Data, Data Science and Exploratory Data Analysis

Siba K Udgata

School of Computer and Information Sciences,
University of Hyderabad, Hyderabad

What is collecting all this data?

Web Browsers **Search Engines**

Microsoft's
Internet Explorer

Mozilla's FireFox
(Non-profit foundation,
used to be Netscape)

Google's Chrome

AOL

Google's

Google

Microsoft's

bing

Yahoo's

YAHOO!

Apple's Safari

IAC Search's

Ask .com

Time-
Warner's
AOL
Explorer

What is collecting all this data?

Smartphones & Apps

Apple's iPhone
(Apple O/S)

Samsung, HTC.
Nokia, Motorola
(Android O/S)

RIM Corp's Blackberry
(BlackBerry O/S)

Tablet Computers & Apps

Apple's iPad

Samsung's Galaxy

Amazon's Kindle Fire

What is collecting all this data?

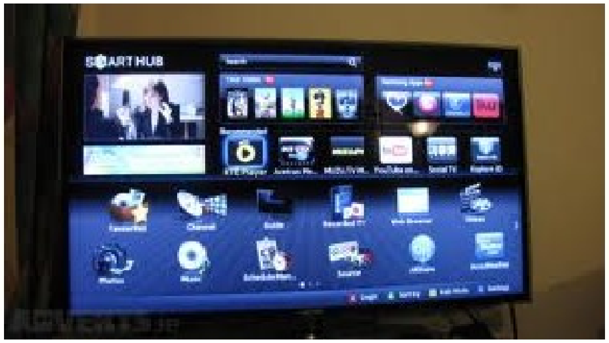
Games Boxes and GPS Systems

Internet Service Providers

What is collecting all this data?

HDTV’s and Blu-Ray Players with built-in Internet connectivity

Movie Rental Sites



What is collecting all this data?

Hospitals & Other Medical Systems

Pharmacies

Laboratories

Imaging Centers

Emergency Medical Services (EMS)

Hospital Information Systems

Doc-in-a-Box

Electronic Medical Records

Blood Banks

Birth & Death Records

Banking & Phone Systems

Branches
ATMs
Call centers
Internet
Relationship managers/ agents

verizon

Can you hear me now?
(Heh heh heh!)

Sprint

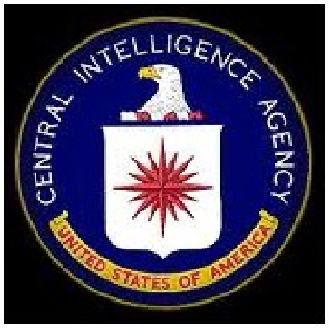
T · · Mobile ·

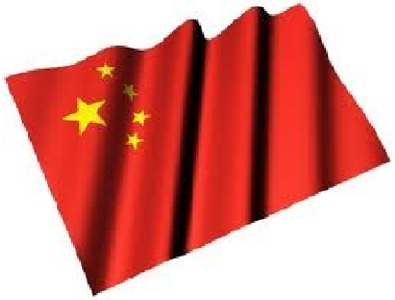
at&t

Who is collecting all of this data?

Government Agencies

IRS
Department of the Treasury
Internal Revenue Service





Big Pharmaceutical Companies

BAYER
BAYER

AstraZeneca

gsk
GlaxoSmithKline

Who is collecting all this data?

Consumer Products Companies

P&G

KRAFT

Big Box Stores

Sam's CLUB

THE HOME DEPOT

COSTCO WHOLESALE

STAPLES

WAL★MART
ALWAYS LOW PRICES.
Always.

Who is collecting what?

Credit Card Companies

MasterCard

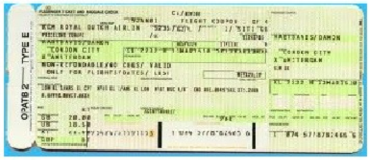
VISA

AMERICAN EXPRESS

DISCOVER

What data are they getting?

Airline ticket



OPATB 2 — TYPE E

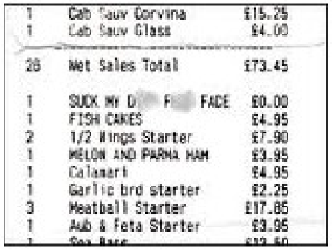
KLM ROYAL DUTCH AIRLINES

LONDON CITY

AMSTERDAM

NON-REFUNDABLE/NO CHGS/ VALID ONLY FOR FLIGHTS/DATES/ LAST

Restaurant check



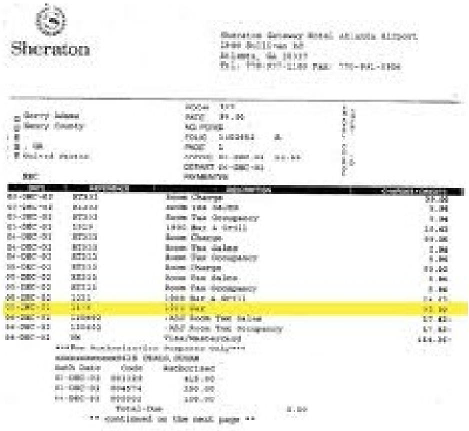
1	Cab Sauv Corvina	£15.25
1	Cab Sauv Glass	£4.00
28	Net Sales Total	£73.45
1	SUCK MY D[redacted] F[redacted] FACE	£0.00
1	FISH CAKES	£4.95
2	1/2 Wings Starter	£7.90
1	MELON AND PARMA HAM	£3.95
1	Calamari	£4.95
1	Garlic brd starter	£2.25
3	Meatball Starter	£17.85
1	Aub & Feta Starter	£3.95

Grocery Bill

Kroger

Right Store. Right Price.

Hotel Bill



Sheraton

Sheraton Gateway Hotel Atlanta Airport

Room Charge

Room Tax Sales

Room Tax Occupancy

Total-Due

** continued on the next page **

Data Analysis Primer

- Data Analysis is an integral part of research
- NOT throwing data to tools and reporting fancy graphs and numbers
- Extracting useful, relevant and meaningful information from observations in a systematic and scientific manner

Example..

Why Data Analytics/ Data Science

- Parameter estimation
  - Parameter estimates (also called coefficients) are the change in the response associated with a one-unit change of the predictor, all other predictors being held constant.
- Model development and forecasting
- Feature extraction and classification
  - **Feature selection** is for filtering irrelevant or redundant **features** from the dataset. The key **difference between feature selection** and **extraction** is that **feature selection** keeps a subset of the original **features** while **feature extraction** creates brand new ones.
- Hypothesis testing (Verification of postulates)
- Fault detection (process monitoring)
- and many more

Difficult to answer WHY NOT?

