Descriptive Statistics and Exploratory Data Analysis

Dean’s Faculty and Resident Development Series

UT College of Medicine Chattanooga

Probasco Auditorium at Erlanger

January 14, 2008

Marc Loizeaux, PhD

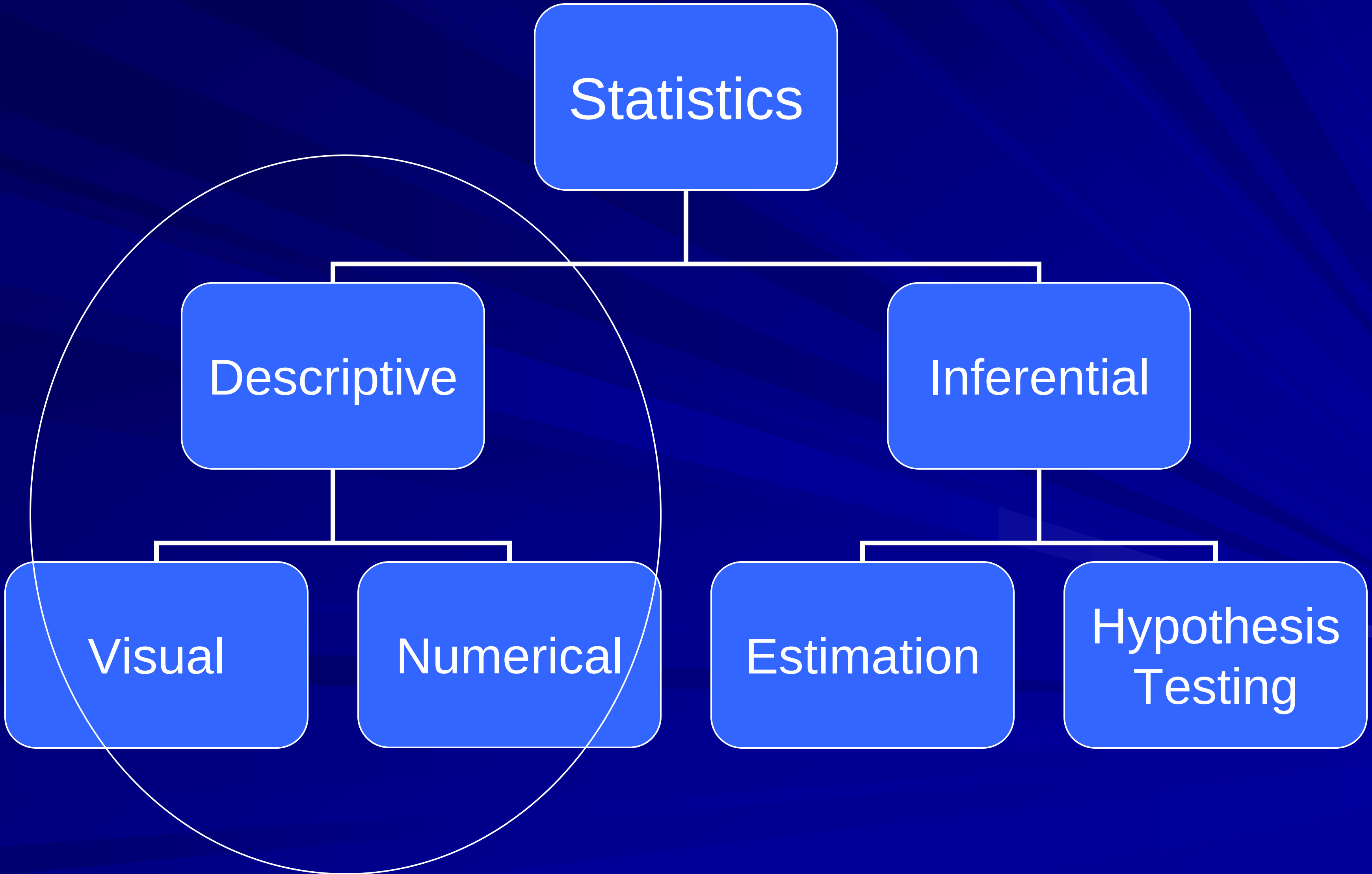
Department of Mathematics

University of Tennessee at Chattanooga

What is descriptive statistics?

- Descriptive statistics <u>describes</u> your data.
    - Visual and Numerical
- Inferential statistics <u>draws inferences</u> about a larger population.
    - Estimation and hypothesis testing

The Big Picture



Statistics

Descriptive

Inferential

Visual

Numerical

Estimation

Hypothesis Testing

Why descriptive statistics?

- To summarize our data
- To help us get to know our data
- To help us describe our data to an audience
- To help us explore our data.

What is Exploratory Data Analysis?

"Exploratory data analysis is detective work

– numerical detective work

– or counting detective work

– or graphical detective work"

- John Wilder Tukey,

Exploratory Data Analysis, page 1

Exploring our data

- Gives us an overall view
- Helps us consider basic assumptions
- Helps us spot oddball values
- Helps us avoid embarrassing oversights
- May help us decide on the next step

Visual Descriptions
(Tools for exploring your data visually)

- Charts and Graphs
  - – Histogram
  - – Dotplot
  - – Stem and leaf plot
  - – Boxplot
  - – Scatterplot
  - – And many more

