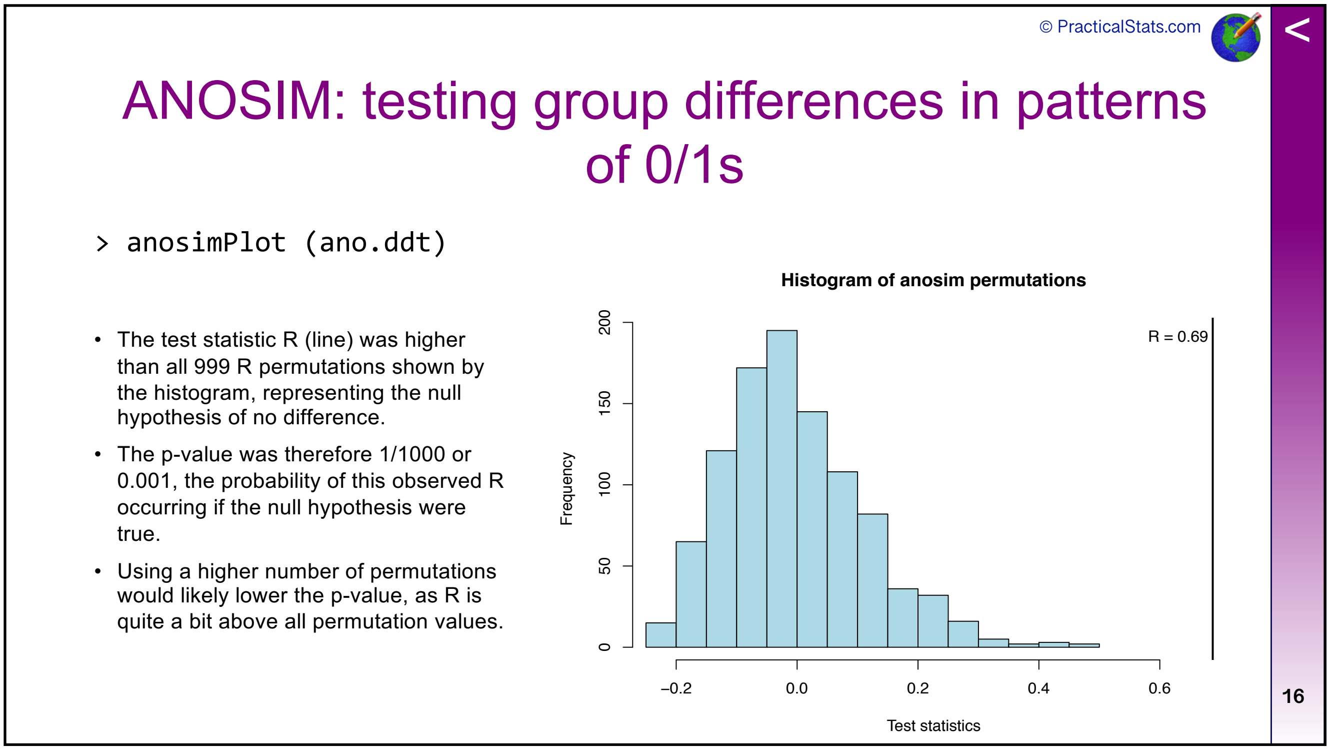
Permutation Test

Randomize the entries in one of the triangles. Compute the test statistic R. Save it.

Do this 999+ times, saving each R.

The set of randomized test statistics represent the distribution of test statistics that can be expected when the null hypothesis is true and there is no difference between groups.

The test’s p-value is the number of randomized R values equal to or exceeding the one observed R from your data.



© PracticalStats.com

ANOSIM: testing group differences in patterns of 0/1s

```
> anosimPlot (ano.ddt)
```

- The test statistic R (line) was higher than all 999 R permutations shown by the histogram, representing the null hypothesis of no difference.
- The p-value was therefore 1/1000 or 0.001, the probability of this observed R occurring if the null hypothesis were true.
- Using a higher number of permutations would likely lower the p-value, as R is quite a bit above all permutation values.

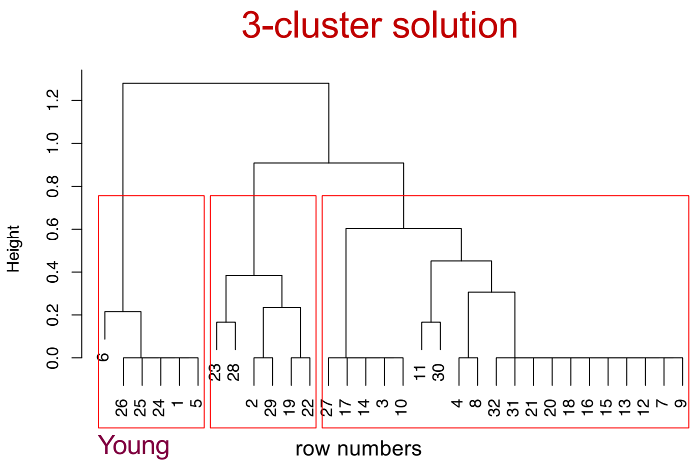
Histogram of anosim permutations

© PracticalStats.com

Clustering 0/1 data

```
> binaryClust(DDTcen, ncluster = 3)
```

All “Young” fish are together in the left-most cluster. The other two clusters are all “Mature” fish.



