Visualizing Data

Ying Zhang
Statistical Consulting Center
Department of Mathematics and Statistics
Acadia University
April 24

Data Liberation Initiative Atlantic, 2009

Outline

- Objective Components of Visualizing Data
- Tools and Data Types
- Case Studies

  Case I: Growth of the Nova Scotia
  Blanding's Turtle

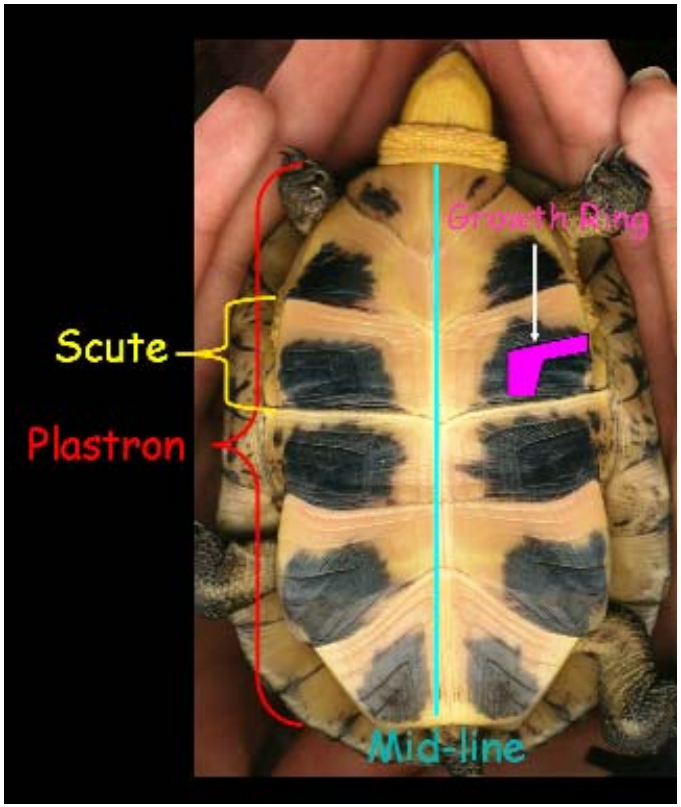
  Case II: Health Utility Index

Objective Components

- Data exploration
- Fitting
- Model diagnostics

Case I: Growth of the Blanding's Turtle

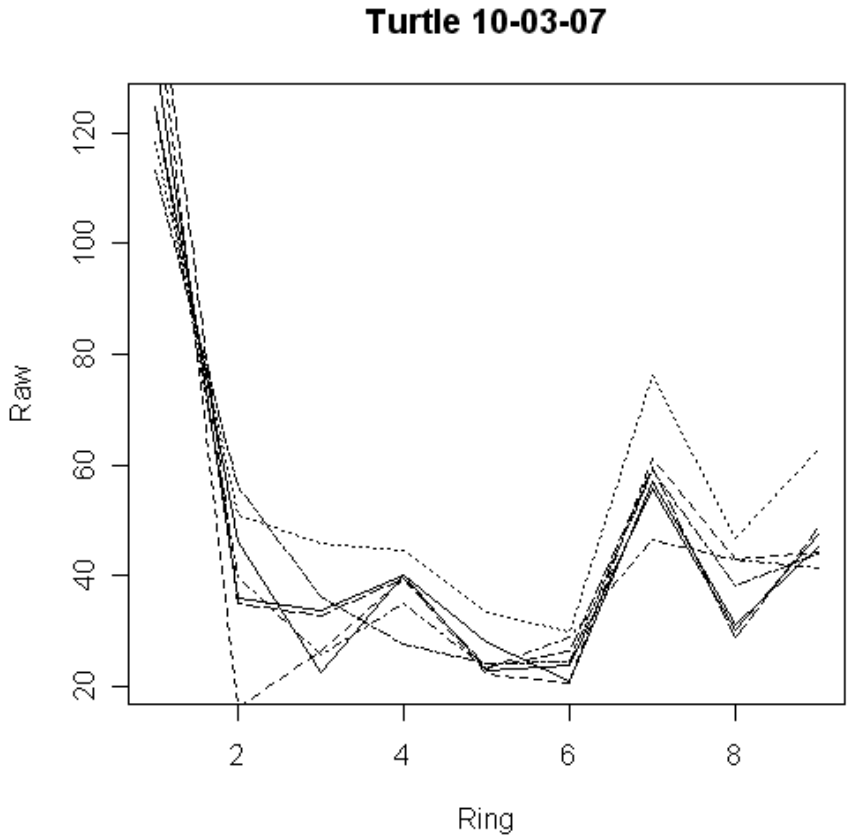
(Huang, Zhang, Cabilio, M. Richard, and Herman, 2009)

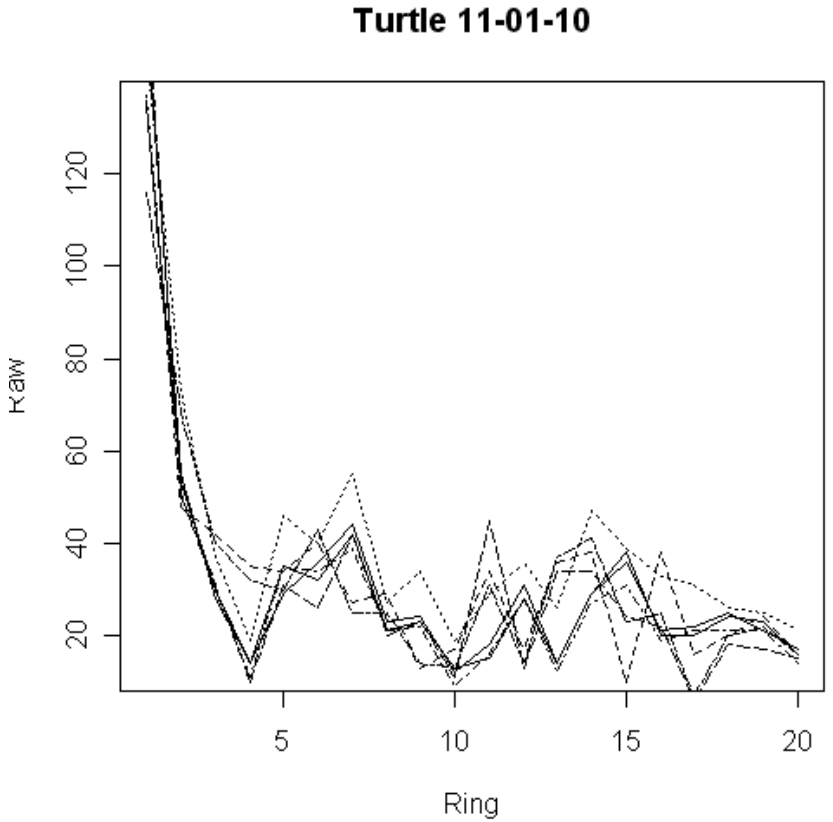


- They mature at age 20-25 (growth stops)
- 123 Blanding's turtles from 3 different locations in NS
  - 73 from Kejimkujik (KNP)
  - 18 from McGowan Lake (ML)
  - 42 from Pleasant River (PR)