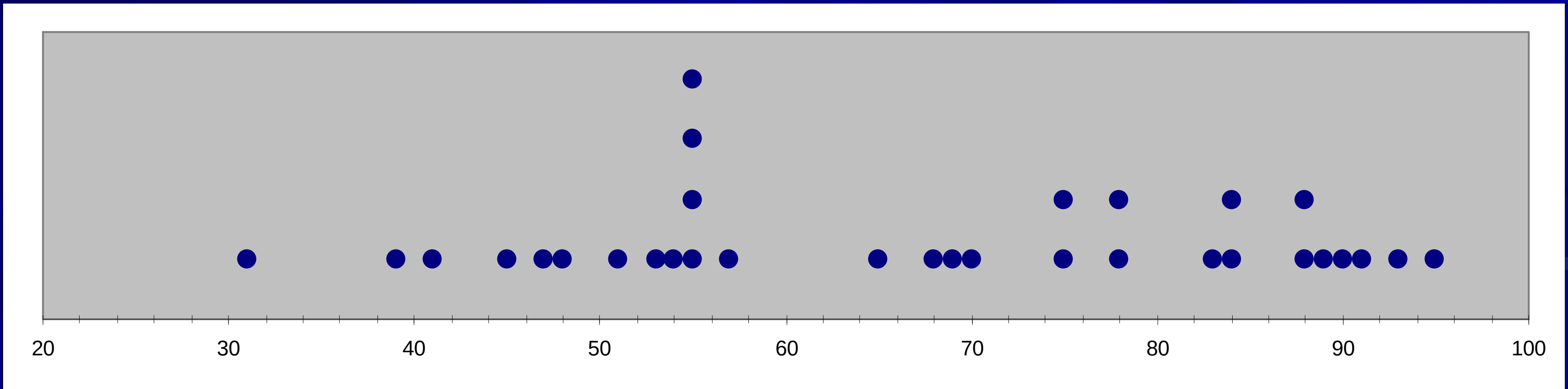
A simple example

Grades on the first exam

84 75 83 48 70 31 39 51 57 68 55

84 89 45 53 55 69 93 54 65 75 78

88 90 91 95 88 55 55 41 47 78



Numerical Descriptions

- (Univariate, interval data)
- We want to describe….
  - The central tendency of the data
    - What is a center point for the data?
    - What is a typical score?
  - The variation of the data?
    - How much spread is there to the data?
    - How far apart are the data values from each other?

Measures of Central Tendency

- The mean is the arithmetic average.
  - Easy to calculate, easy to understand
  - The balance point of the data

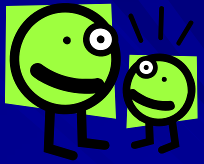
- The median is the score in the middle.
  - Resistant to extreme scores

Measures of Dispersion

- The range.
  - – Easy to calculate and quick

    Range = high score – low score

  - – Limited – only considers two scores



- The standard deviation.
  - – More complicated, but…
  - – Indicates a “typical” deviation from the mean

Childhood Respiratory Disease

(playing with the data)

Data available from OzDASL, StatSci.org

- FEV (forced expiratory volume) is an index of pulmonary function that measures the volume of air expelled after one second of constant effort.
- The data: determinations of FEV on 654 children ages 6-22 who were seen in the Childhood Respiratory Desease Study in 1980 in East Boston, Massachusetts. The data are part of a larger study to follow the change in pulmonary function over time in children.
- Source:
  - Tager, I. B., Weiss, S. T., Rosner, B., and Speizer, F. E. (1979). Effect of parental cigarette smoking on pulmonary function in children. *American Journal of Epidemiology*, **110**, 15-26.
  - Rosner, B. (1990). *Fundamentals of Biostatistics, 3rd Edition*. PWS-Kent, Boston, Massachusetts.