© PracticalStats.com

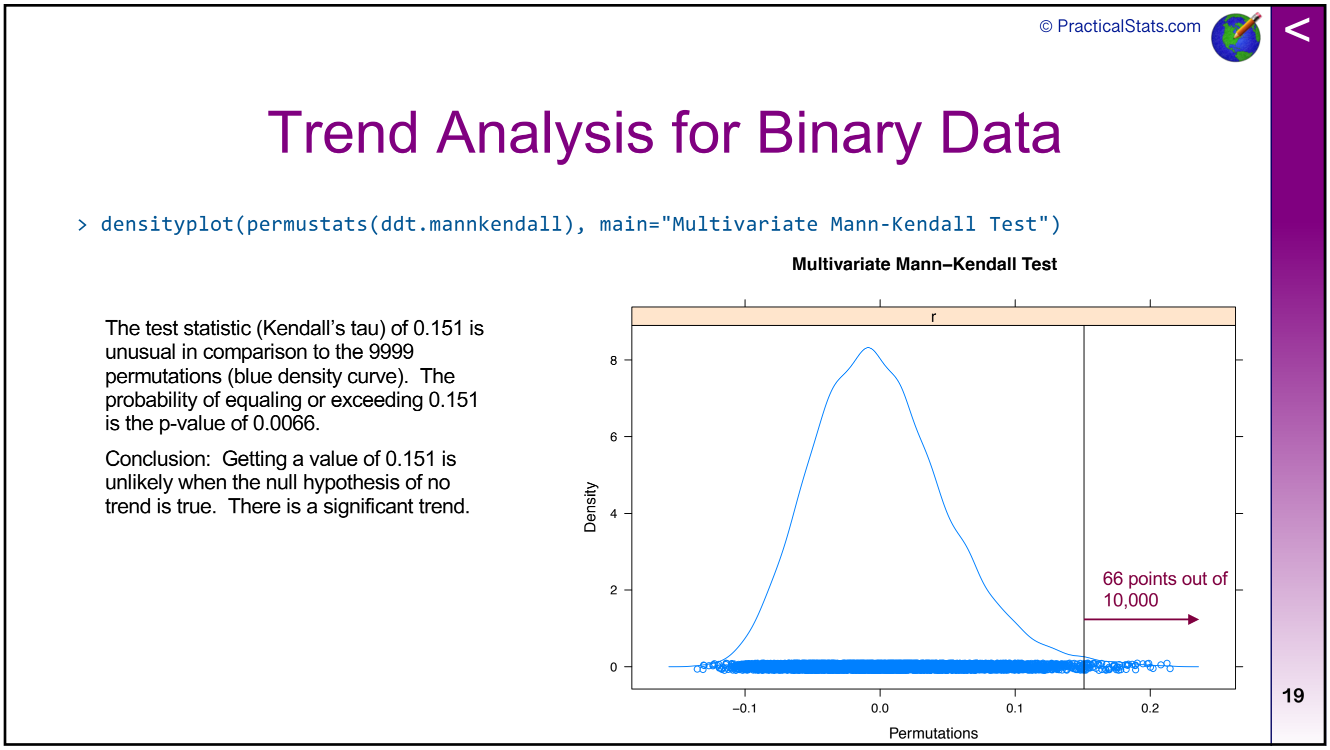
Trend Analysis for Binary Data

A multivariate version of the Mann-Kendall test for trend.

1. Compute the “X” triangular matrix of differences in the time variable. Use Manhattan distances (the difference between 1998 and 1999 will equal 1).
2. Compute a mantel test using Kendall’s tau between the distance matrix of the “Y” censoring pattern (DDTdissim) and the time distance matrix (time.dist).

```
> time.dist <- dist(Date, method = "manhattan")
> ddt.mannkendall <- mantel(time.dist, DDTdissim, method="kendall", permutations = 9999)
> ddt.mannkendall

Mantel statistic based on Kendall's rank correlation tau

Mantel statistic r: 0.151      Significance: 0.0066
Upper quantiles of permutations (null model):
    90%     95%   97.5%     99%
 0.0654  0.0886  0.1097  0.1380
Permutation: free Number of permutations: 9999
```

A significant change in the 0/1 pattern over time.



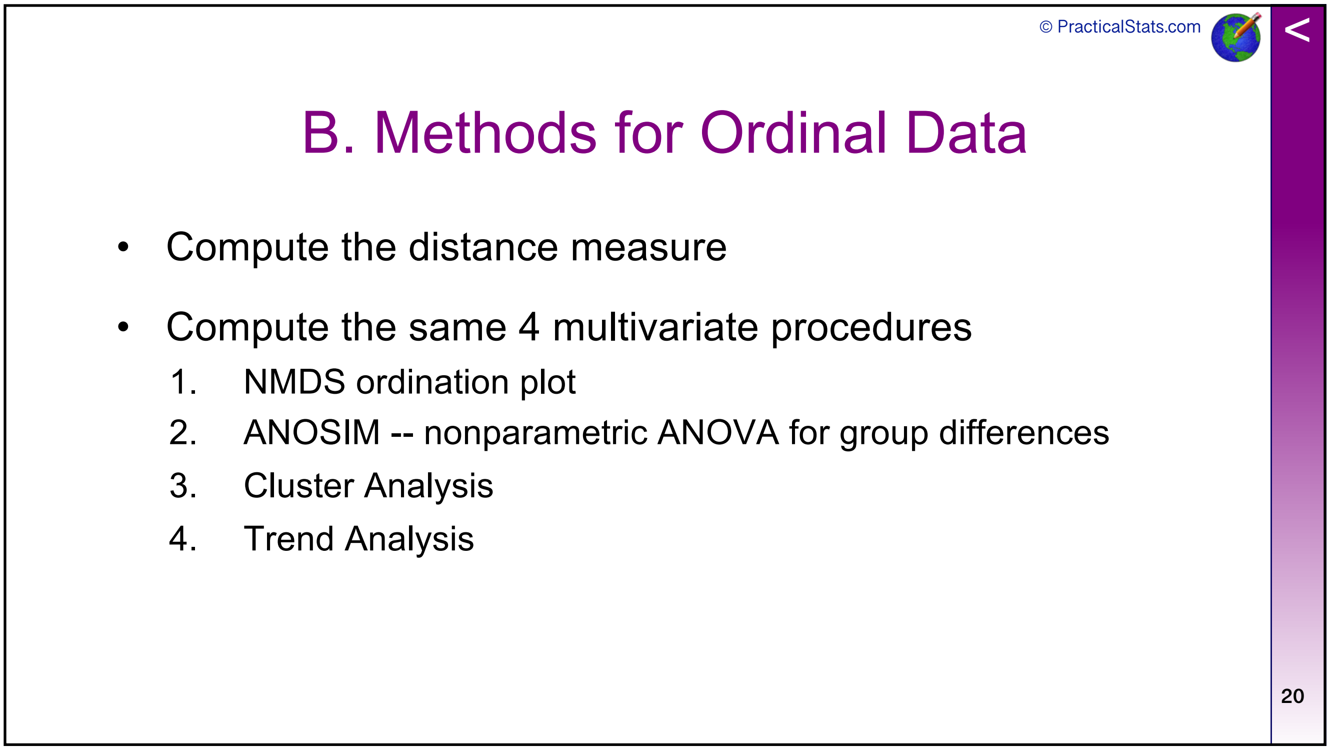
Trend Analysis for Binary Data

```
> densityplot(permustats(ddt.mannkendall), main="Multivariate Mann-Kendall Test")
```

The test statistic (Kendall's tau) of 0.151 is unusual in comparison to the 9999 permutations (blue density curve). The probability of equaling or exceeding 0.151 is the p-value of 0.0066.

Conclusion: Getting a value of 0.151 is unlikely when the null hypothesis of no trend is true. There is a significant trend.