Exploratory Data Analysis

Exploratory Data Analysis refers to a set of techniques originally developed by John Tukey to display data in such a way that interesting features will become apparent.

Unlike classical methods which usually begin with an assumed model for the data, EDA techniques are used to encourage the data to suggest models that might be appropriate.

Sample Dataset

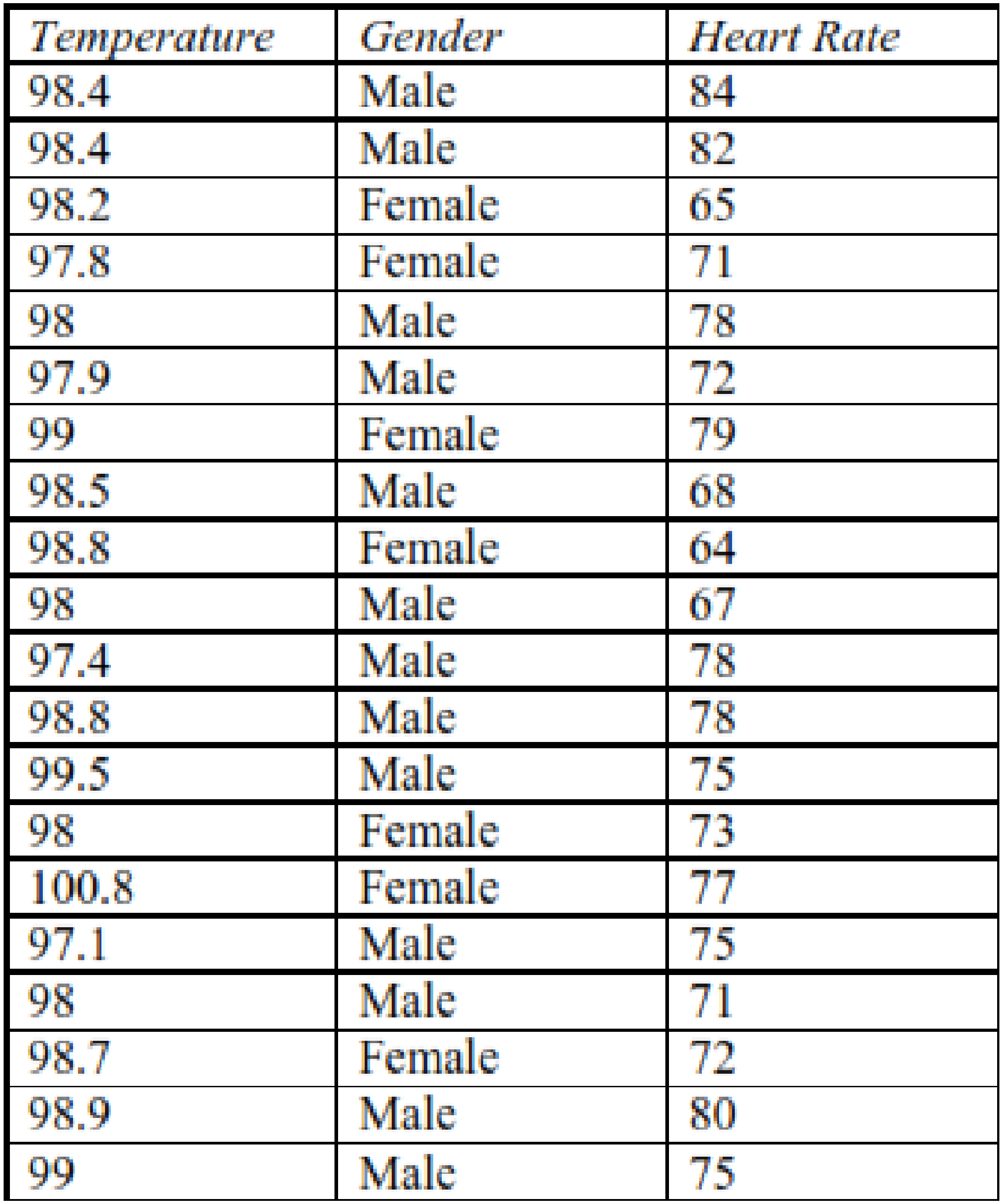
data describing the body temperature of a sample of n = 130 people.

It was obtained from the Journal of Statistical Education Data Archive

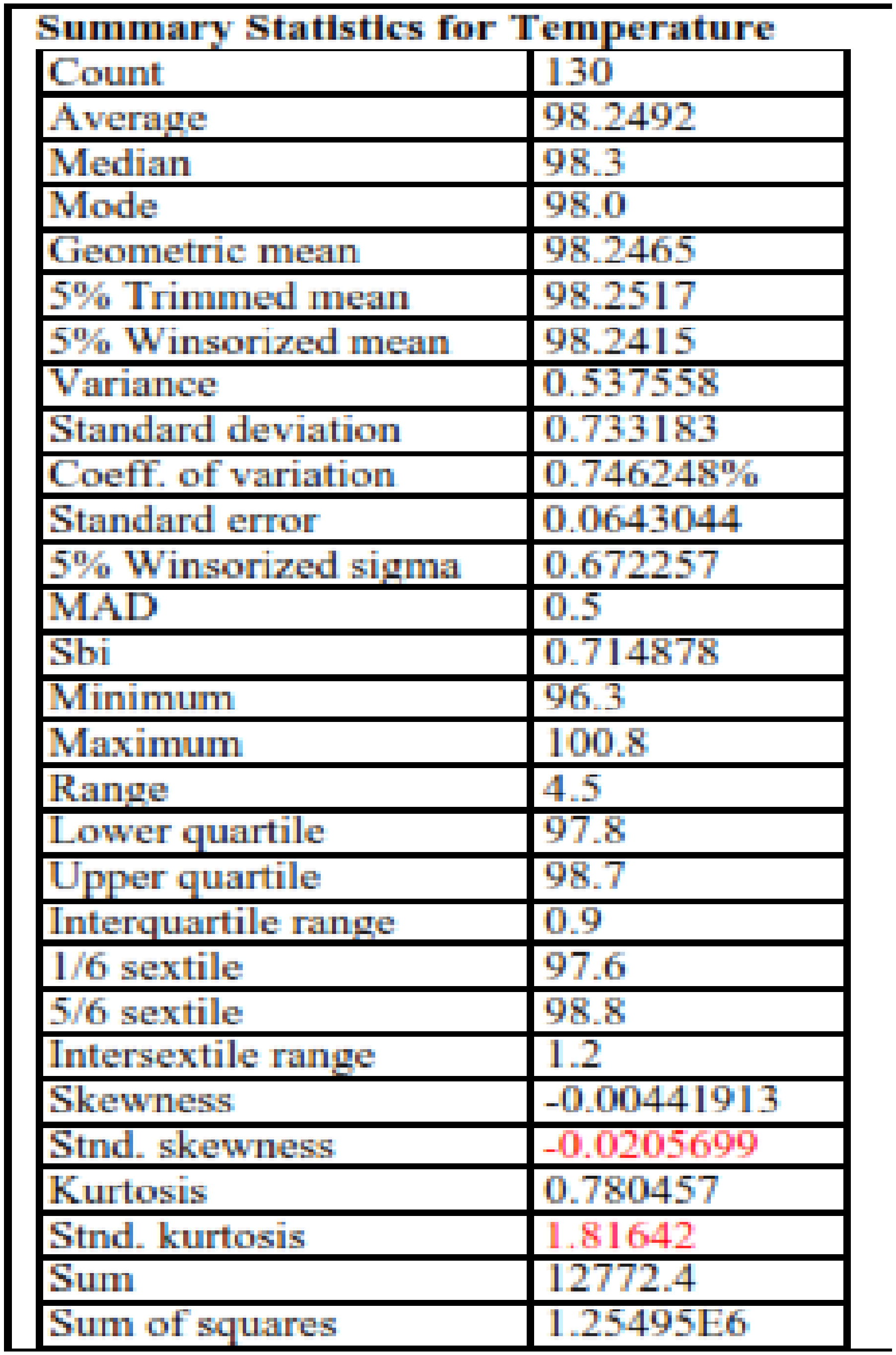
(www.amstat.org/publications/jse/jse_data_archive.html)