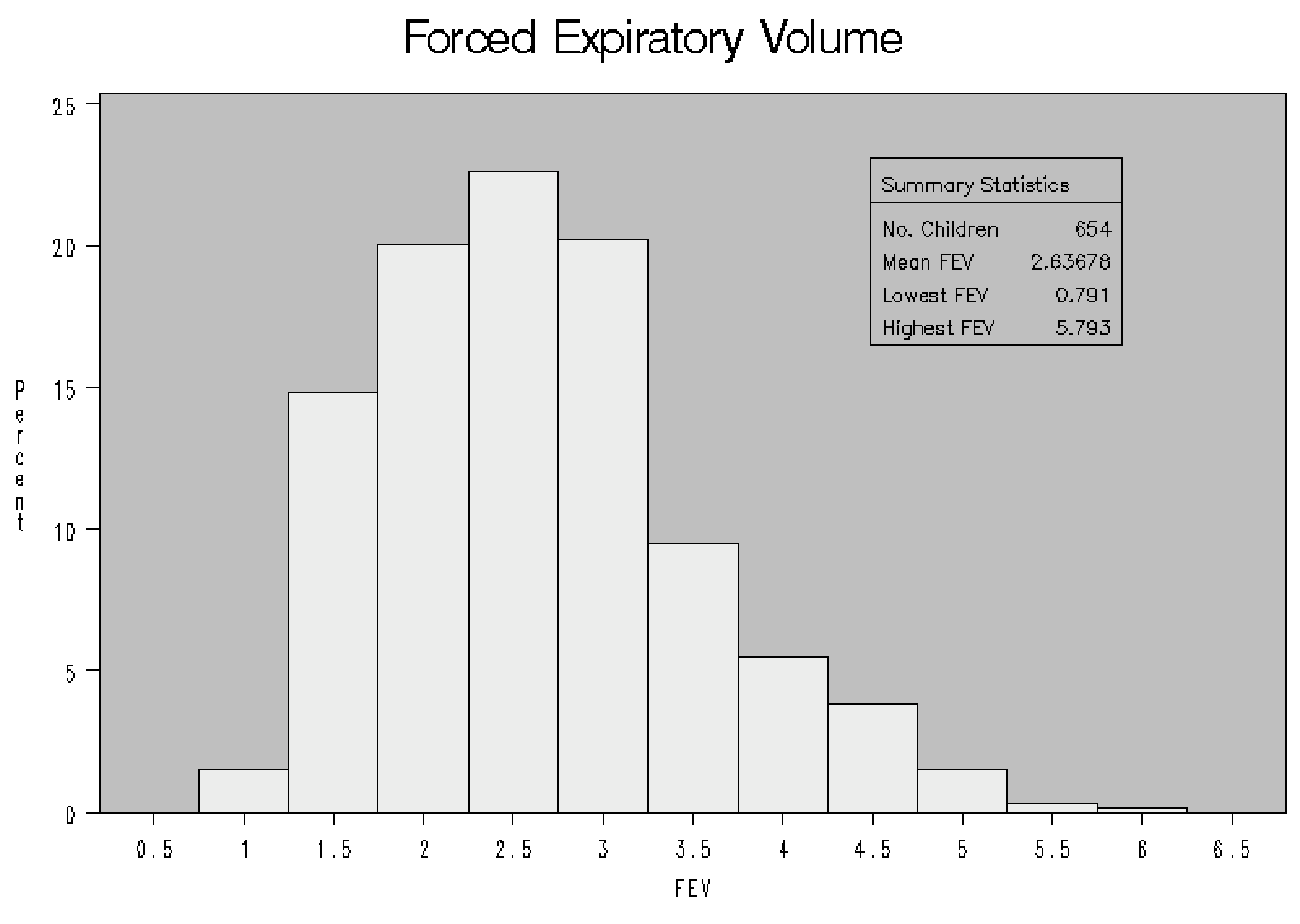
Some of the Data

ID	Age	FEV	Height	Sex	Smoker
46951	12	3.082	63.5	Female	Non
47051	13	3.297	65	Female	Current
47052	11	3.258	63	Female	Non
72901	12	2.935	65.5	Male	Non
73041	16	4.27	67	Male	Current
73042	15	3.727	68	Male	Current
73751	18	2.853	60	Female	Non
75852	16	2.795	63	Female	Current
77151	15	3.211	66.5	Female	Non

Descriptive Statistics

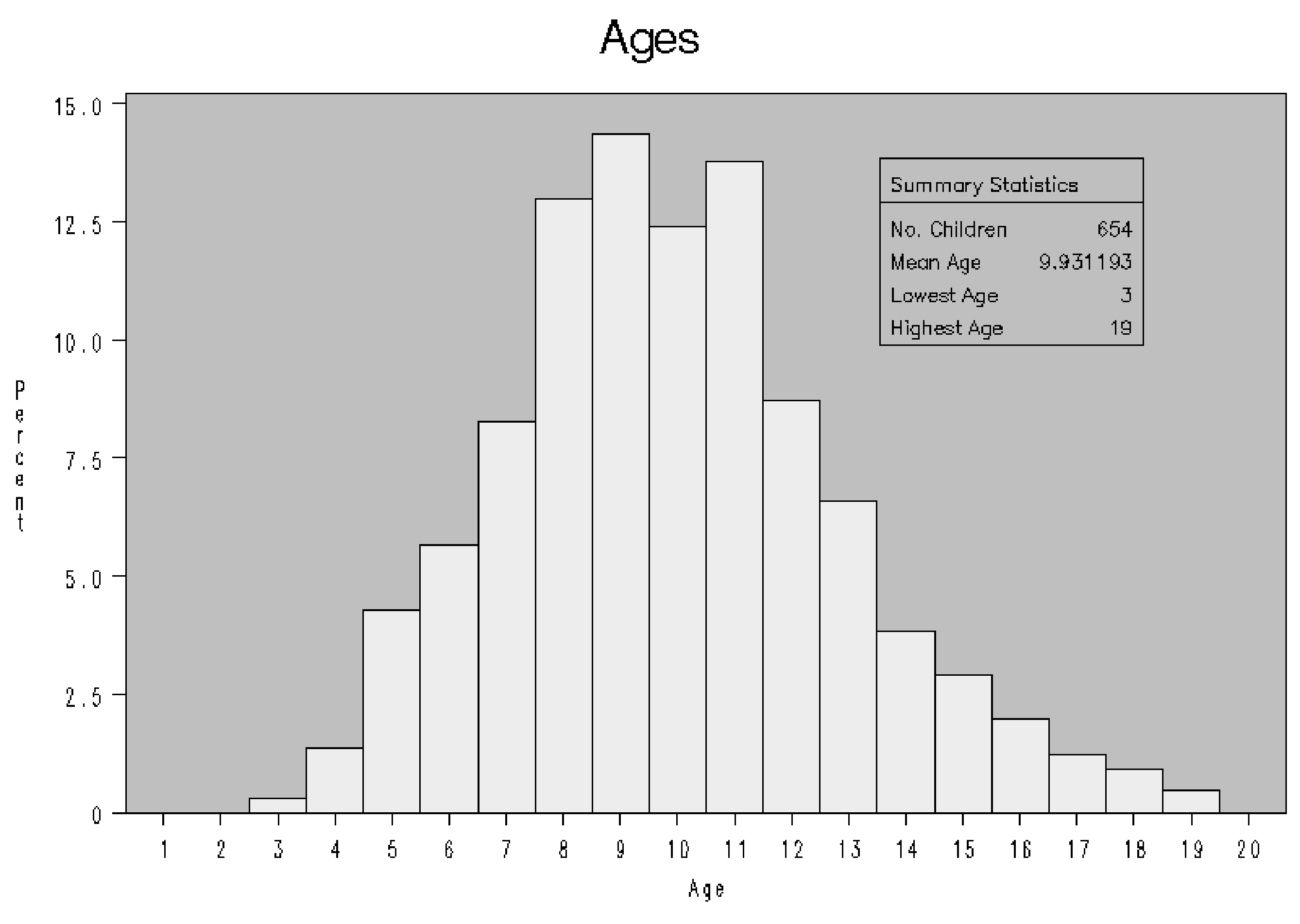
	Age	*FEV*	*Height*
Mean	9.93	2.64	61.14
Median	10.00	2.55	61.50
Mode	9	3.08	63
Standard Deviation	2.95	0.87	5.70
Range	16	5.00	28
Minimum	3	0.79	46
Maximum	19	5.79	74