Case I: Data Exploration

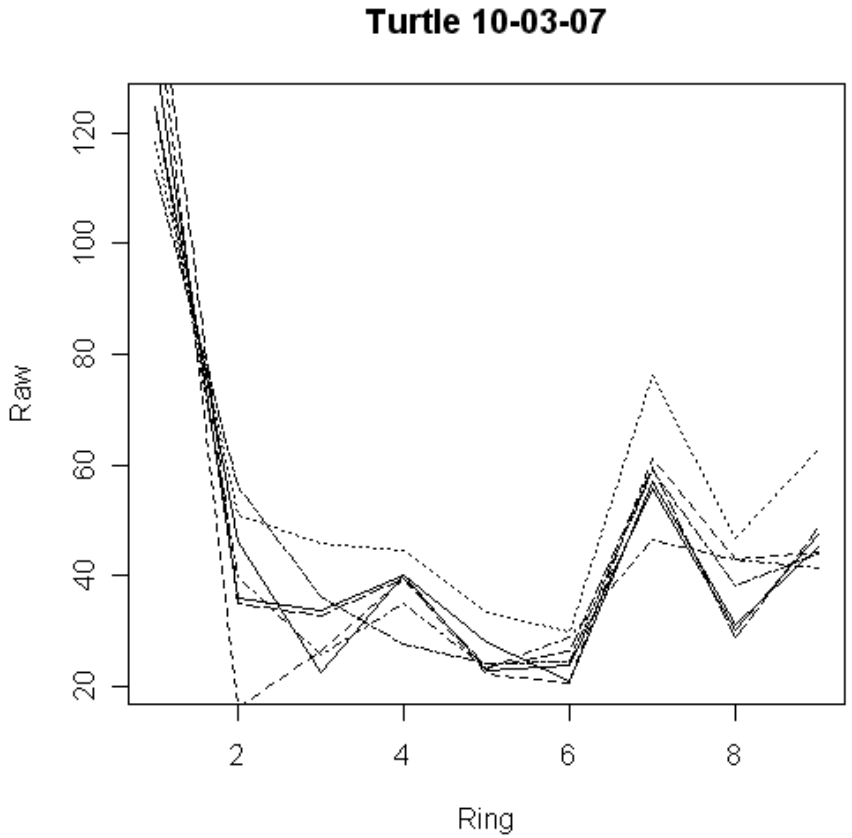
Growth Ring

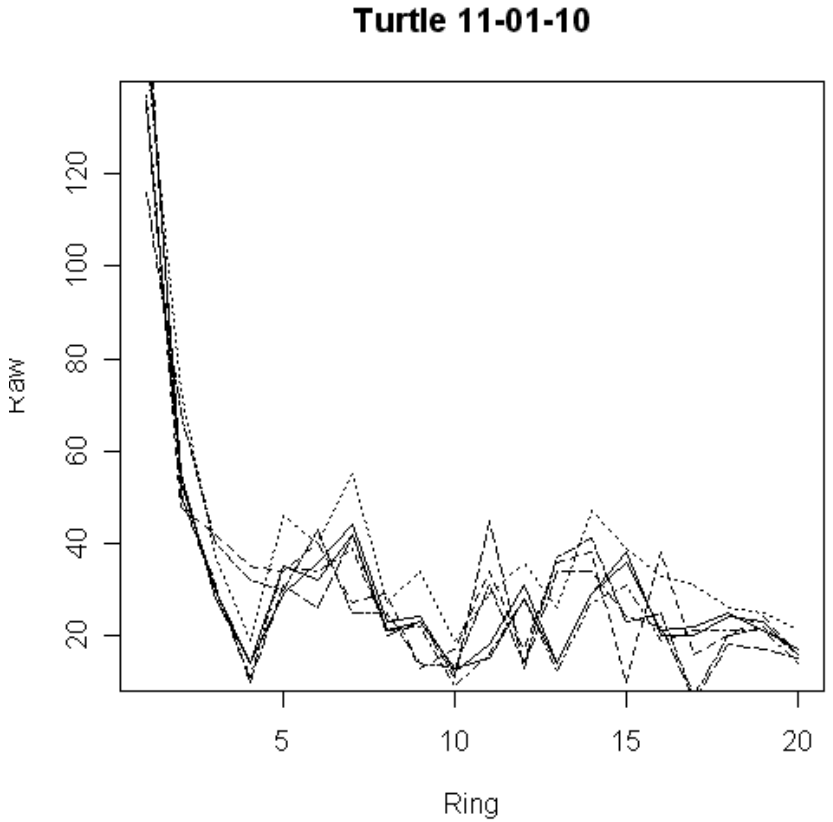




Case I: Data Exploration

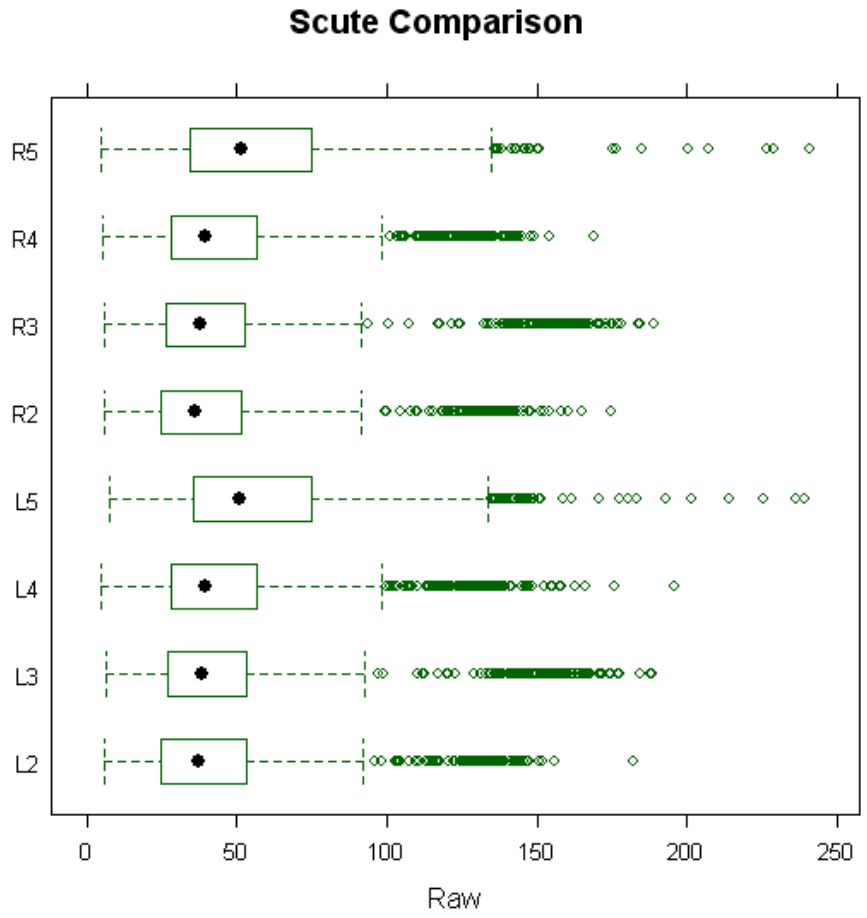
Growth Ring





Case I: Data Exploration

Growth Ring by Scute



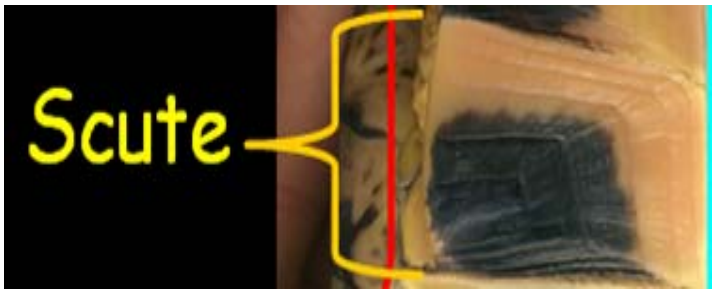
Case I: Fitting

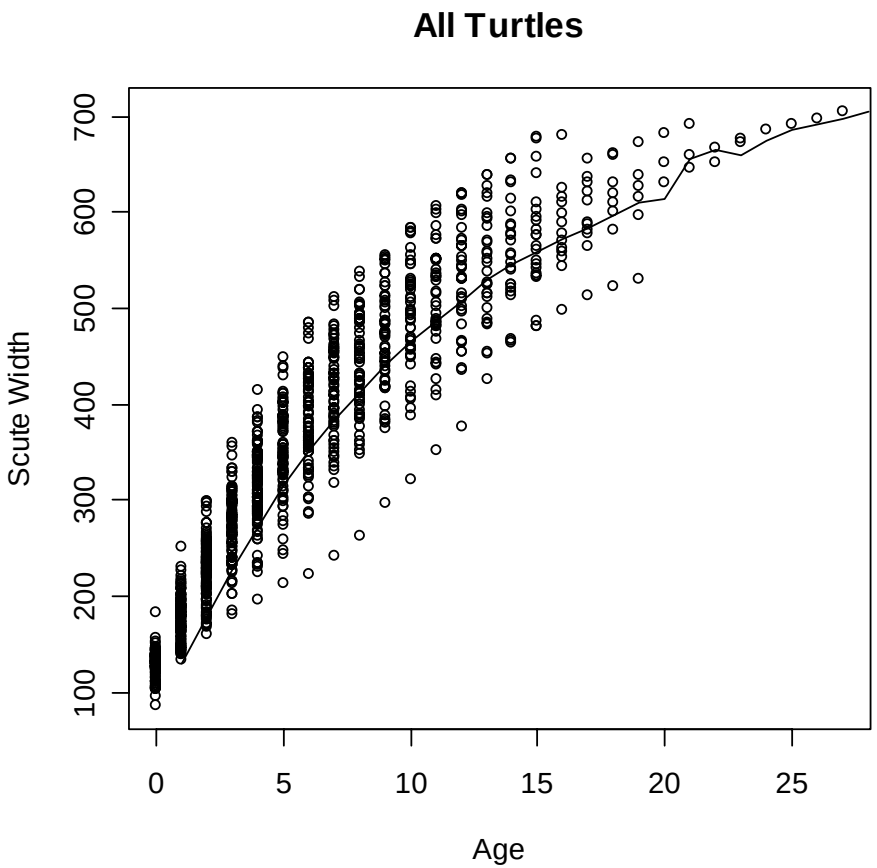
Accumulated Growth Curve

Accumulated Growth:

Scute Width

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





Case I: Fitting

Exponential Growth

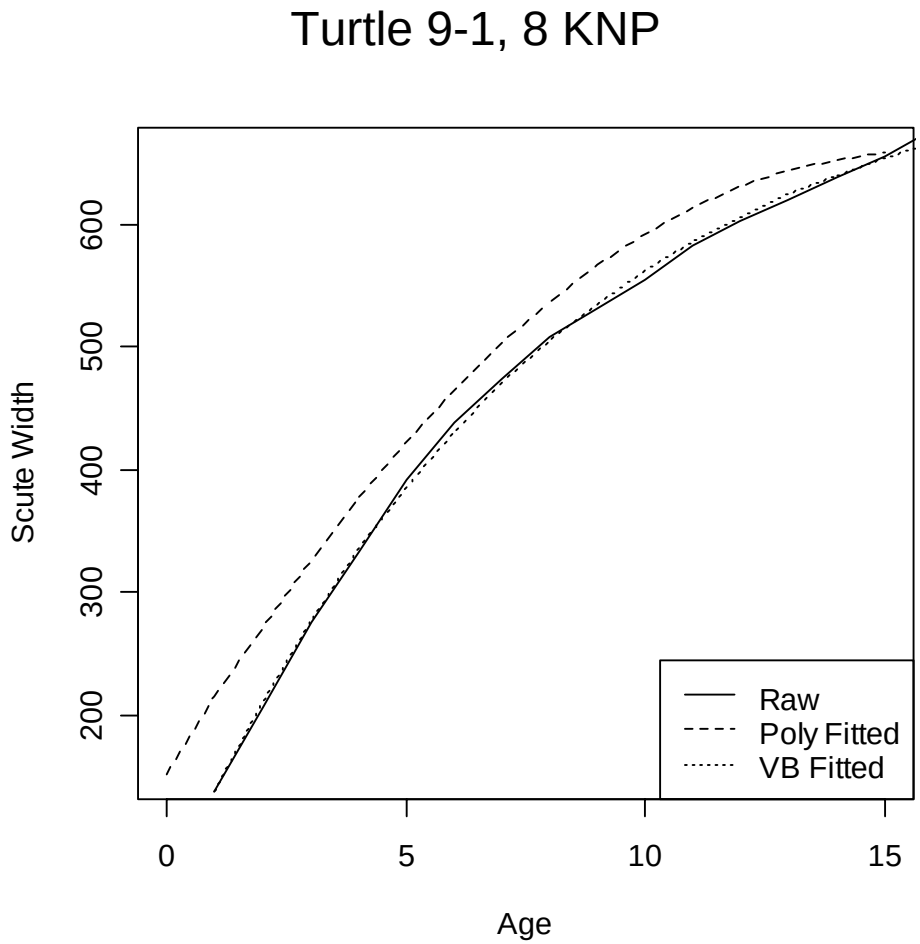
$$SW_t = a\,(1 - be^{rt})$$

a is asymptotic length;

b is a variable related to size at birth;

e is the base of the natural logarithms;

r is an intrinsic growth factor



Case II: Health Utilities Index (HUI)

(**Bu, 2006**)

- To measure health-related quality-of-life/ functional-status in clinical trials
- To describe and monitor the health of general populations
- http://healthutilities.biz/ lists lots of papers of the validation of HUI.

Case II: Health Utilities Index (HUI)

Attributes (7):	Attributes (8):
Sensation (4 levels)	Vision (6 levels)
Mobility (5 levels)	Hearing (6 levels)
Emotion (5 levels)	Speech (5 levels)
Cognition (4 levels)	Ambulation (6 levels)
Self-Care (4 levels)	Dexterity (6 levels)
Pain (5 levels)	Emotion (5 levels)
Fertility (3 levels)	Cognition (6 levels)
	Pain (5 levels)
*Note: HUI2 describes 24,000 unique health states.	*Note: HUI3 describes 972,000 unique health states.

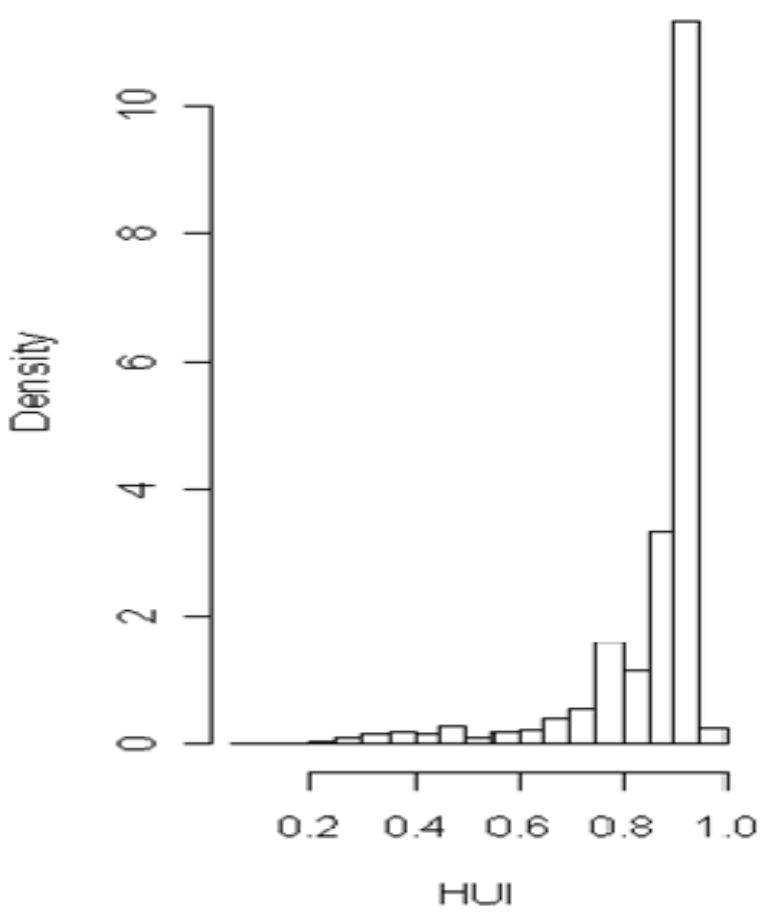
Case II: Health Utilities Index (HUI)

HUI is applicable to most people. It provides descriptive health profile measures and HRQL scores on a generic scale. HUI also provides single-attribute scores of morbidity for each attribute.

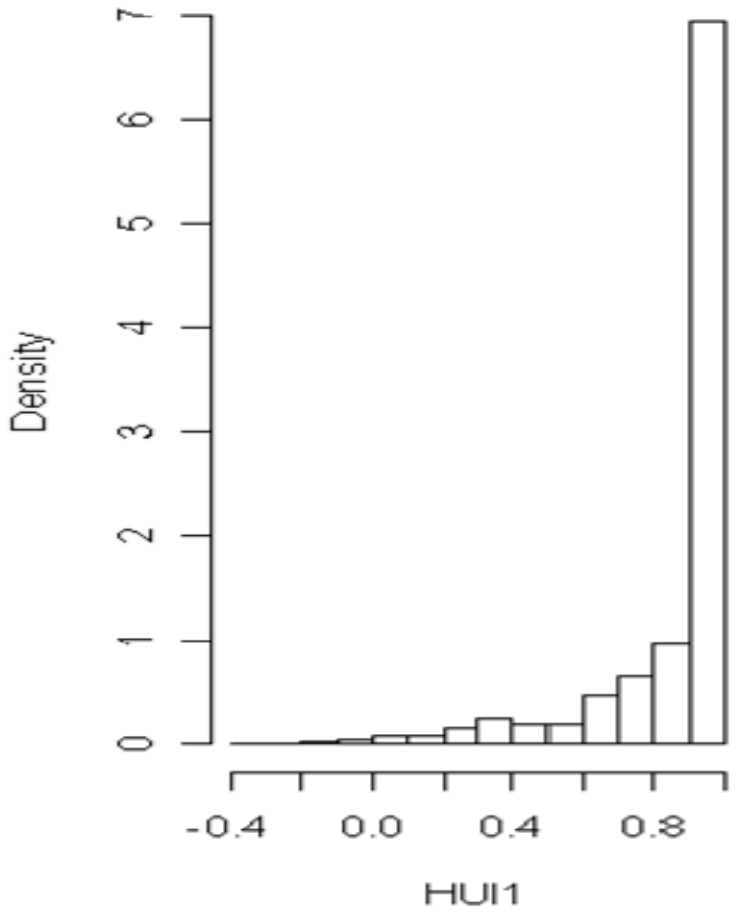
Each HUI attribute (dimension) has 3-6 levels of discrimination and is very responsive to changes in health caused by treatment therapies or other influences.