and originally appeared in the Journal of the American Medical Association.

The first 20 rows of the file are shown



Temperature	*Gender*	*Heart Rate*
98.4	Male	84
98.4	Male	82
98.2	Female	65
97.8	Female	71
98	Male	78
97.9	Male	72
99	Female	79
98.5	Male	68
98.8	Female	64
98	Male	67
97.4	Male	78
98.8	Male	78
99.5	Male	75
98	Female	73
100.8	Female	77
97.1	Male	75
98	Male	71
98.7	Female	72
98.9	Male	80
99	Male	75



Summary Statistics for Temperature

Count	130
Average	98.2492
Median	98.3
Mode	98.0
Geometric mean	98.2465
5% Trimmed mean	98.2517
5% Winsorized mean	98.2415
Variance	0.537558
Standard deviation	0.733183
Coeff. of variation	0.746248%
Standard error	0.0643044
5% Winsorized sigma	0.672257
MAD	0.5
Sbi	0.714878
Minimum	96.3
Maximum	100.8
Range	4.5
Lower quartile	97.8
Upper quartile	98.7
Interquartile range	0.9
1/6 sextile	97.6
5/6 sextile	98.8
Intersextile range	1.2
Skewness	-0.00441913
Stnd. skewness	-0.0205699
Kurtosis	0.780457
Stnd. kurtosis	1.81642
Sum	12772.4
Sum of squares	1.25495E6

Most of the statistics fall into one of three categories:

1. measures of central tendency – statistics that characterize the “center” of the data.

2. measure of dispersion – statistics that measure the spread of the data.

3. measures of shape – statistics that measure the shape of the data relative to a normal distribution.

α% Trimmed Mean (measure of central tendency) – the mean of the sample after removing a fraction α each of the smallest and largest data values:

$$T(\alpha) = \frac{1}{n(1-2\alpha)}\left[k\left(x_{(r+1)} + x_{(n-r)}\right) + \sum_{i=r+2}^{n-r-1} x_{(i)}\right] \tag{4}$$

where $r = \lfloor \alpha n \rfloor$ and $k = 1 - (\alpha n - r)$. By default, STATGRAPHICS trims 15% from each

Winsorized mean (measure of central tendency) – a resistant measure obtained by calculating the sample mean after copies of $x_{(r+1)}$ and $x_{(n-r)}$ have replaced the data values which would be trimmed away by a trimmed mean:

$$T_W = \frac{1}{n}\left\{\sum_{i=r+1}^{n-r} x_{(i)} + r\left[x_{(r+1)} + x_{(n-r)}\right]\right\} \tag{5}$$

- **Lower quartile** - the 25-th percentile. Approximately 25% of the data values will lie below this value.

- **Upper quartile** - the 75-th percentile. Approximately 75% of the data values will lie below this value.

- **Interquartile range** (measure of dispersion) - the distance between the quartiles:

$$\text{IQR} = \text{upper quartile - lower quartile} \tag{15}$$

- **1/6 sextile** - the 16.67-th percentile.

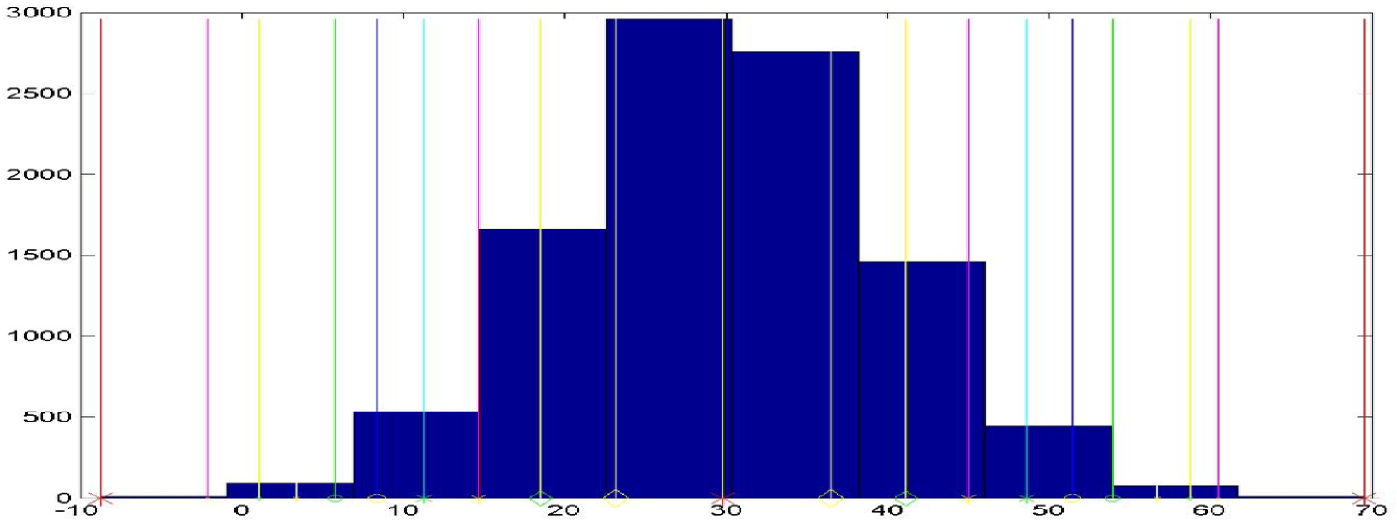
- **5/6 sextile** - the 83.33-th percentile.