Pictures may say more



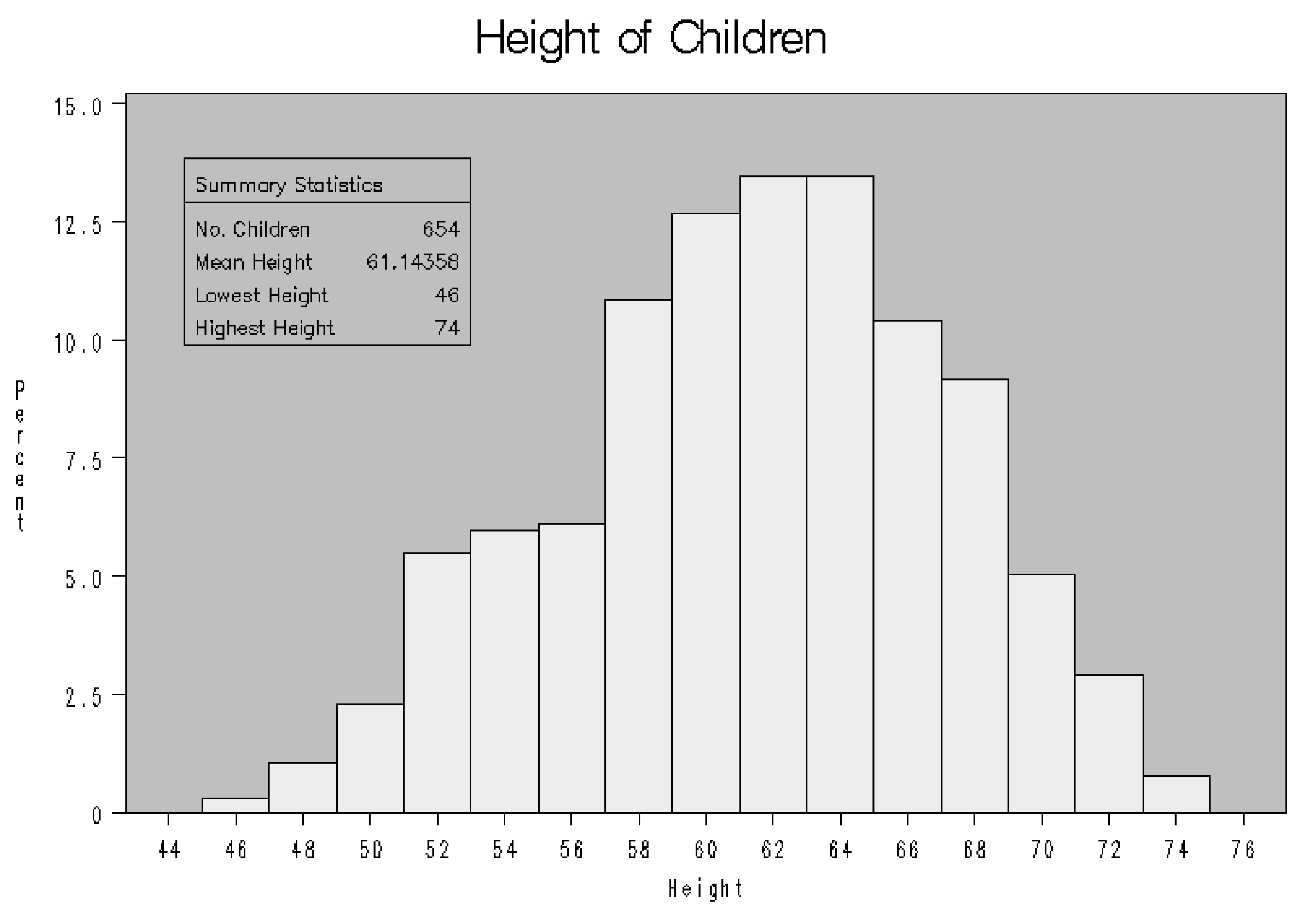
Summary Statistics	
No. Children	654
Mean FEV	2.63678
Lowest FEV	0.791
Highest FEV	5.793

The ages look like this



Summary Statistics	
No. Children	654
Mean Age	9.931193
Lowest Age	3
Highest Age	19