66 points out of 10,000



B. Methods for Ordinal Data

- Compute the distance measure
- Compute the same 4 multivariate procedures
  1. NMDS ordination plot
  2. ANOSIM -- nonparametric ANOVA for group differences
  3. Cluster Analysis
  4. Trend Analysis

© PracticalStats.com

B. Distance Measure for Ordinal methods

- Rank data within each variable (column). The `ordranks.R` script makes this quick.
- Input to ordranks is a dataframe with both the concentration and associated indicator columns. Format can be `C1 I1 C2 I2 C3 I3` (`paired = TRUE`, the default) or `C1 C2 C3 I1 I2 I3` (`paired = FALSE`).
- If there are multiple RLs, the highest must be used and data re-censored to it. All <RL are assigned a tied rank. The `ordranks.R` script makes this quick.
- Euclidean distance E on ranks used for MV analysis

$$E = \sqrt{\sum_i (y_{i1} - y_{i2})^2}$$

21

© PracticalStats.com

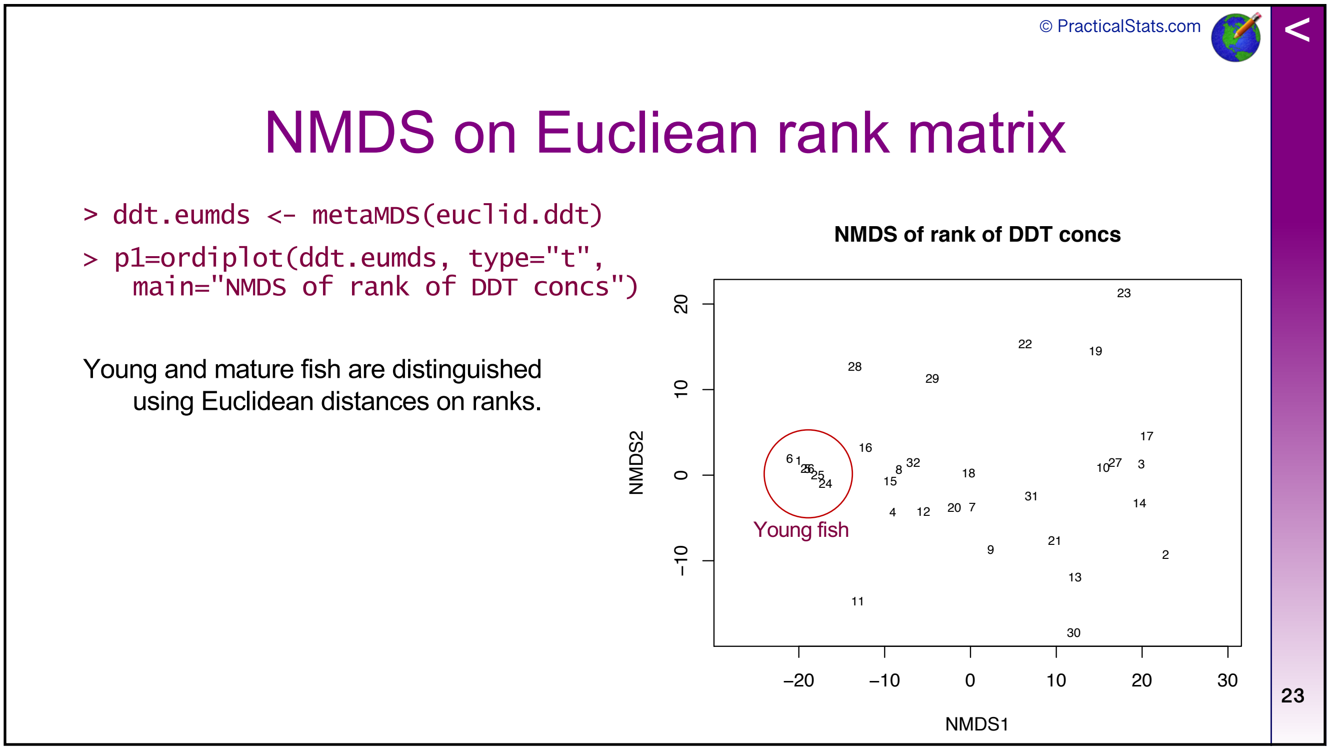
B. Distance Measure for Ordinal methods

```
> newFish <- FishDDT[, -(7:8)]  # removes Age, Date columns
> attach(newFish)
> ranks.ddt <- ordranks(newFish, paired = FALSE)   # will censor at highest DL in column
> euclid.ddt <- dist(ranks.ddt) # euclidean dist is default
> euclid.ddt
          1         2         3         4         5         6
2  46.502688
3  42.772655 26.692696
4  17.888544 36.366193 29.077483
5   2.500000 45.318319 41.118731 15.692355
6   2.000000 47.523678 44.153143 19.697716  4.500000
7  24.052027 26.748832 25.308101 16.807736 22.688103 25.268558
8  16.650826 35.654593 29.685855  6.726812 15.116216 18.034689
9  28.792360 26.485845 23.674881 18.681542 27.023138 30.282008
10 39.287403 23.097619 11.874342 29.385371 38.141185 40.292679
11 24.667793 41.922548 35.064227 17.219175 22.907422 26.162951  . . . . . . . . . . .
```