- **Intersextile range** (measure of dispersion) - the distance between the sextiles:

$$\text{ISR} = \text{upper sextile - lower sextile} \tag{16}$$

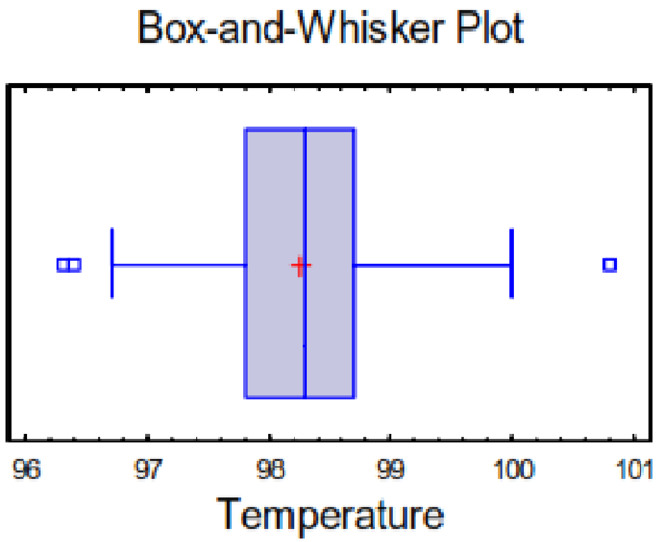
2k+1 Point Representation

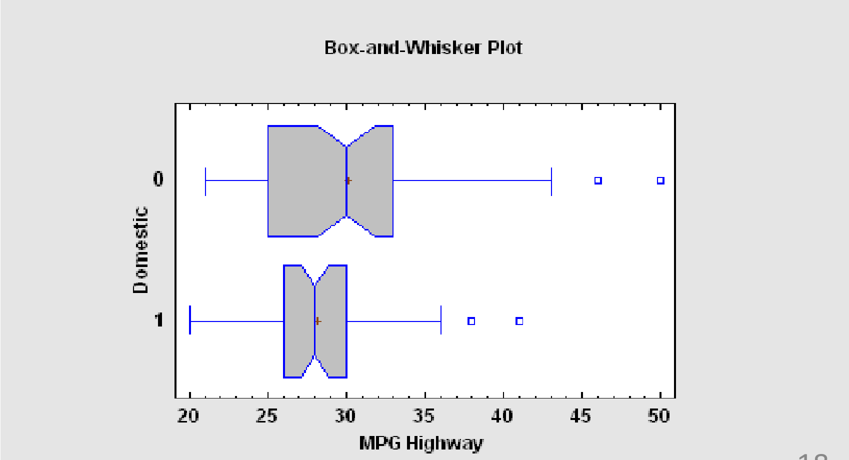
- 3 point Summery
- 5 point Summery
- 7 point Summery



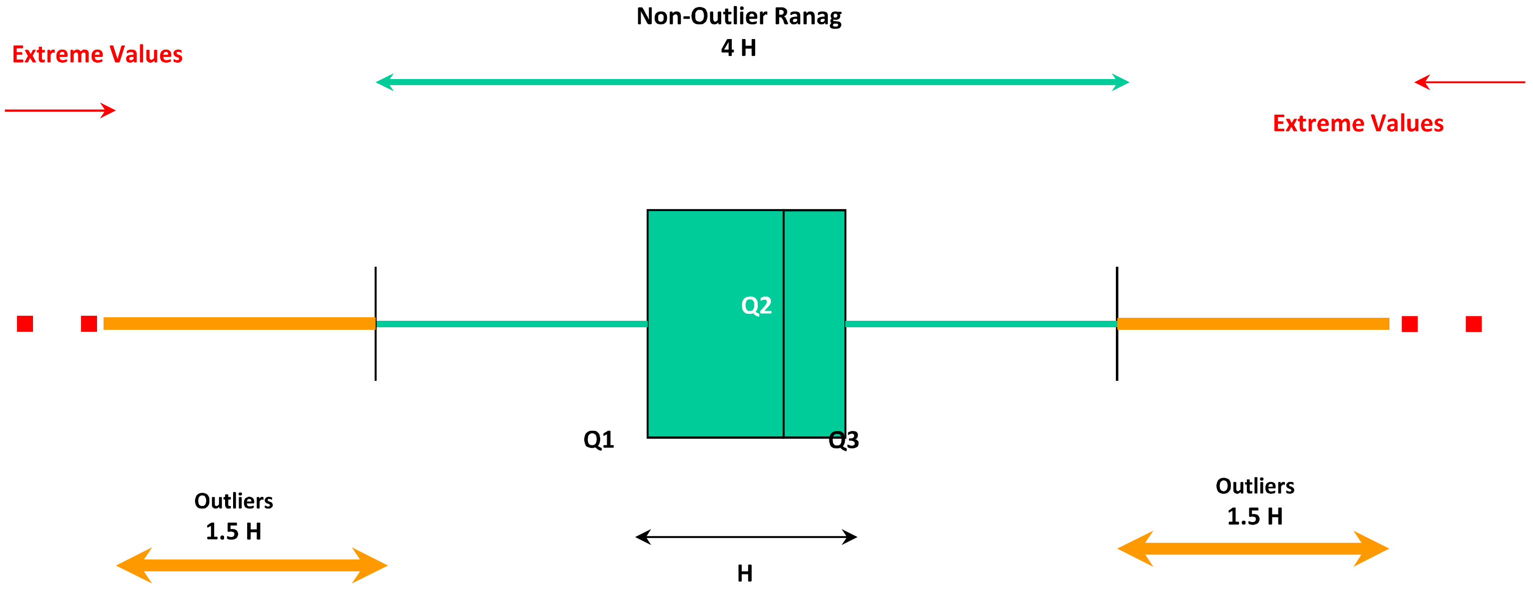
Box-and-Whisker Plots

- Box-and-whisker plots are graphical displays based upon Tukey's 5-number summary of a data sample. In his original plot, a box is drawn covering the center 50% of the sample. A vertical line is drawn at the median, and whiskers are drawn from the central box to the smallest and largest data values. If some points are far from the box, these "outside points" may be shown as separate point symbols. Later analysts have added notches showing approximate confidence intervals for the median, and plus signs at the sample mean.