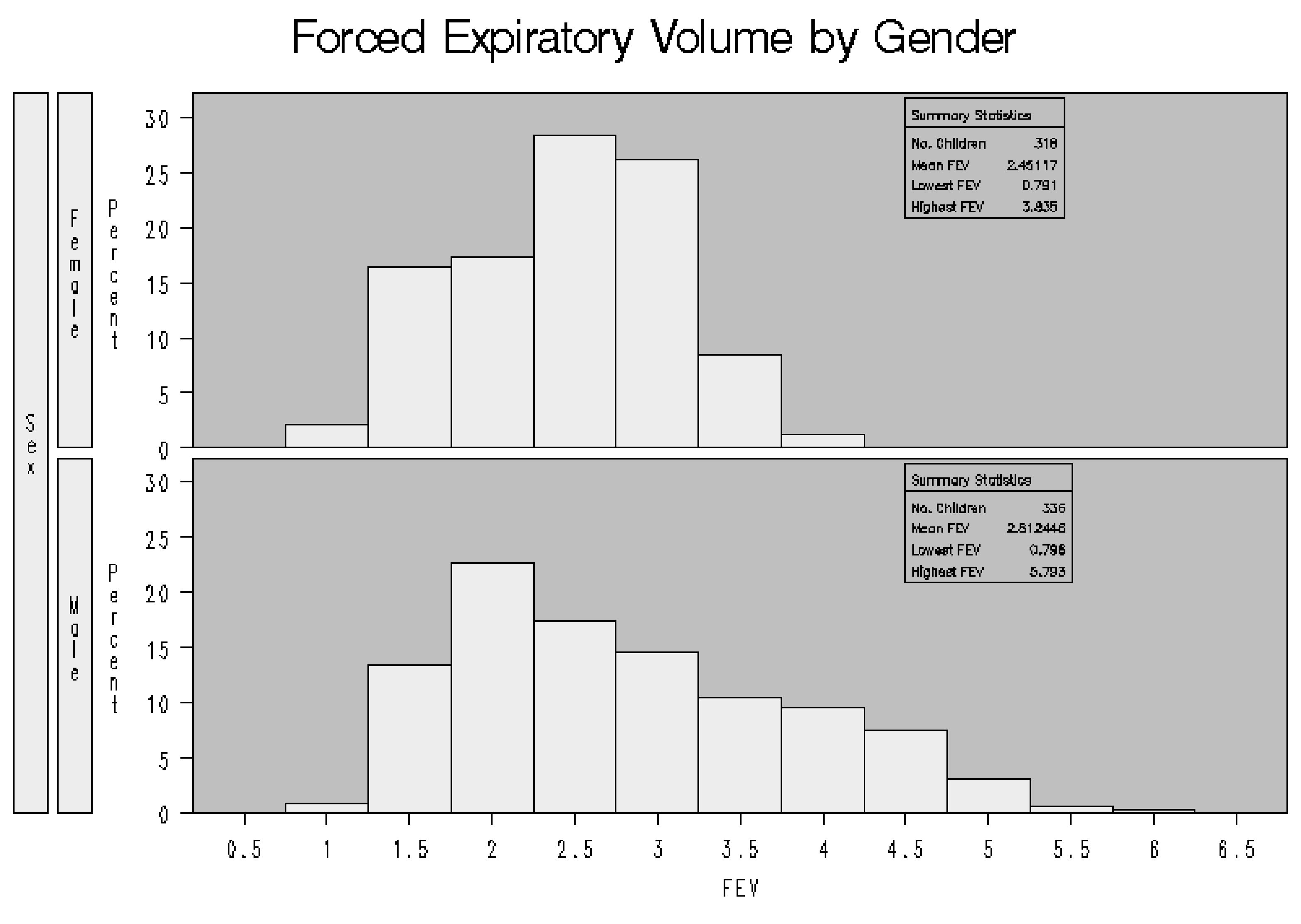
And again


Height of Children
Summary Statistics
No. Children 654
Mean Height 61.14358
Lowest Height 46
Highest Height 74
15.0
12.5
10.0
7.5
5.0
2.5
0
Percent
44
46
48
50
52
54
56
58
60
62
64
66
68
70
72
74
76
Height

One variable, then two…

- A univariate exploration
  - – Explore each data column individually
- A multivariate exploration
  - – Explore the relationships between two data columns

Consider natural subgroups

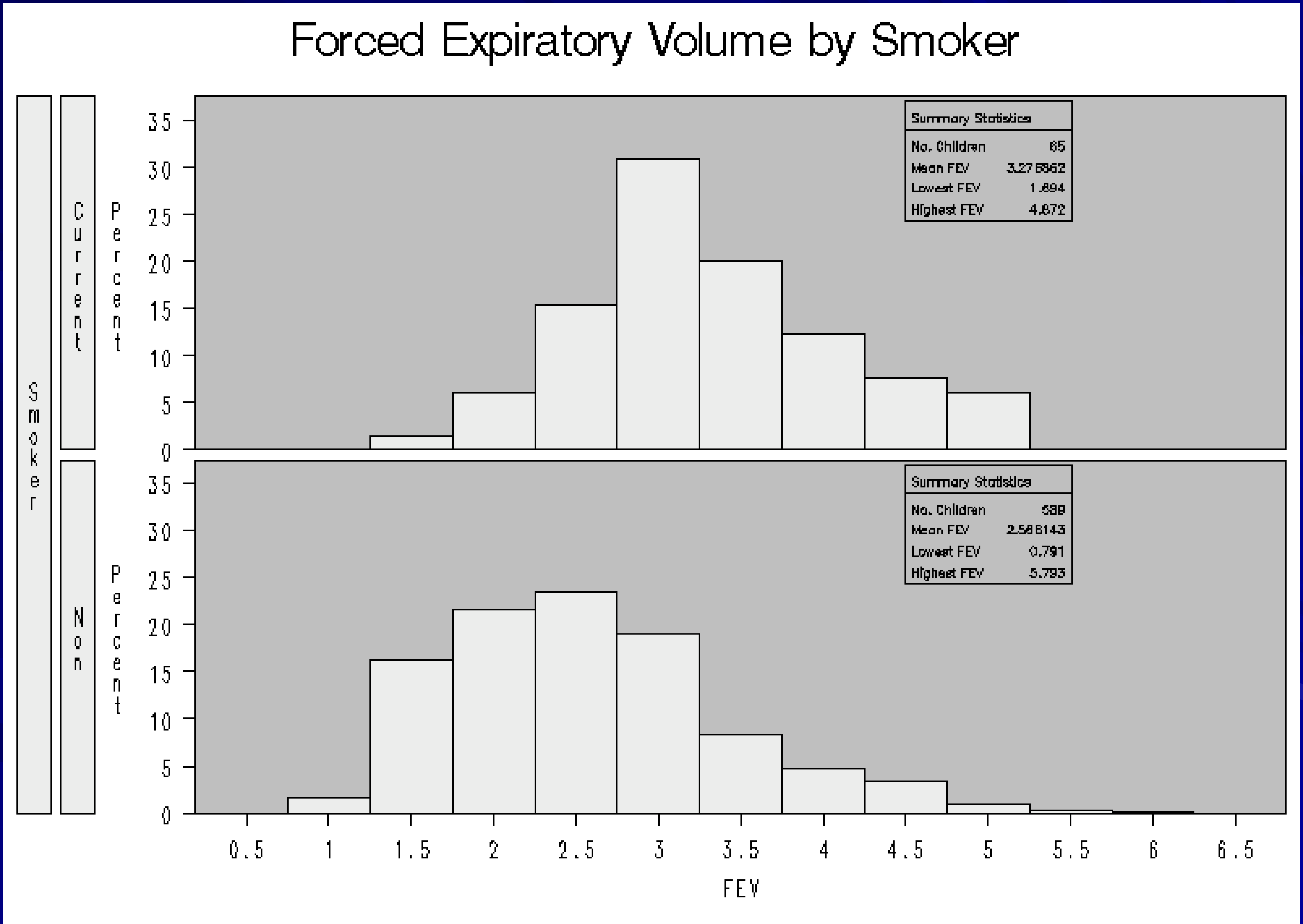


Forced Expiratory Volume by Gender

Summary Statistics	
No. Children	318
Mean FEV	2.45117
Lowest FEV	0.791
Highest FEV	3.835

Summary Statistics	
No. Children	336
Mean FEV	2.812446
Lowest FEV	0.796
Highest FEV	5.793

Raising more questions?