Case II: Model diagnostics

HUI 1996



HUI 2000



Data resource: National Population Health Survey (NPHS) 1996 & Canadian Community Health Survey (CCHS) 2000

Tools and Data Types

(Cleveland, 1993)

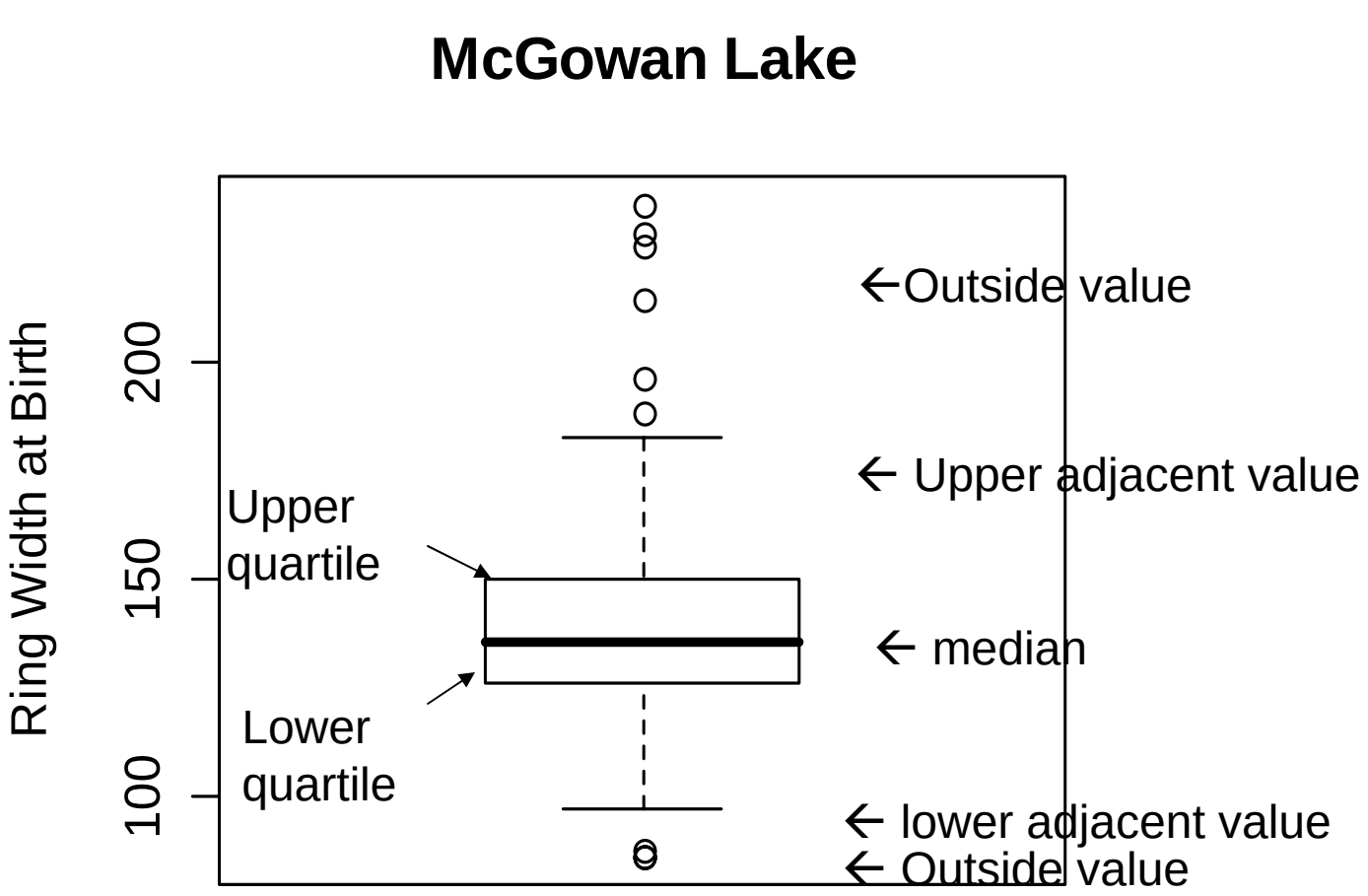
- Univariate Data
- Bivariate Data
- Multi-way Data

Learning the Distribution of the Univariate Data

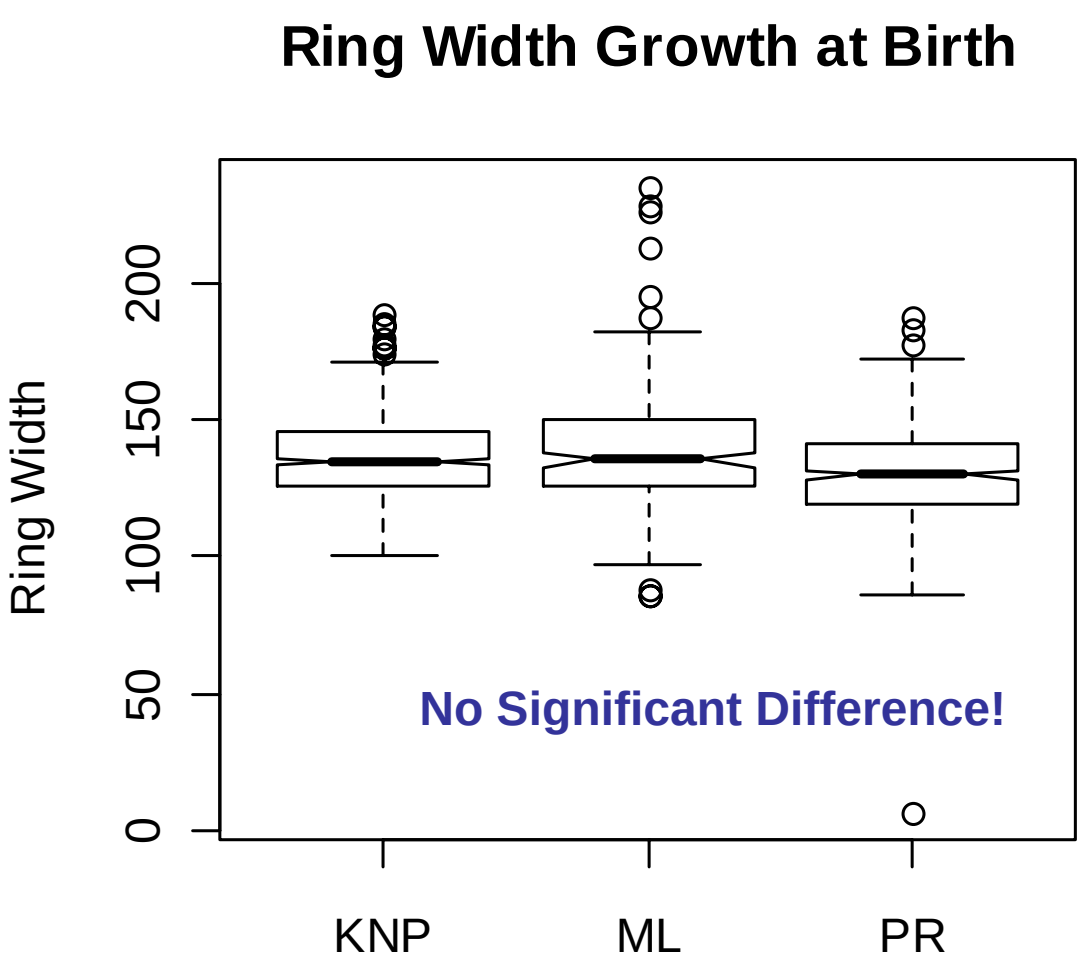
Box Plot

- A graphical display of the quartiles and the extreme values of a sample.
  - To catch features of a sample: *Location, Spread, Skewness, Outliers*
  - Very useful for comparing several data sets with respect to these characteristics