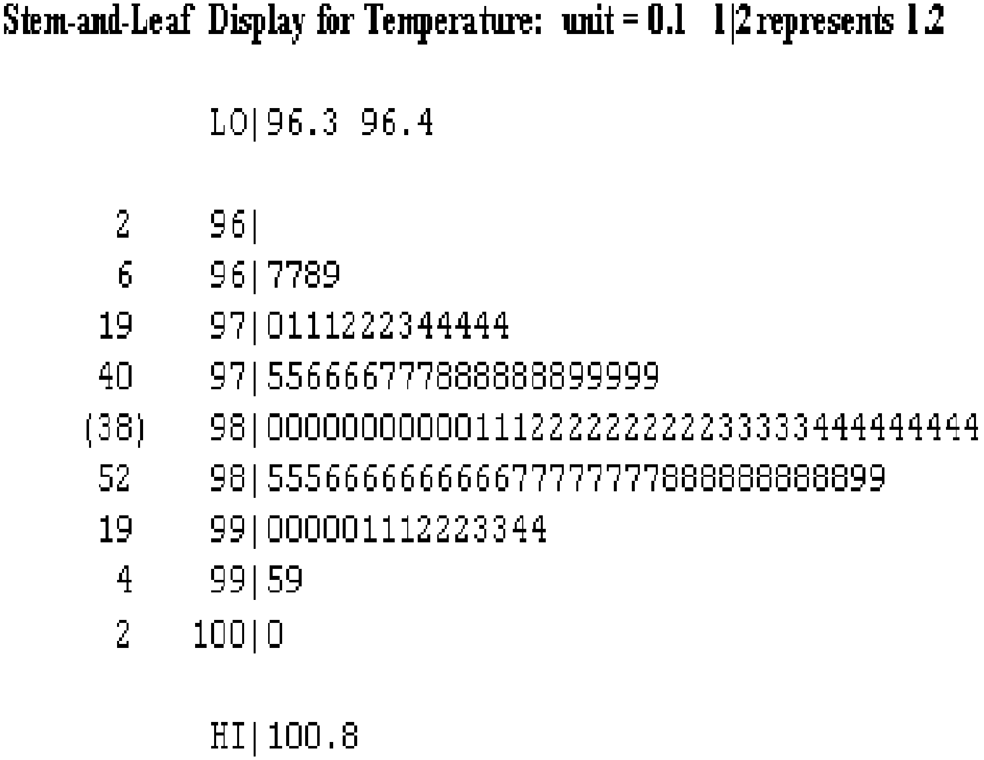


- The plot is constructed in the following manner:
- A box is drawn extending from the lower quartile of the sample to the upper quartile.
- This is the interval covered by the middle 50% of the data values when sorted from smallest to largest.
- A vertical line is drawn at the median (the middle value).
- If requested, a plus sign is placed at the location of the sample mean.
- Whiskers are drawn from the edges of the box to the largest and smallest data values, unless there are values unusually far away from the box (which Tukey calls outside points). Outside points, which are points more than 1.5 times the interquartile range
- (box width) above or below the box, are indicated by point symbols. Any points more than 3 times the interquartile range above or below the box are called far outside points, and are indicated by point symbols with plus signs superimposed on top of them. If outside points are present, the whiskers are drawn to the largest and smallest data values which are not outside points.


Non-Outlier Ranag
4 H
Extreme Values
Extreme Values
Q2
Q1
Q3
Outliers
1.5 H
H
Outliers
1.5 H

Stem-and-Leaf Display

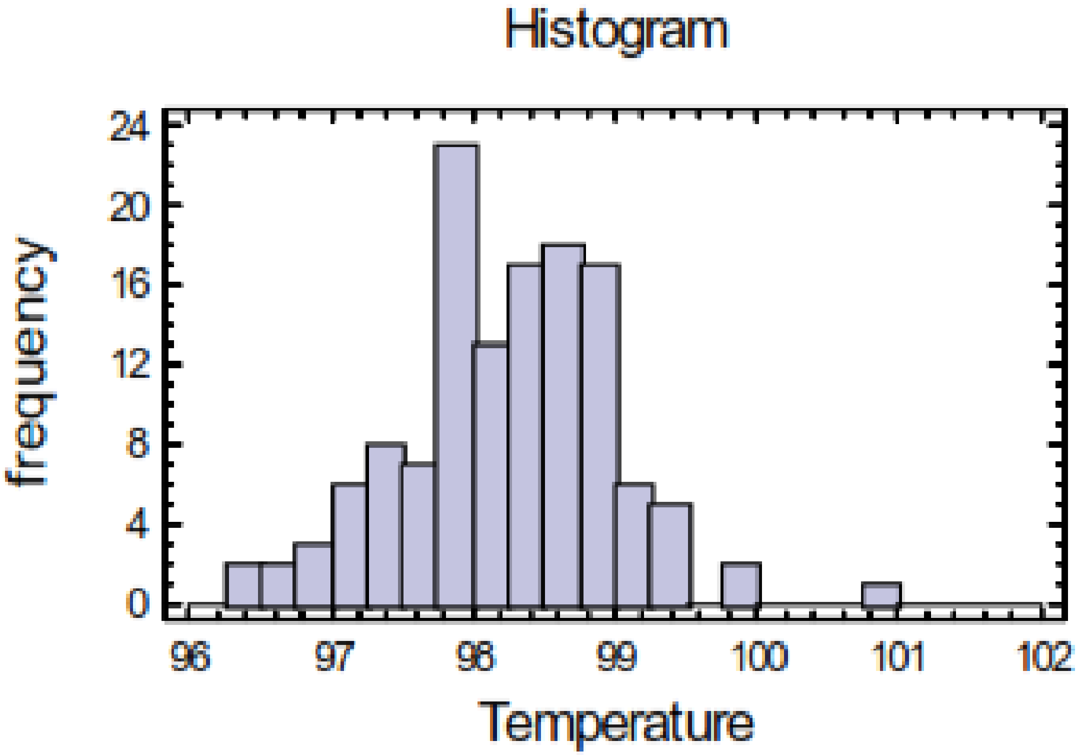
Stem-and-leaf displays take each data value and divide it into a stem and a leaf. For example, the temperature of the first subject in the data sample to the left had a body temperature of 98.4 degrees. The first two digits (“98”) are called the stem and plotted at the left, while the third digit (“4”) is called the leaf. Although similar to a histogram turned on its side, Tukey thought that the stem-and-leaf plot was preferable to a barchart since the data values could be recovered from the display.



```
Stem-and-Leaf Display for Temperature:  unit = 0.1   1|2 represents 1.2

          LO|96.3 96.4

   2      96|
   6      96|7789
  19      97|0111222344444
  40      97|556666777888888899999
 (38)     98|00000000000111222222222233333444444444
  52      98|555666666666677777777888888888899
  19      99|000001112223344
   4      99|59
   2     100|0

          HI|100.8
```

Histogram

The Frequency Histogram displays the results of the tabulation in the form of a barchart or lineplot