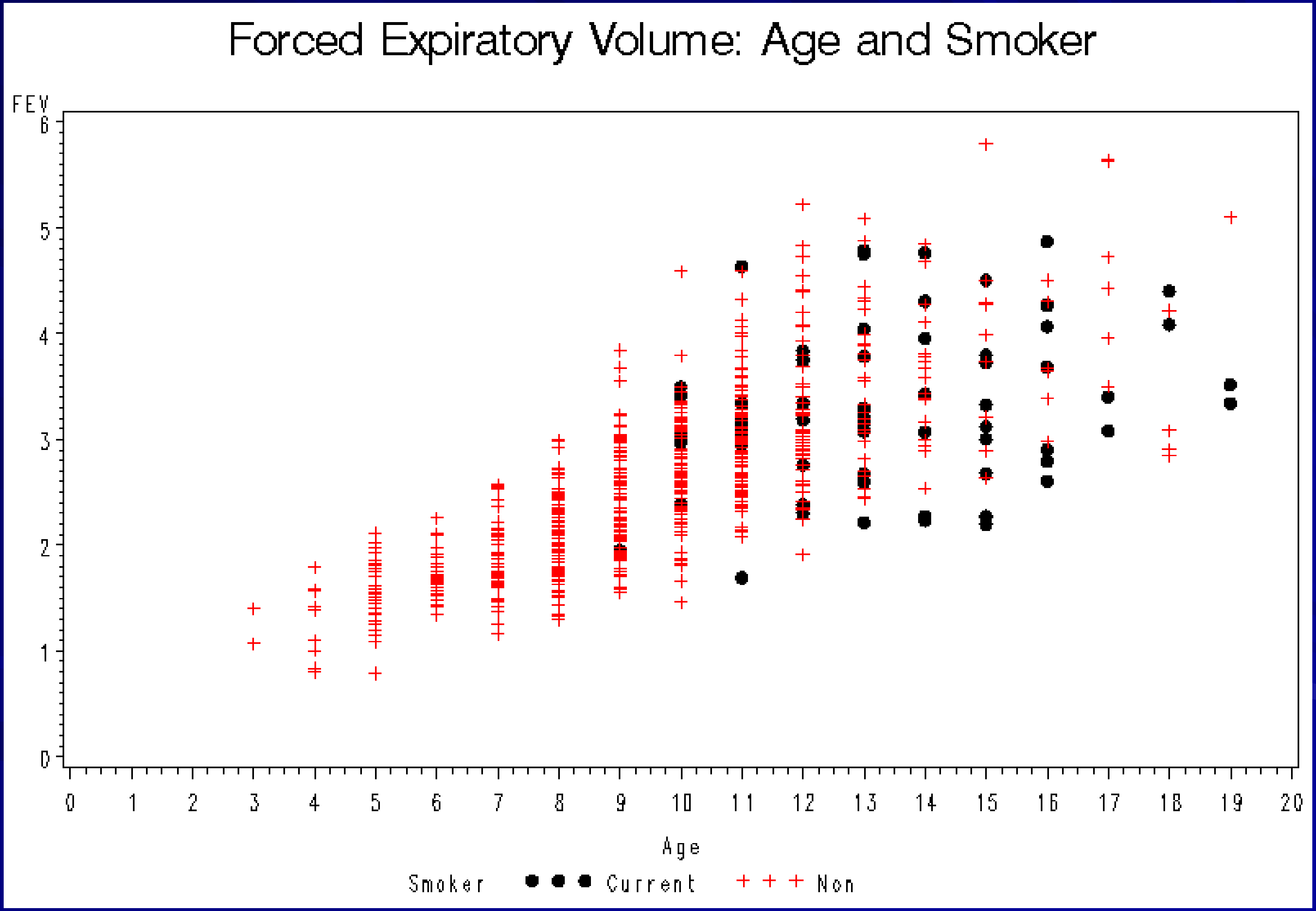


Forced Expiratory Volume by Smoker

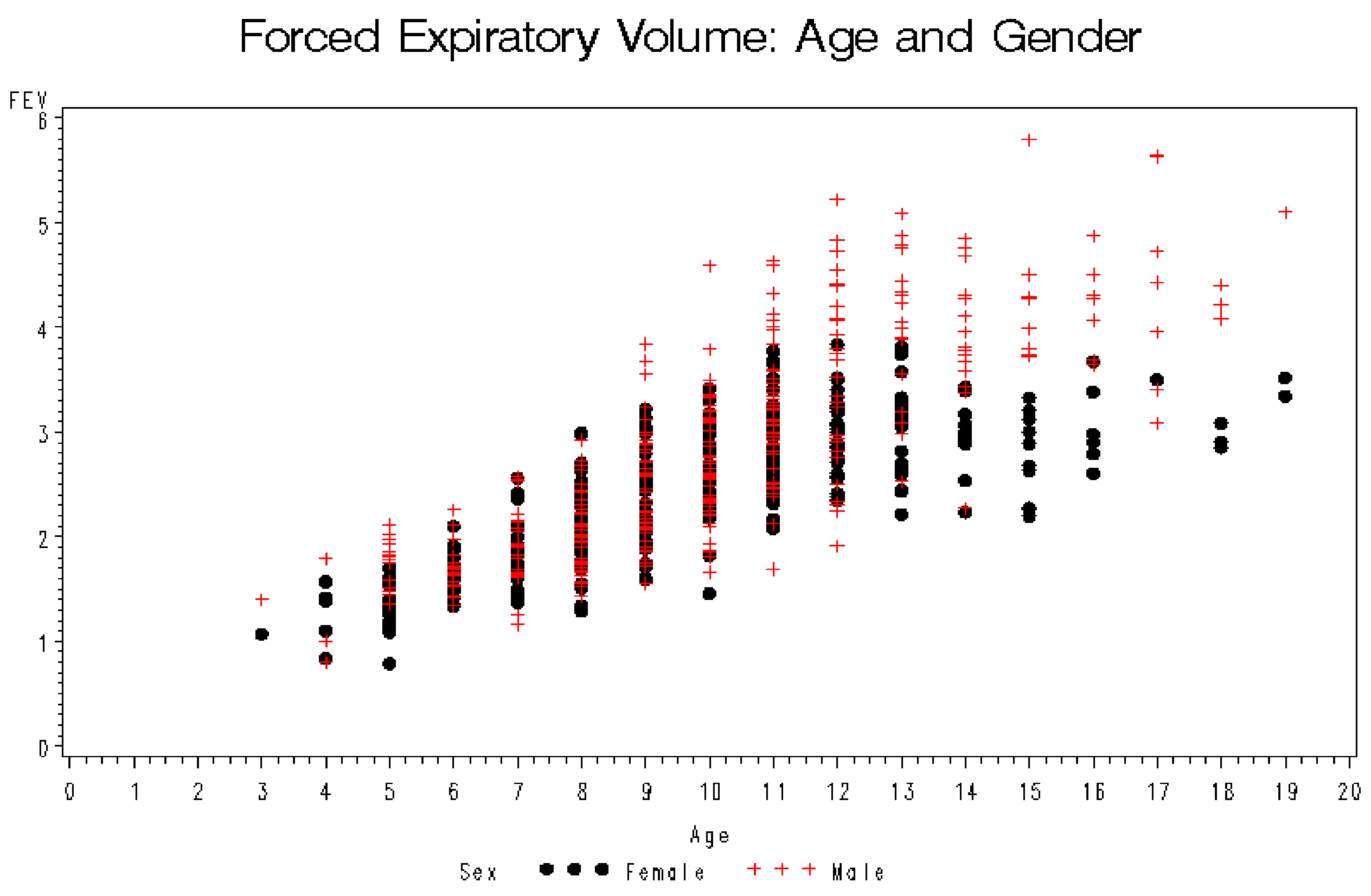
Summary Statistics	
No. Children	65
Mean FEV	3.2768962
Lowest FEV	1.694
Highest FEV	4.872

Summary Statistics	
No. Children	589
Mean FEV	2.566143
Lowest FEV	0.791
Highest FEV	5.793