a triangular distance matrix

sites 1, 5, 6 are all Young fish and similar

22

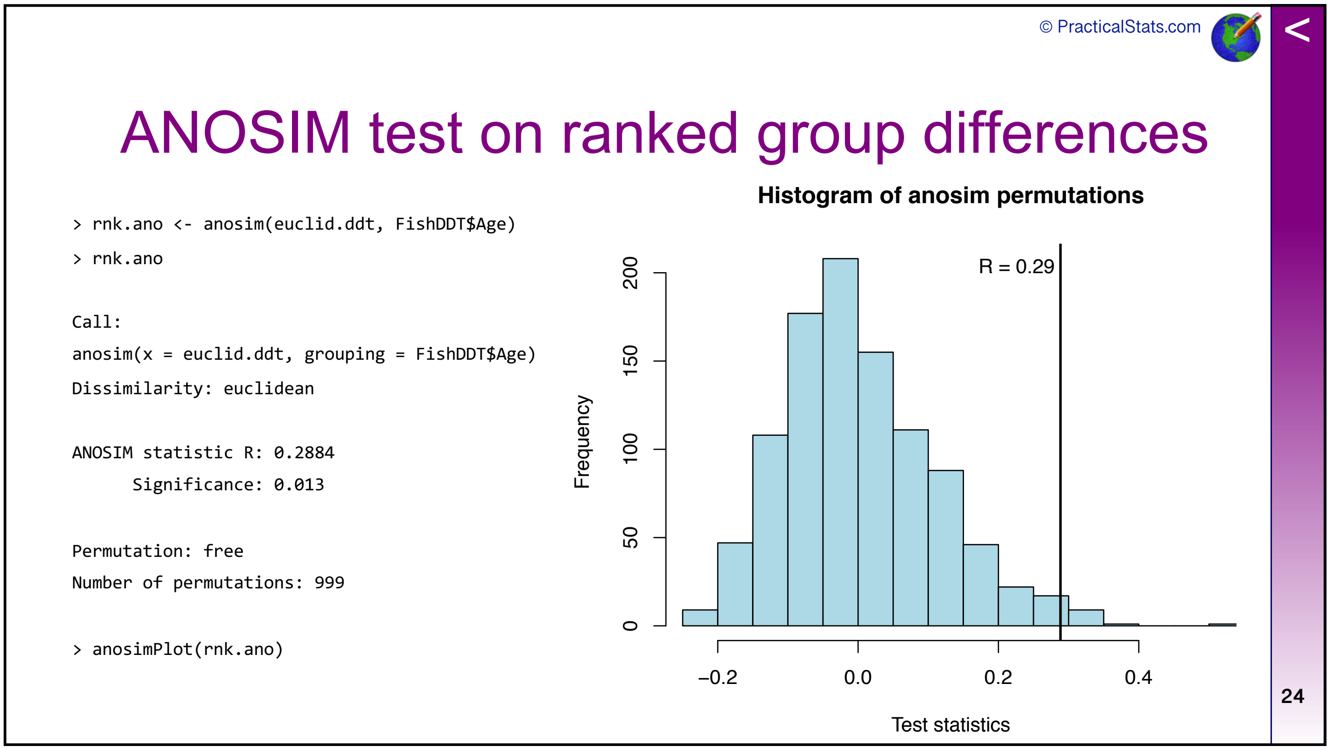


NMDS on Euclean rank matrix

```
> ddt.eumds <- metaMDS(euclid.ddt)
> p1=ordiplot(ddt.eumds, type="t",
    main="NMDS of rank of DDT concs")
```

Young and mature fish are distinguished using Euclidean distances on ranks.

NMDS of rank of DDT concs



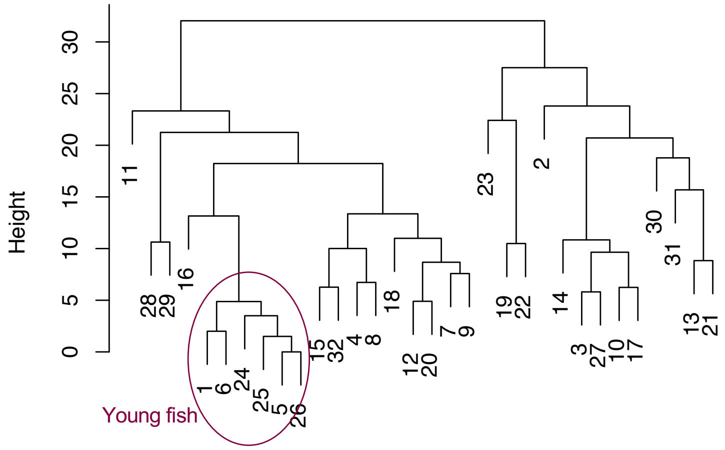
ANOSIM test on ranked group differences

```
> rnk.ano <- anosim(euclid.ddt, FishDDT$Age)
> rnk.ano

Call:
anosim(x = euclid.ddt, grouping = FishDDT$Age)
Dissimilarity: euclidean

ANOSIM statistic R: 0.2884
      Significance: 0.013

Permutation: free
Number of permutations: 999

> anosimPlot(rnk.ano)
```

Histogram of anosim permutations

Cluster on ranks of concentrations

```
> rankclust <- hclust(euclid.ddt, method = "average")
> plot(rankclust)
```



Trend test on ranks

- one triangle of distances between ranks of concentrations
- second triangle of distances between times (manhattan distances between times in years)
- Kendall's tau between the triangles. Are distances between ranks correlated with distances in time?

Trend test on ranks

```
> time.dist <- dist(FishDDT$Date, method = "manhattan")

> ddt.ranks <- mantel(time.dist, euclid.ddt, method="kendall", permutations =9999)

> ddt.ranks

Mantel statistic based on Kendall's rank correlation tau

mantel(xdis = time.dist, ydis = euclid.ddt, method = "kendall",      permutations = 9999)

Mantel statistic r: 0.3903                 r is higher than all of the permutation results!

      Significance: 1e-04                  resulting in a p-value < 0.001


Upper quantiles of permutations (null model):

   90%    95%  97.5%    99%

0.0538 0.0715 0.0893 0.1078

Number of permutations: 9999
```

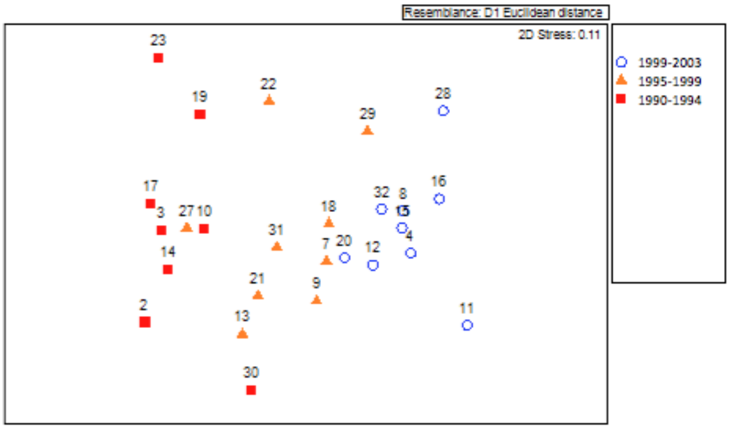
One way to illustrate trend test results

I created a category of 3 time periods to illustrate a time vector.

Drew an NMDS and colored the points by the time periods.

Multivariate trend test had p<0.001

Pattern in the ranks of concentrations changes with time, as shown by the first NMDS direction (x axis)



earlier → later

C. u-scores for multiple DL data

- Compute the distance measure
- Compute the same 4 multivariate procedures
  1. NMDS ordination plot
  2. ANOSIM -- nonparametric ANOVA for group differences
  3. Cluster Analysis
  4. Trend Analysis

Example Data with Multiple RLs

The FishDDT dataset was altered. Some <5 values were changed to have values from 0 to the method DL of 2 (0,2) and some to have values between the MDL of 2 and the Quant. Limit of 5 (2,5). Other <5 values were left alone, so are between 0 and 5 (0,5).