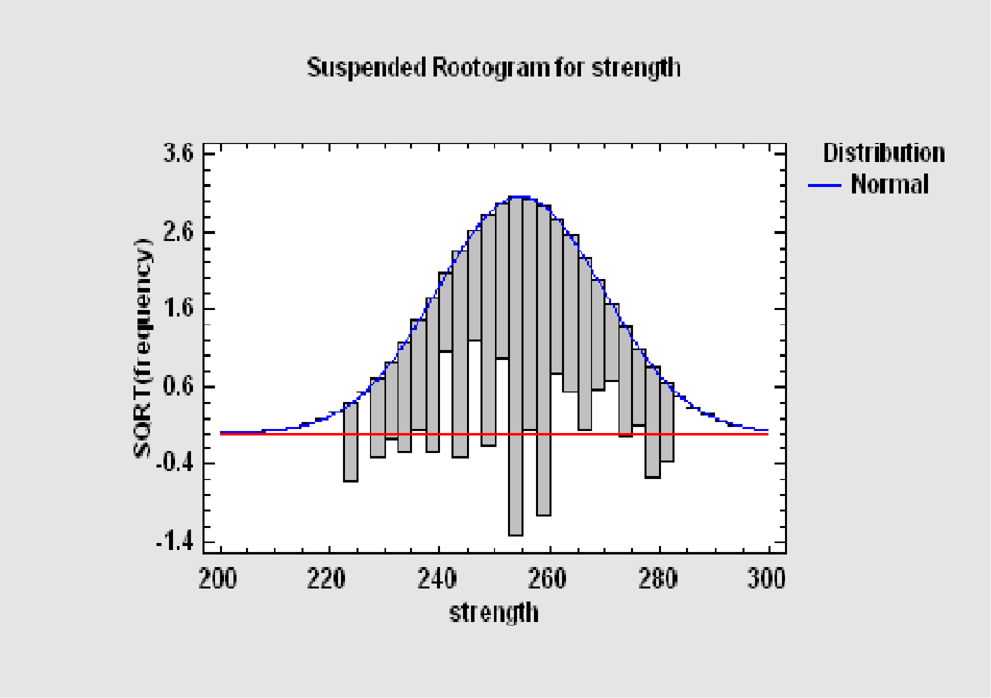


Rootogram

A rootogram is similar to a histogram, except that it plots the square roots of the number of observations observed in different ranges of a quantitative variable.

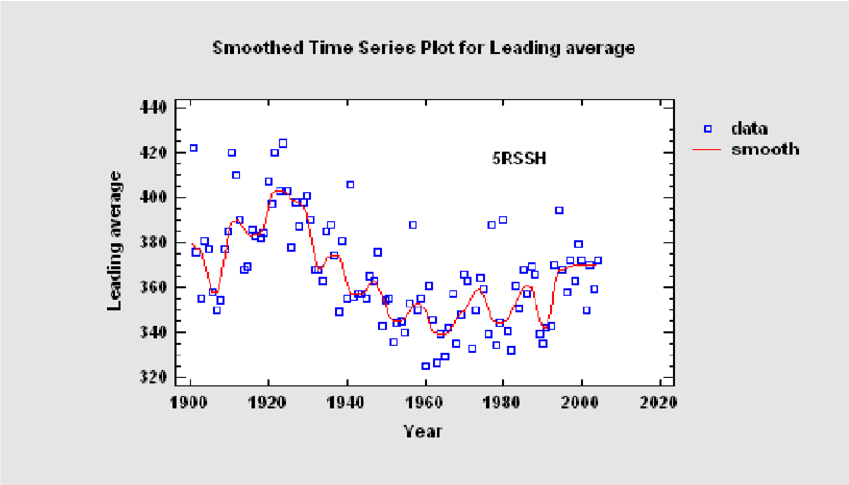
It is usually plotted together with a fitted distribution. The idea of using square roots is to equalize the variance of the deviations between the bars and the curve, which otherwise would increase with increasing frequency.

Sometimes, the bars are suspending the from the fitted distribution, which allows for easier visual comparison with the horizontal line drawn at 0, since visual comparison with a curved line may be deceiving.



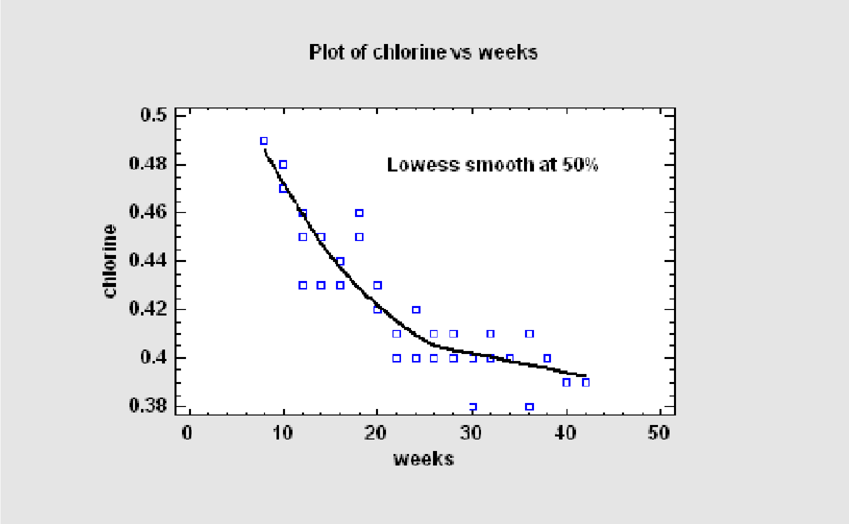
Resistant Time Series Smoothing

- Tukey invented a number of nonlinear smoothers, used to smooth sequential time series data, that are very good at ignoring outliers and are often applied as a first step to reduce the influence of potential outliers before a moving average is applied.
- These include 3RSS, 3RSSH, 5RSS, 5RSSH, and 3RSR smoothers. Each symbol in the name of the smoother indicates an operation that is applied to the data.



Scatterplot Smoothing

- X-Y scatterplots may be smoothed using any of several methods: running means, running lines, LOWESS (locally weighted scatterplot smoothing), and resistant LOWESS.
- Smoothers are useful for suggesting the type of regresson model that might be appropriate to describe the relationship between two variables.



Median Polish

- The Median Polish procedure constructs a model for data contained in a two-way table. The model represents the contents of each cell in terms of:

	1	2	3	4	5	6
A	70	16	3	57	71	29
B	25	4	54	16	45	48
C	3	49	53	93	52	23
D	67	63	10	85	16	45
E	83	16	30	45	8	5

Median polish:

(1) find the row medians for each row, find the median of the row medians, record this as the **overall effect**.