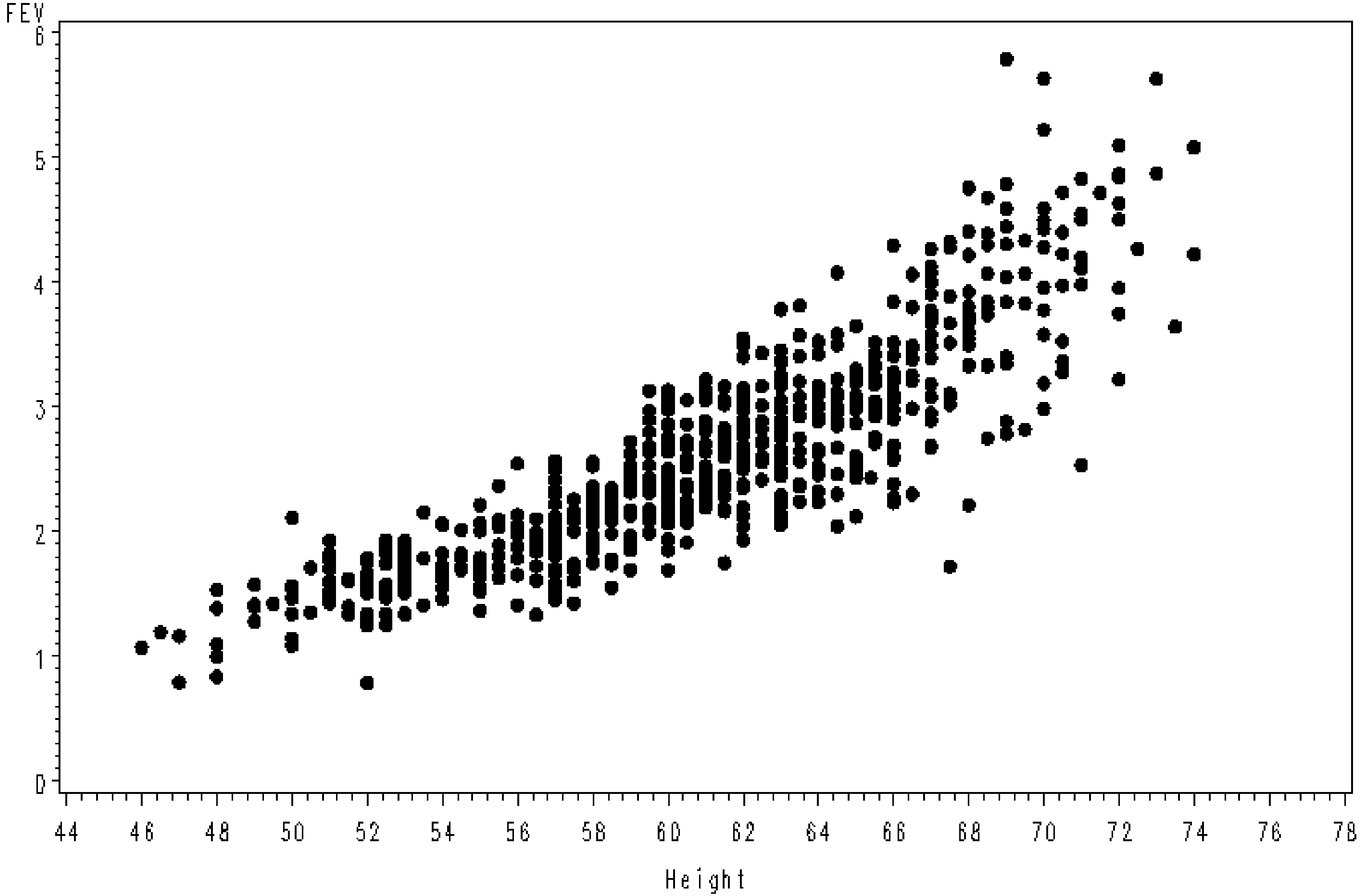
It starts to make sense



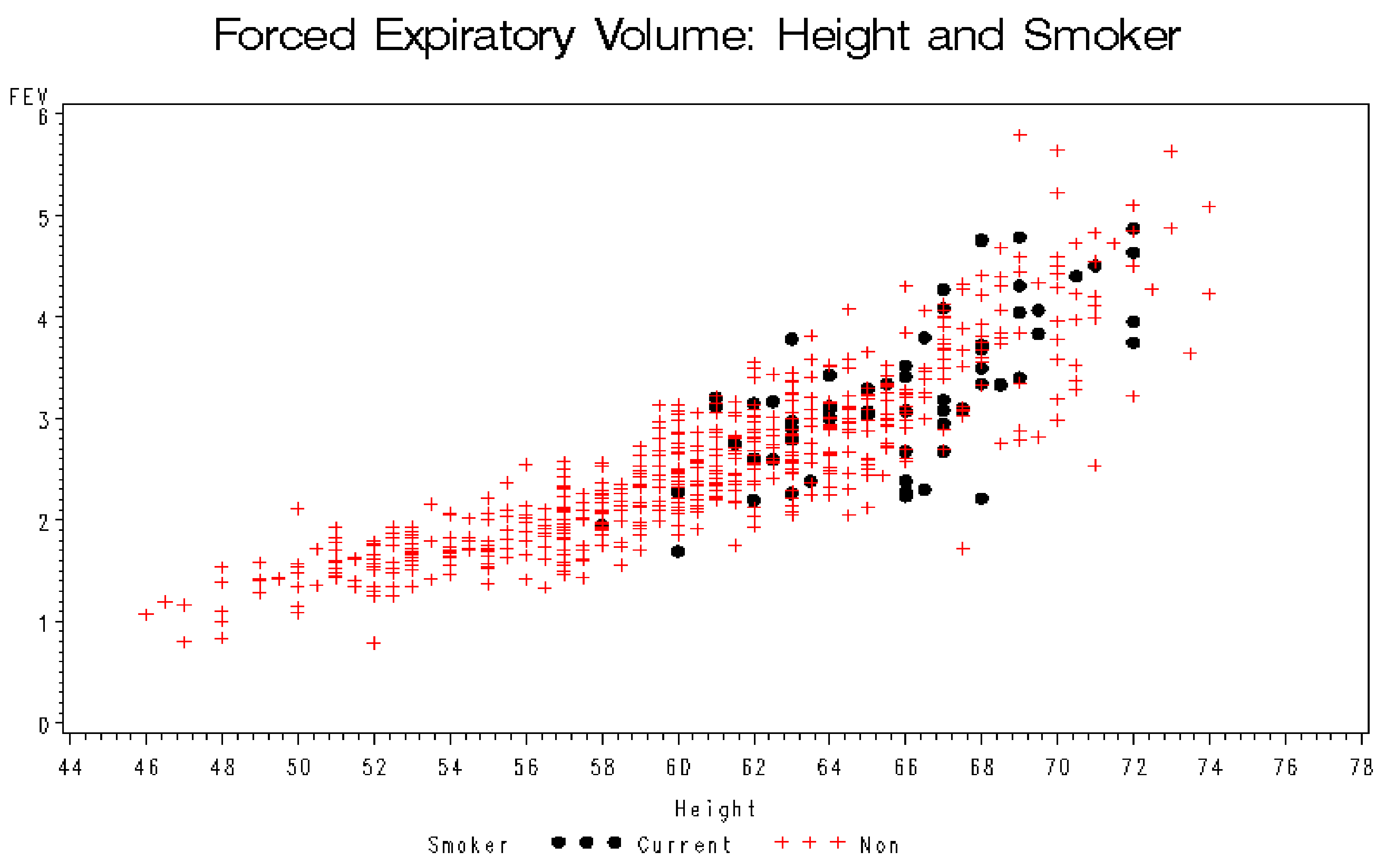
Something else to study?



Forced Expiratory Volume by Height



Differentiating Subgroups



Preparing for an Audience

- Some Do’s
  - Pick and choose your graphs
  - Include appropriate numbers for your type of data
  - Include narrative
    - Does the histogram indicate asymmetry?
    - Are there unexpected values in the data set?
    - Are there special problems you had to deal with to describe the data?

Preparing for an Audience (2)

- Some Don’ts
  - Don’t include everything – that just confuses us.
  - Don’t be redundant – some graphs say the same thing.
  - Don’t include descriptors you don’t understand (kurtosis?) – ask the chauffeur

Points to Remember

(in no particular order)

- Don’t skip the simple stuff!
- Spend time playing with your data.
- Pictures say a lot.
- Describe the spread as well as the center.
- Consider the natural subgroups in your data.

Next Time

- Confidence Intervals,
  Hypothesis Tests,
  and Statistical Significance
- 2 x 2 tables

Monday, February 11