We’ll use the dataset FishDDTalt.rda (R format dataset)

By mistake, one value was also deleted, producing an NA for an upper concentration in row 6. The unintended consequences are discussed in the R example of the video.

1. Multiple RLs: u-scores

- U-scores are a measure of order within a variable
- They are the sum of the sign of differences between the ith observation and all other observations in the data set

$$u_i = \sum_{i \neq k} sign(x_i - x_k)$$

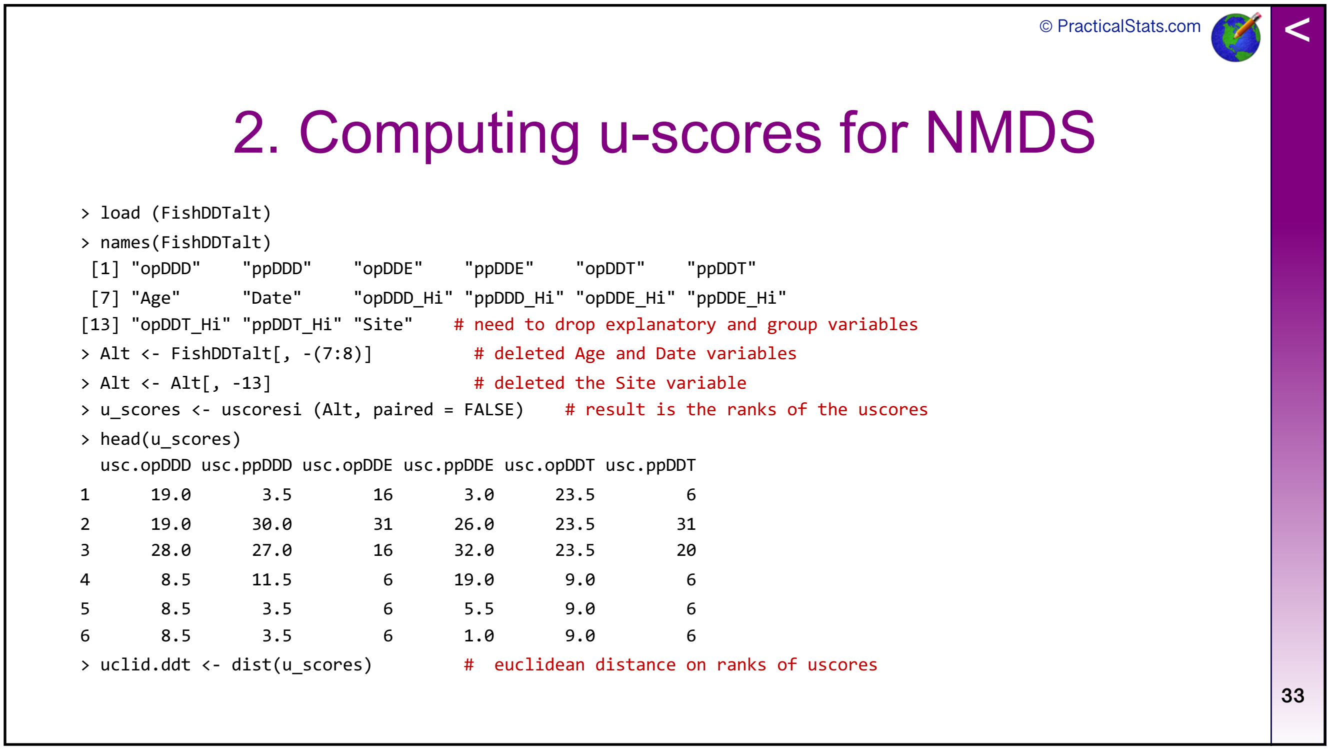
- To say it another way, the u-score is the number of observations the i^{th} value is above, minus the number of observations the i^{th} value is below. The larger the i^{th} observation's value, the higher the u-score. The median observation would have a u-score of 0.
- u-scores are the basis for the Mann-Whitney test, is the numerator in computation of Kendall's tau correlation, and are related to Kaplan-Meier and other nonparametric methods
- The distance measure will be the Euclidean distance on either the u-scores, or the ranks of the u-scores. The latter insures all values are positive, which may be required by some multivariate software.

U-score functions in NADA2

uscores.R: computes u-scores for data using the 0/1 indicator column format. A <1 means the value is between 0 and 1 (the lower limit is always 0)

uscoresi.R: computes u-scores for data using the (lo, hi) interval censoring format. This allows inclusion of data in the interval (Detect limit, Quantitation limit) and other nonzero lower limits.

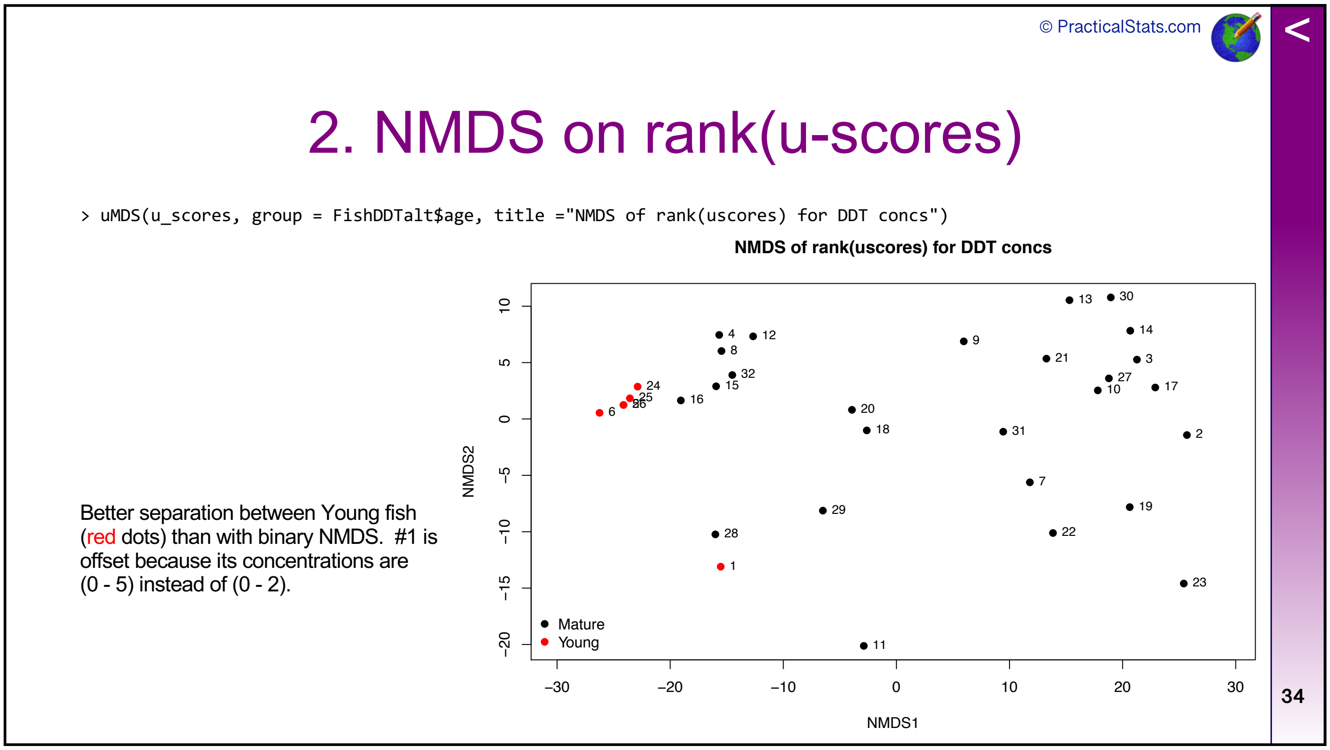
uMDS: computes Euclidean distances between scores or rank(scores) and plots the results on a Nonmetric Multi-Dimensional Scale plot. Has an option to color the points by a grouping variable.