(2) subtract each element in a row by its row median, do this for all rows.

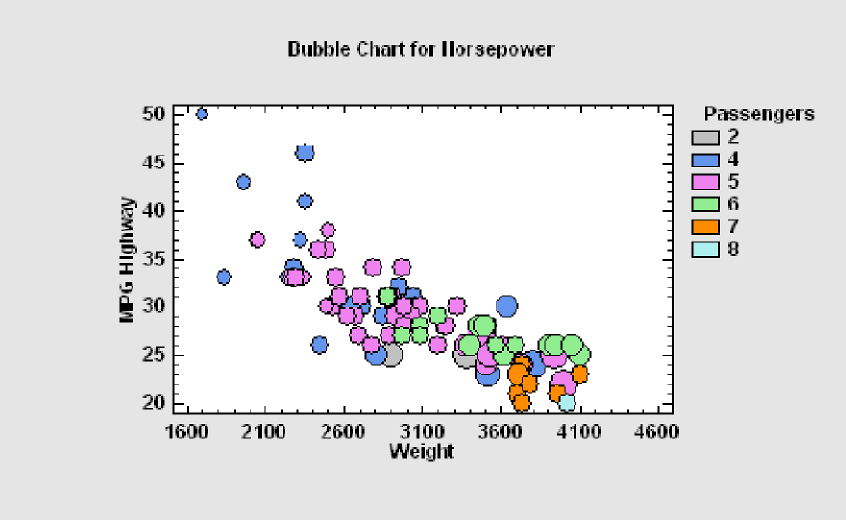
(3) subtract the **overall effect** from each row median.

(4) do the same for each column, and add the **overall effect** from column operations to the **overall effect** generated from row operations.

(5) repeat (1)-(4) until negligible change occur with row or column medians

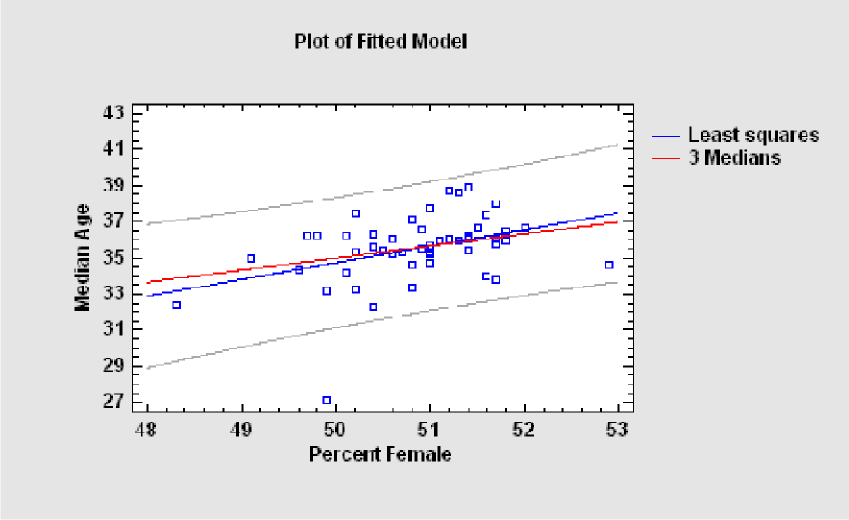
Bubble Chart

- The Bubble Chart is an X-Y scatterplot on which the value of a third and possibly fourth variable is shown by changing the size and/or color of the point symbols.
- It is one way to plot multivariate data in 2 dimensions.



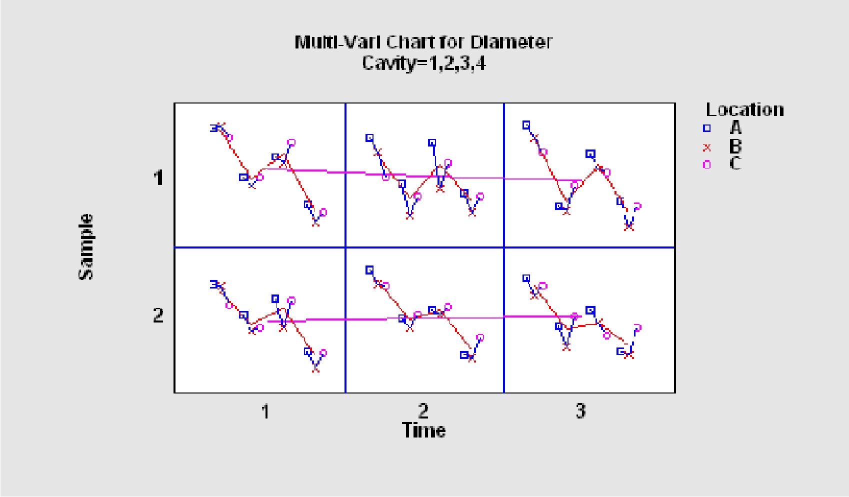
Resistant Curve Fitting

- Tukey proposed a method for fitting lines and other curves that is less influenced by any outliers that might be present.
- Called the method of 3 medians, the data are first divided into 3 groups according to the value of X.
- Medians are then computed within each group, and the curve is determined from the 3 medians.