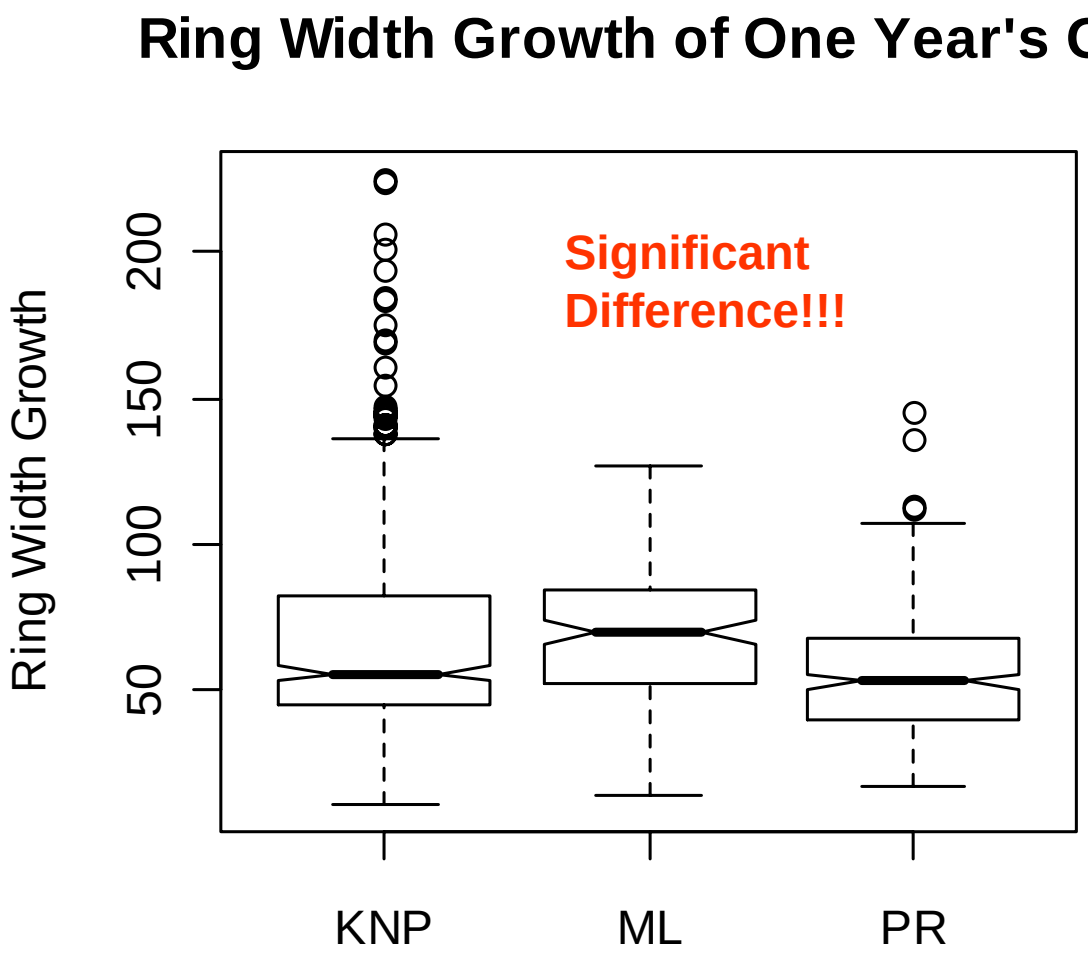
Box plot


McGowan Lake
Ring Width at Birth
100
150
200
←Outside value
← Upper adjacent value
Upper
quartile
← median
Lower
quartile
← lower adjacent value
← Outside value

Box Plot

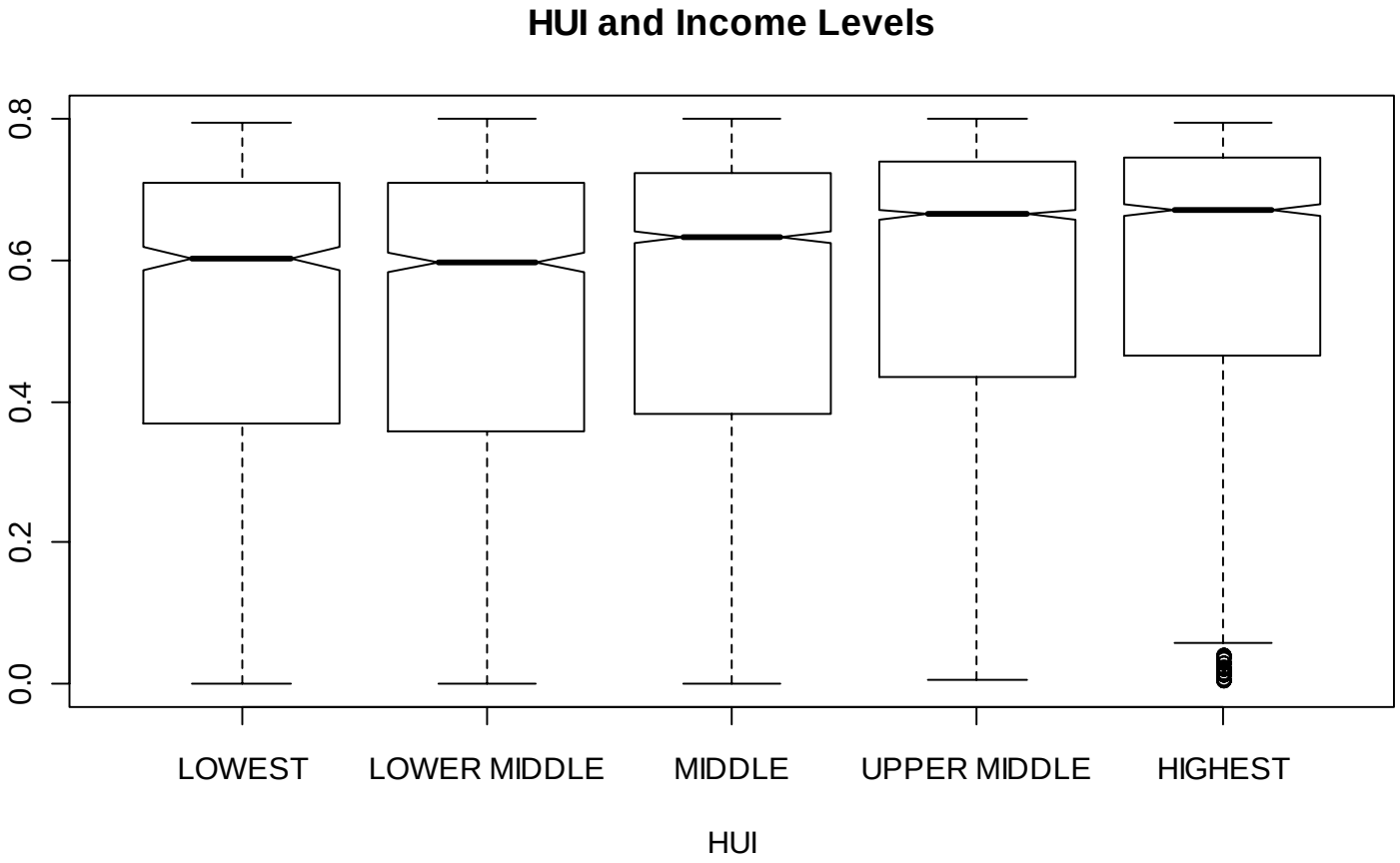

Ring Width Growth at Birth
Ring Width
0
50
100
150
200
KNP
ML
PR
No Significant Difference!

Box Plots


Ring Width Growth of One Year's Old
Ring Width Growth
50
100
150
200
Significant
Difference!!!
KNP
ML
PR

Box Plot

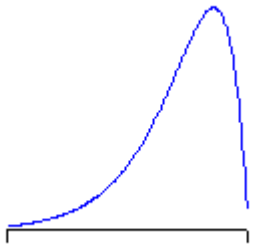
- **The richer, the healthier?**



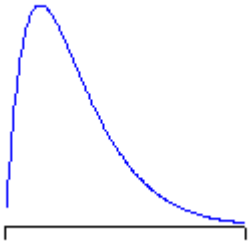
Histogram

- Relative frequency histogram
  - A picture of the relative frequency of that sample
- Density histogram
  - The relative frequency of a class is equal to the area of the rectangle above that class.
- Very useful for learning distribution shape