© PracticalStats.com

2. Computing u-scores for NMDS

```
> load (FishDDTalt)
> names(FishDDTalt)
 [1] "opDDD"    "ppDDD"    "opDDE"    "ppDDE"    "opDDT"    "ppDDT"
 [7] "Age"      "Date"     "opDDD_Hi" "ppDDD_Hi" "opDDE_Hi" "ppDDE_Hi"
[13] "opDDT_Hi" "ppDDT_Hi" "Site"    # need to drop explanatory and group variables
> Alt <- FishDDTalt[, -(7:8)]          # deleted Age and Date variables
> Alt <- Alt[, -13]                    # deleted the Site variable
> u_scores <- uscoresi (Alt, paired = FALSE)    # result is the ranks of the uscores
> head(u_scores)
  usc.opDDD usc.ppDDD usc.opDDE usc.ppDDE usc.opDDT usc.ppDDT
1      19.0       3.5        16       3.0      23.5         6
2      19.0      30.0        31      26.0      23.5        31
3      28.0      27.0        16      32.0      23.5        20
4       8.5      11.5         6      19.0       9.0         6
5       8.5       3.5         6       5.5       9.0         6
6       8.5       3.5         6       1.0       9.0         6
> uclid.ddt <- dist(u_scores)         #  euclidean distance on ranks of uscores
```



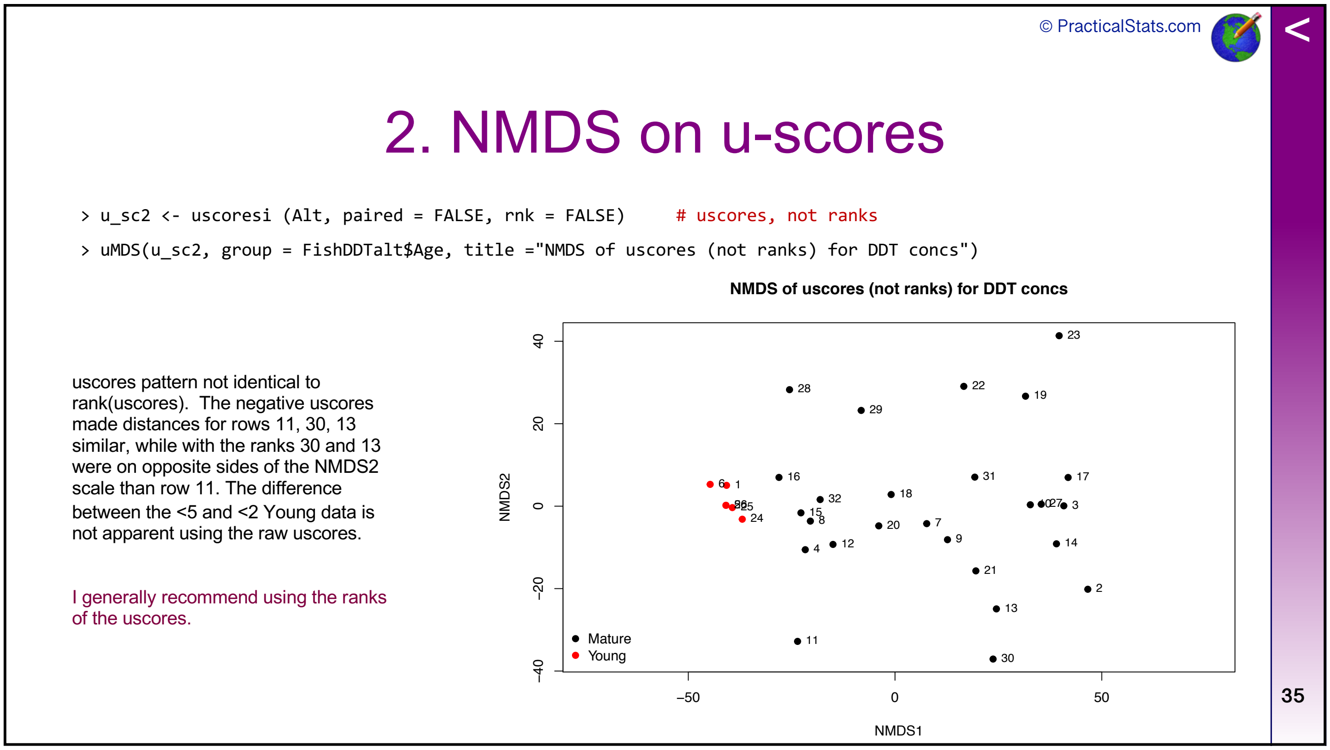
© PracticalStats.com

2. NMDS on rank(u-scores)

```
> uMDS(u_scores, group = FishDDTalt$age, title ="NMDS of rank(uscores) for DDT concs")
```

NMDS of rank(uscores) for DDT concs

Better separation between Young fish (red dots) than with binary NMDS. #1 is offset because its concentrations are (0 - 5) instead of (0 - 2).

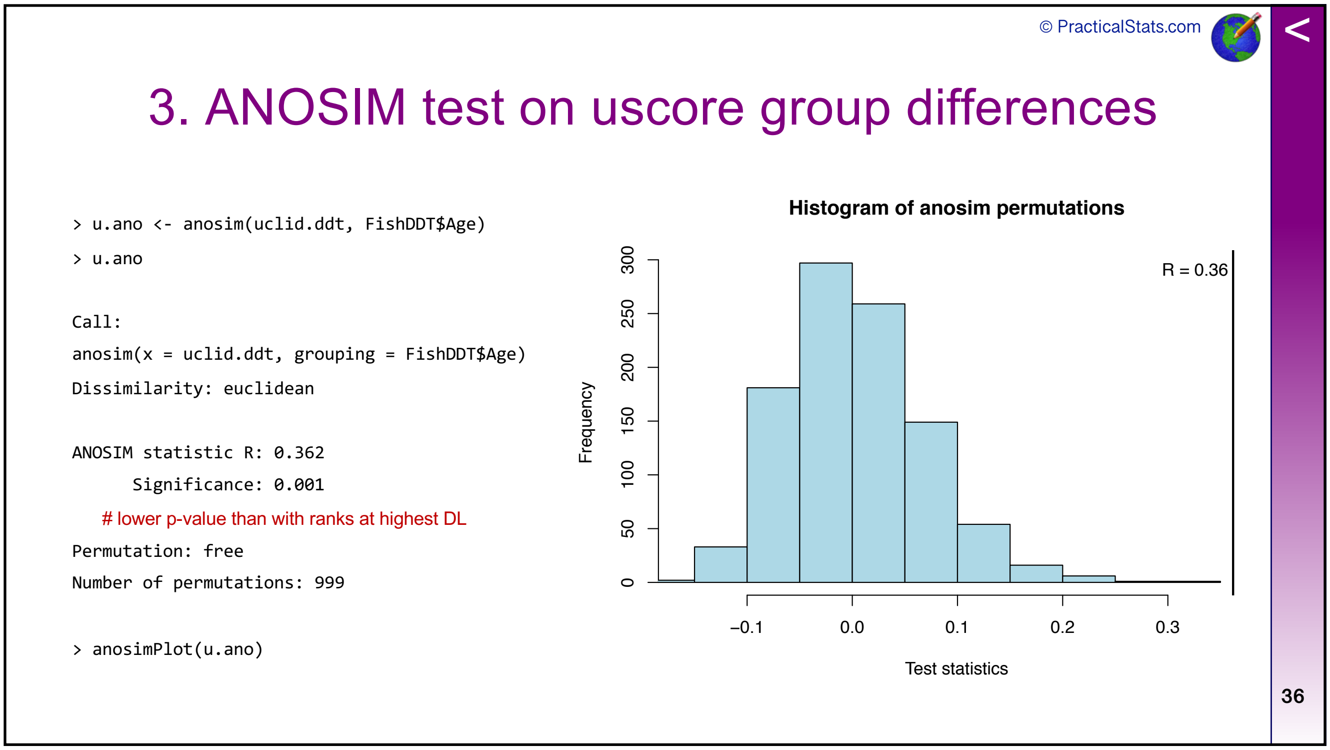


2. NMDS on u-scores

```
> u_sc2 <- uscoresi (Alt, paired = FALSE, rnk = FALSE)     # uscores, not ranks
> uMDS(u_sc2, group = FishDDTalt$Age, title ="NMDS of uscores (not ranks) for DDT concs")
```

uscores pattern not identical to rank(uscores). The negative uscores made distances for rows 11, 30, 13 similar, while with the ranks 30 and 13 were on opposite sides of the NMDS2 scale than row 11. The difference between the <5 and <2 Young data is not apparent using the raw uscores.

I generally recommend using the ranks of the uscores.

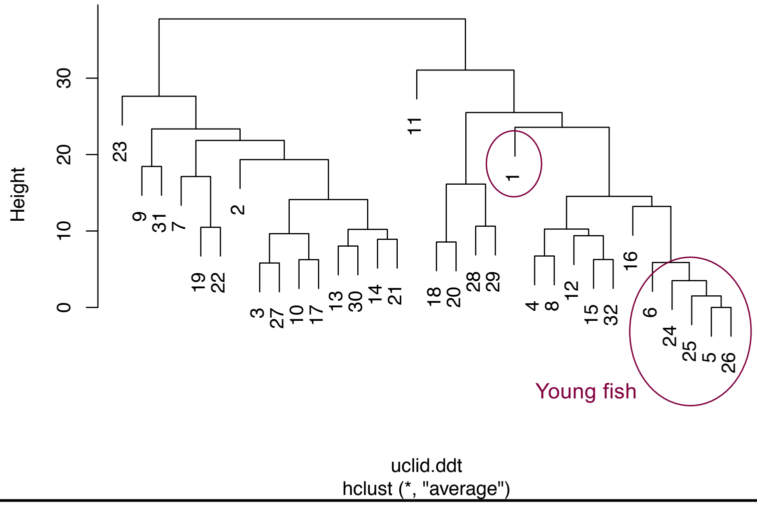


3. ANOSIM test on uscore group differences

```
> u.ano <- anosim(uclid.ddt, FishDDT$Age)
> u.ano

Call:
anosim(x = uclid.ddt, grouping = FishDDT$Age)
Dissimilarity: euclidean

ANOSIM statistic R: 0.362
      Significance: 0.001
   # lower p-value than with ranks at highest DL
Permutation: free
Number of permutations: 999

> anosimPlot(u.ano)
```