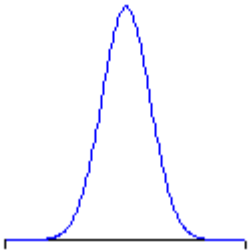
Some Typical Sample Shapes



Right Skewed



Left Skewed

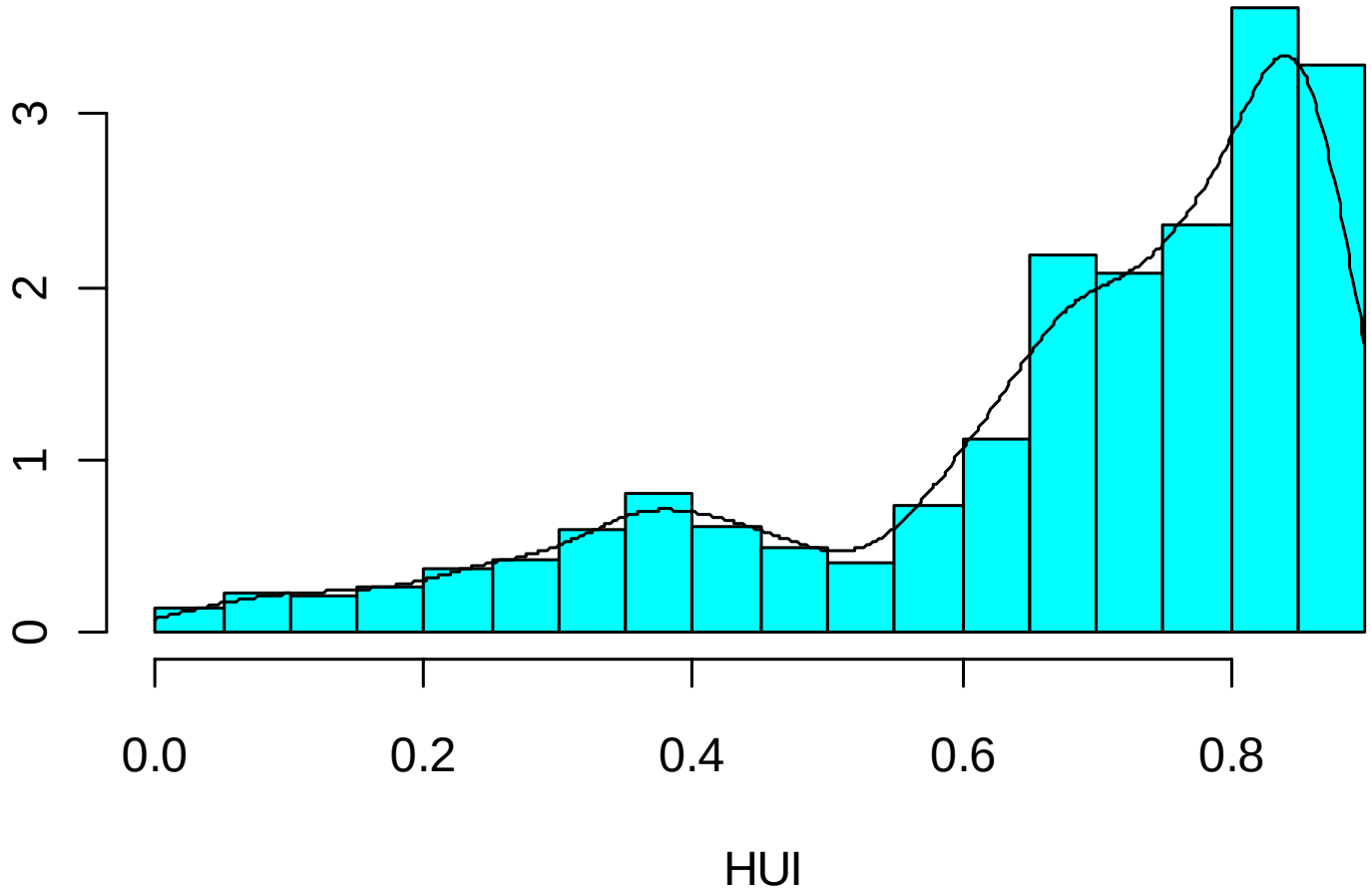


Bell Shaped

--- From Math 2233, Statistics for Life Sciences 1, Course Notes (Cabilio and Masaro)

Histogram and Density Plot

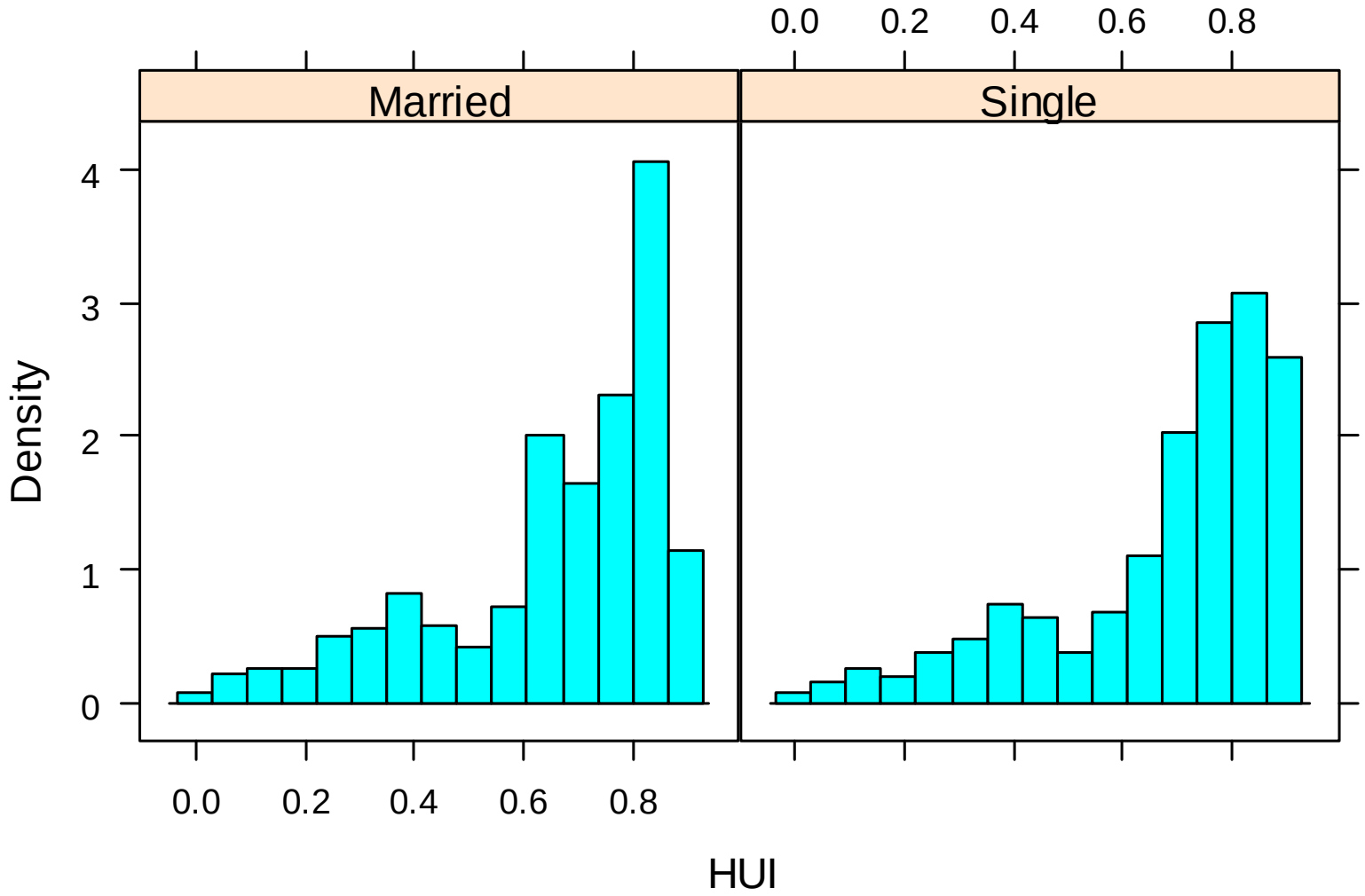
Female



- Not normal
- Bi-mode
- Left skewed

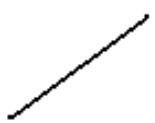
Histogram and Density Plot

Female



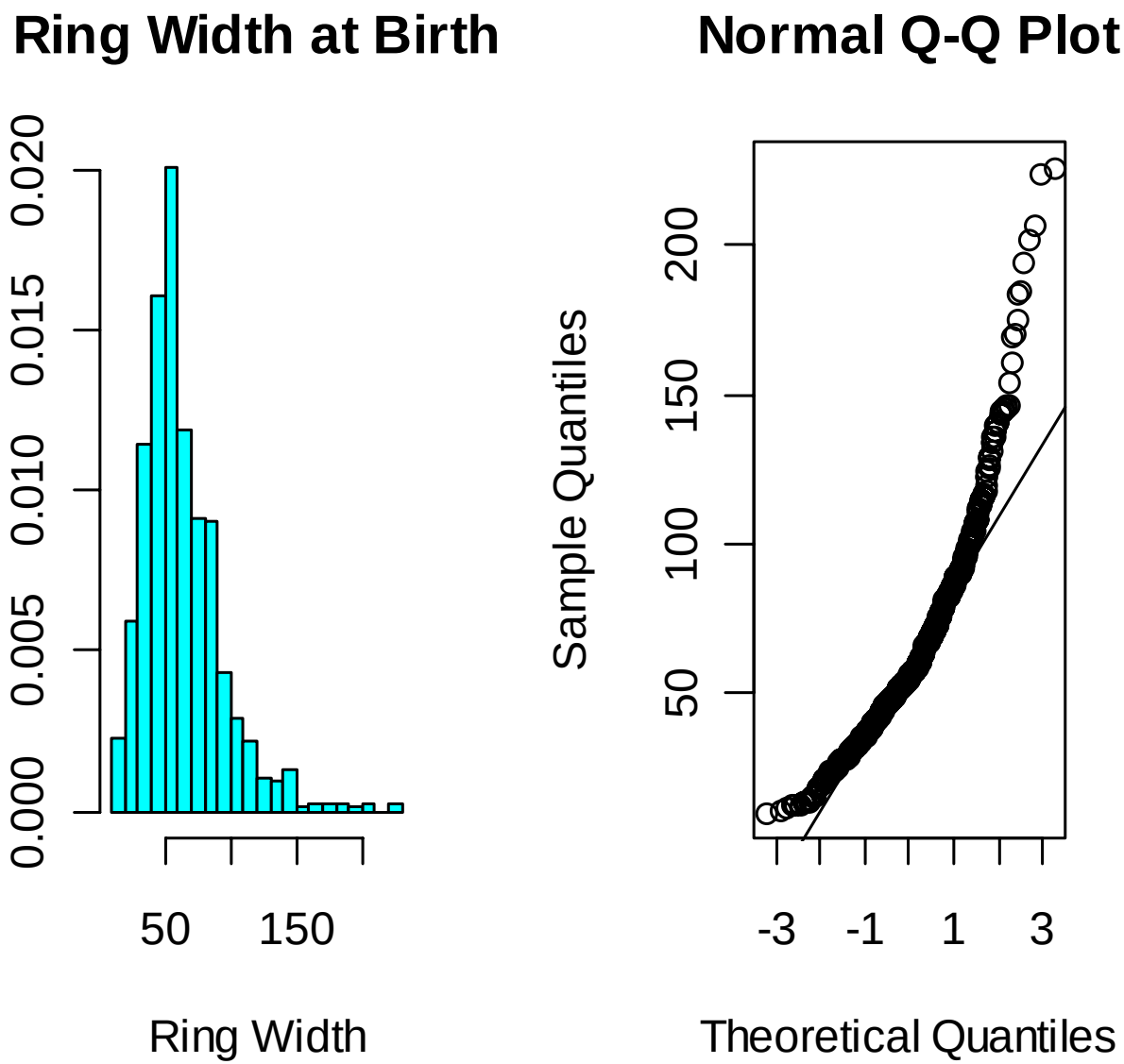
Probability Plot

- The normal distribution scores vs the sort sample
- Prototype probability plots

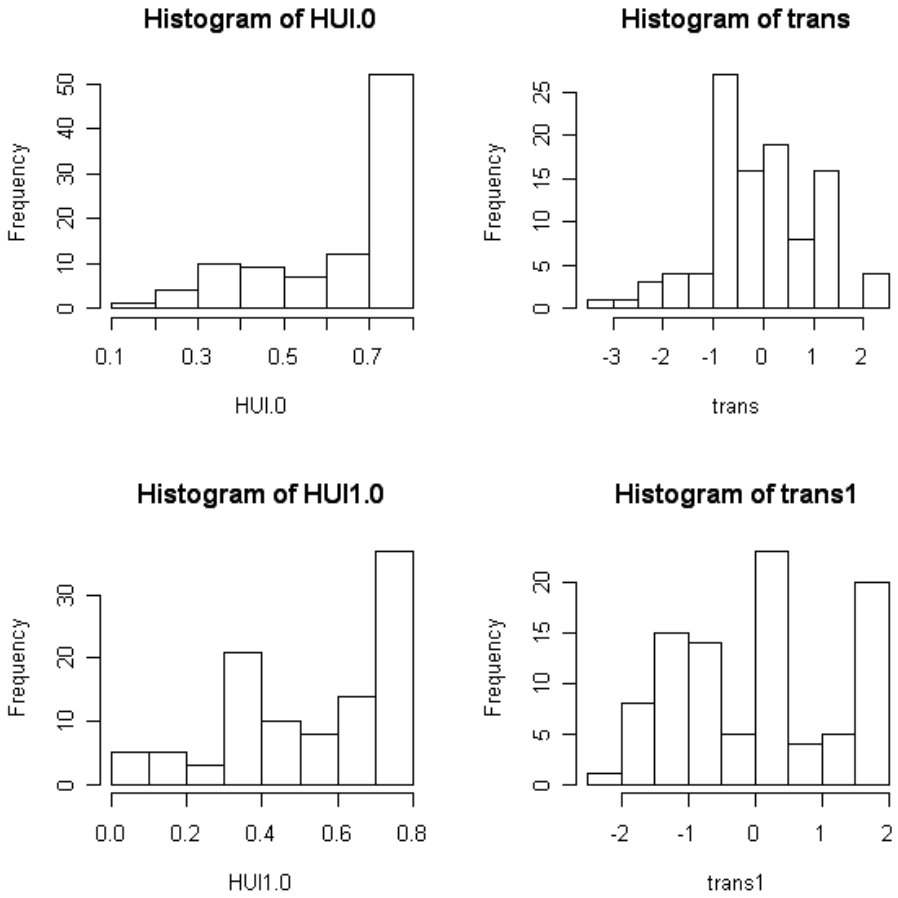
Normal	Left Skewed	Right Skewed	Light Tailed	Heavy Tailed
				

--- From MATH 2243, Statistics for Life Sciences 2, Course Notes (Cabilio and Masaro)