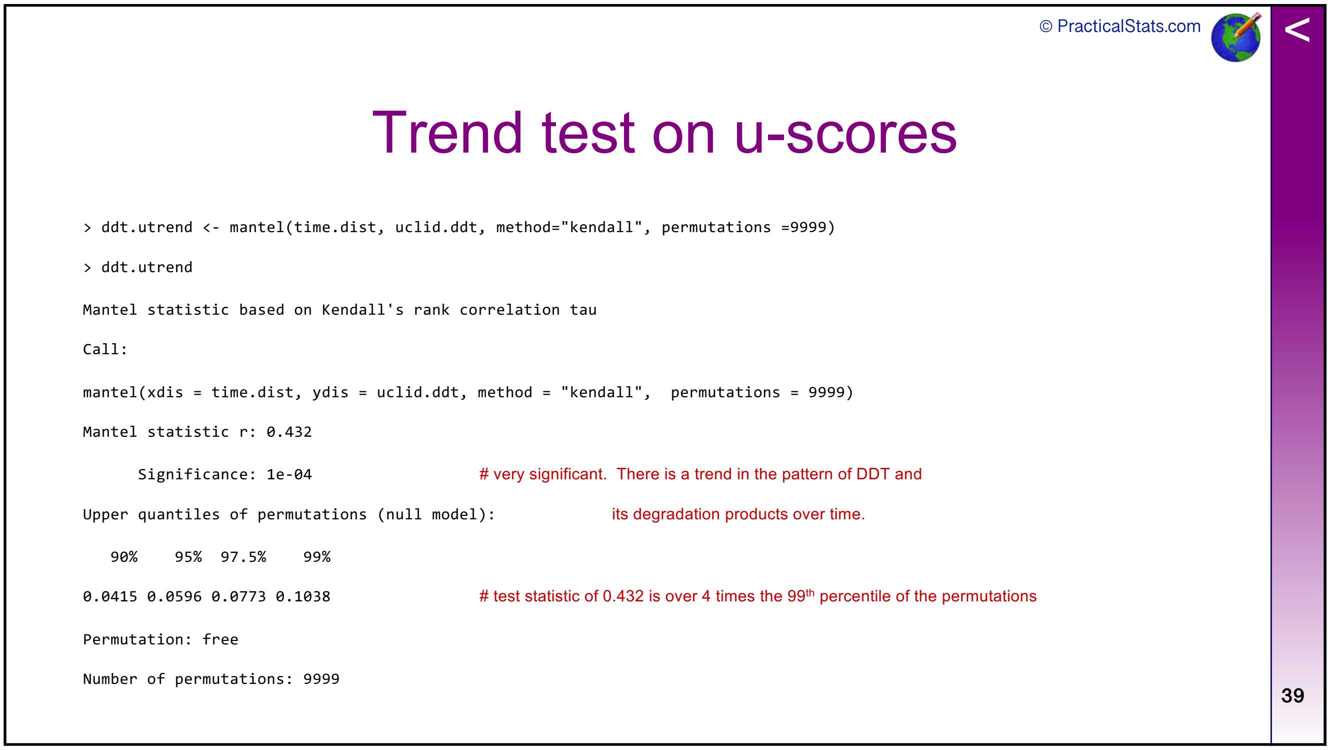
4. Cluster analysis using uscores

```
> uclust <- hclust(uclid.ddt, method = "average")
> plot(uclust)
```



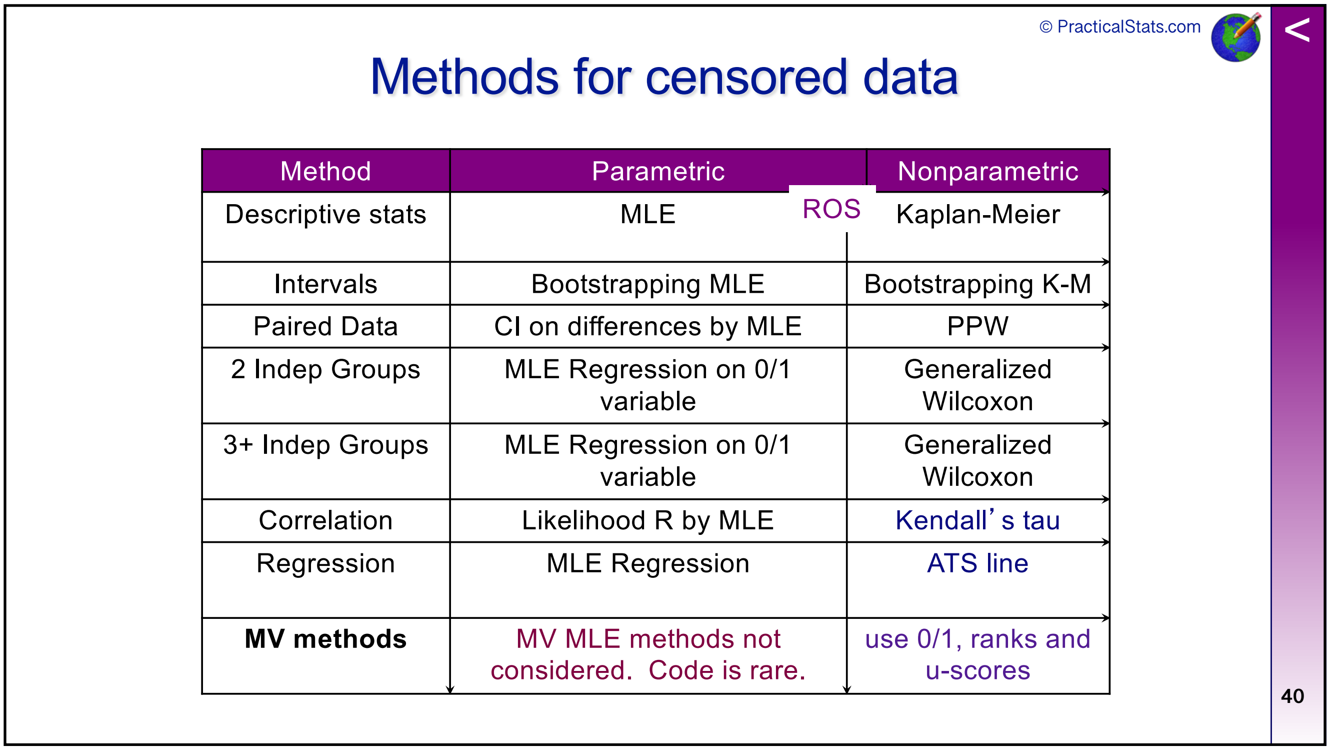
5. Trend analysis using u-scores

- one triangle of distances between u-scores of concentrations
- second triangle of distances between times (manhattan distances between times in years)
- Kendall's tau between the triangles. Are distances between ranks correlated with distances in time?



Trend test on u-scores

```
> ddt.utrend <- mantel(time.dist, uclid.ddt, method="kendall", permutations =9999)

> ddt.utrend

Mantel statistic based on Kendall's rank correlation tau

Call:

mantel(xdis = time.dist, ydis = uclid.ddt, method = "kendall",  permutations = 9999)

Mantel statistic r: 0.432

      Significance: 1e-04                   # very significant.  There is a trend in the pattern of DDT and

Upper quantiles of permutations (null model):              its degradation products over time.

   90%    95%  97.5%    99%

0.0415 0.0596 0.0773 0.1038                 # test statistic of 0.432 is over 4 times the 99th percentile of the permutations

Permutation: free

Number of permutations: 9999
```



Methods for censored data

Method	Parametric	Nonparametric
Descriptive stats	MLE ROS	Kaplan-Meier
Intervals	Bootstrapping MLE	Bootstrapping K-M
Paired Data	CI on differences by MLE	PPW
2 Indep Groups	MLE Regression on 0/1 variable	Generalized Wilcoxon
3+ Indep Groups	MLE Regression on 0/1 variable	Generalized Wilcoxon
Correlation	Likelihood R by MLE	Kendall's tau
Regression	MLE Regression	ATS line
MV methods	MV MLE methods not considered. Code is rare.	use 0/1, ranks and u-scores