Histogram and Probability Plot

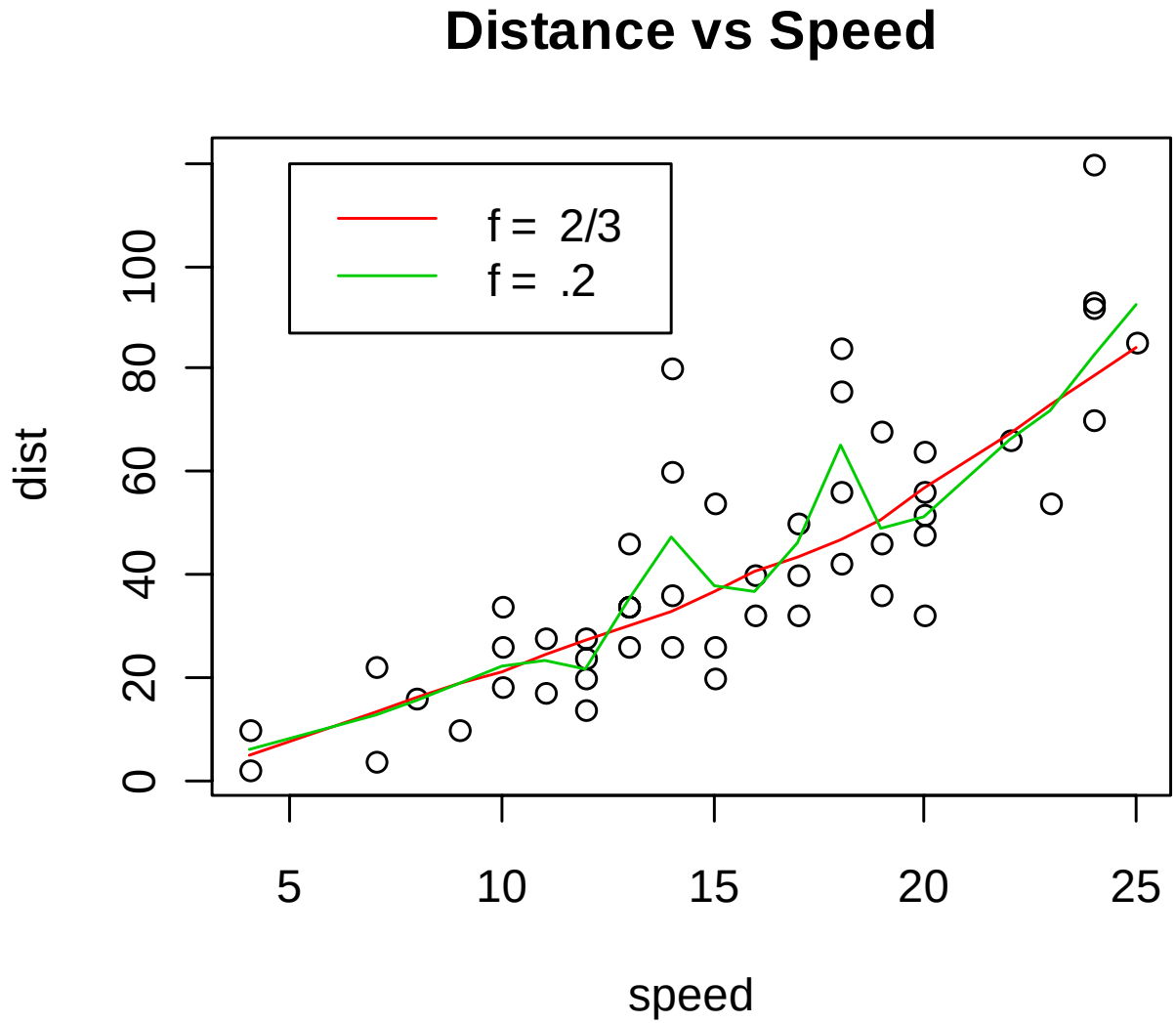


Histogram and Data Transformation

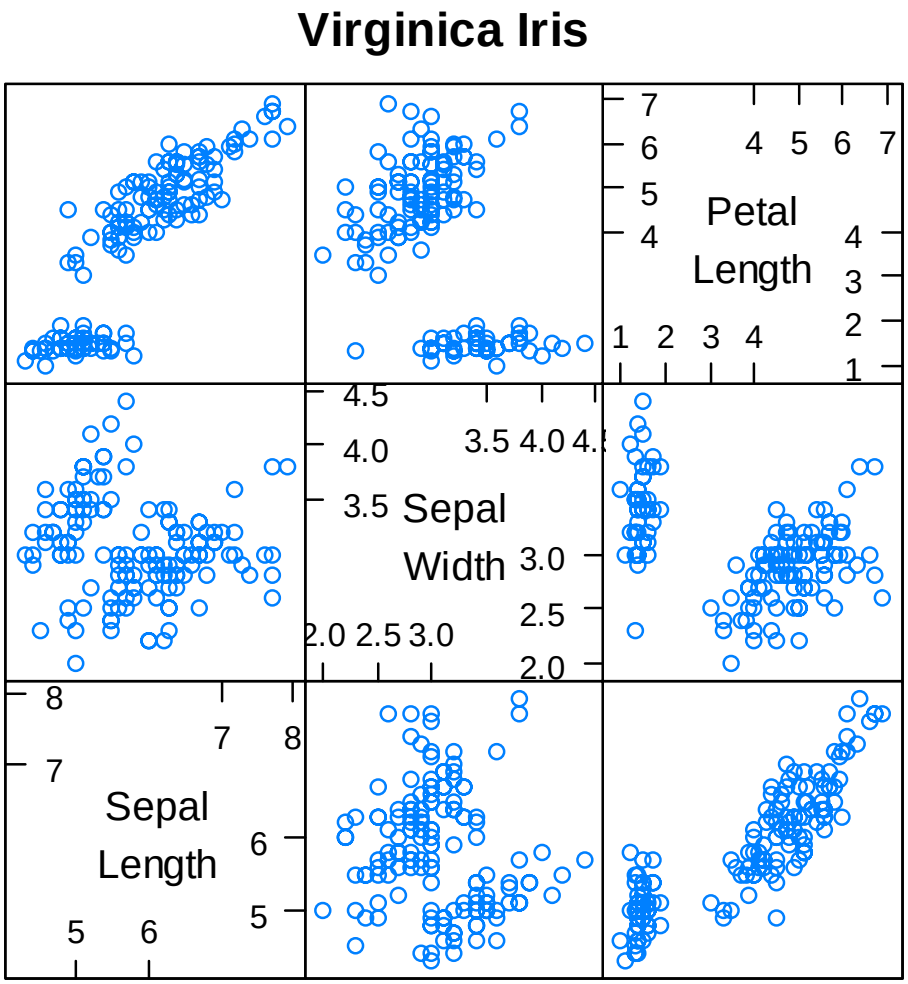


Learning the Relationship from the Bivariate Data

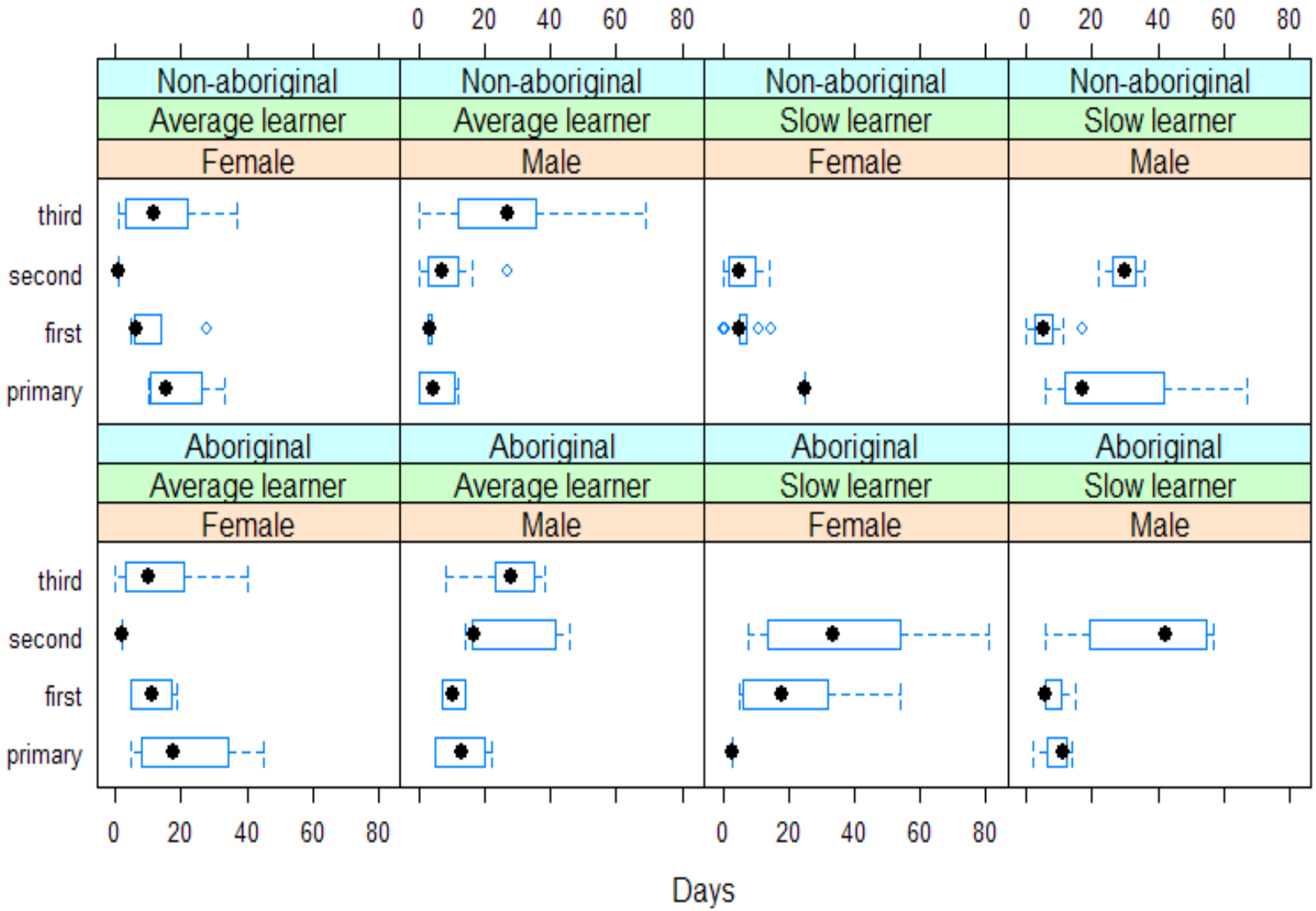
Scatter Plot Smoothing



Scatterplot Matrix



Learning the Interaction Effect from the Mutiway Data


Non-aboriginal
Average learner
Female
Non-aboriginal
Average learner
Male
Non-aboriginal
Slow learner
Female
Non-aboriginal
Slow learner
Male
Aboriginal
Average learner
Female
Aboriginal
Average learner
Male
Aboriginal
Slow learner
Female
Aboriginal
Slow learner
Male
third
second
first
primary
0
20
40
60
80
Days

References

- **J. Bu** (2006). Statistical Modeling for Nursing Human Health Resources. *A **t**echnical report for MITACS graduate student internship grant*.

- **P. Cabilio and J. Masaro**. *Statistics for Life Sciences 1 & 2, Course Notes*. Department of Mathematics and Statistics, Acadia University

- **R Development Core Team** (2007). *R: A Language and Environment for Statistical Computing*. Note that most of graphs in this presentation were produced in R. URL http: //www.R-project.org

- **W. S. Cleveland** (1993). *Visualizing Data.* AT & T Laboratories, Murray Hill, New Jersey.

- **Y. C. Huang, Y. Zhang, P. Cabilio, M. Richard, and T. Herman** (2009) . Analysis of the Growth of the Nova Scotia Blanding's Turtle. *Atlantic Electronic Journal of Mathematics.* To appear.