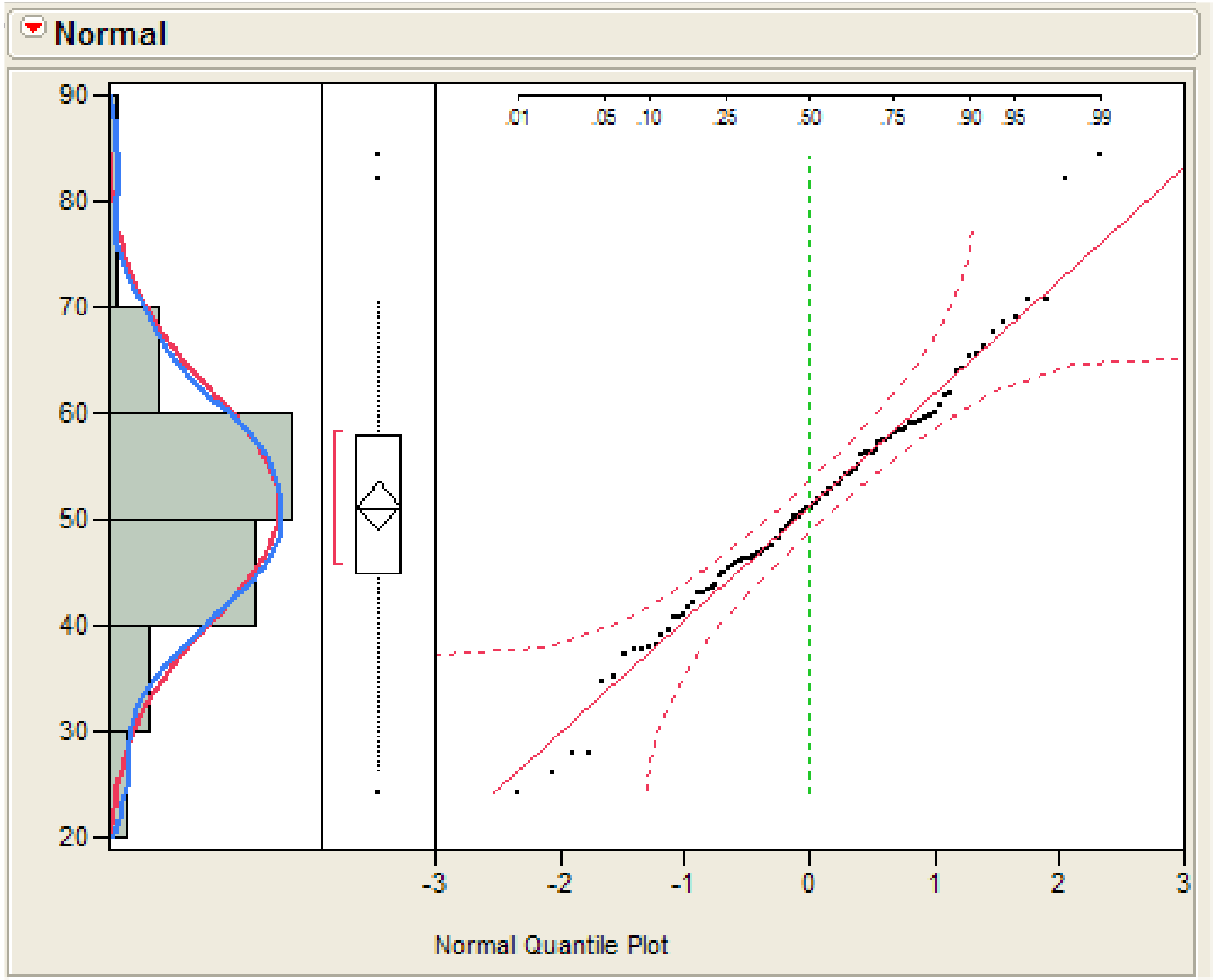


Multi-Vari Chart

- A Multi-Vari Chart is a chart designed to display multiple sources of variability in a way that enables the analyst to identify easily which factors are the most important.
- It is commonly used to display EDA data from a designed experiment prior to performing a formal statistical analysis.



Normal Quantile Plot

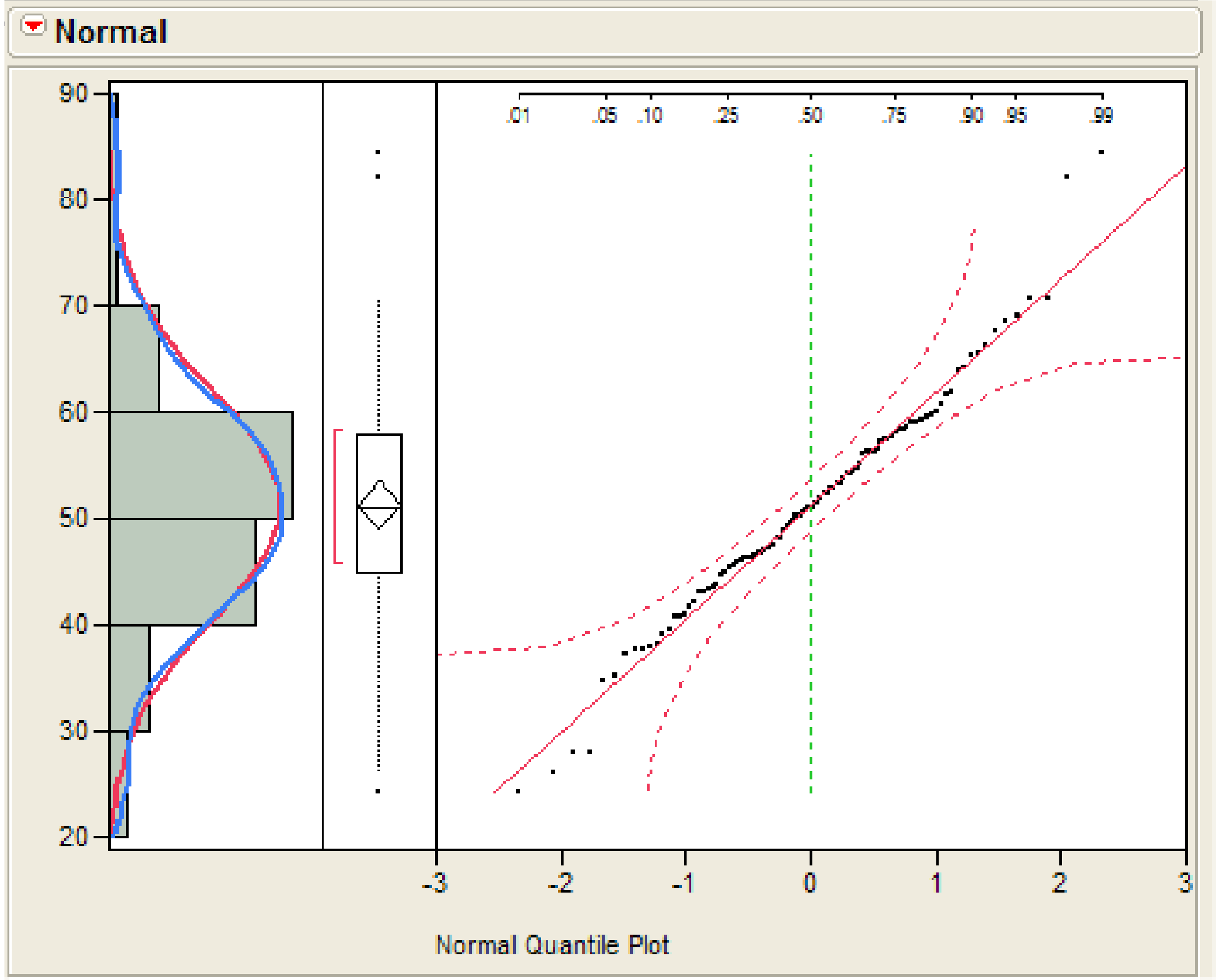


THE IDEAL PLOT:

Here is an example where the data is perfectly normal. The plot on right is a normal quantile plot with the data on the vertical axis and the expected z-scores if our data was normal on the horizontal axis.

When our data is approximately normal the spacing of the two will agree resulting in a plot with observations lying on the reference line in the normal quantile plot. The points should lie within the dashed lines.

Normal Quantile Plot

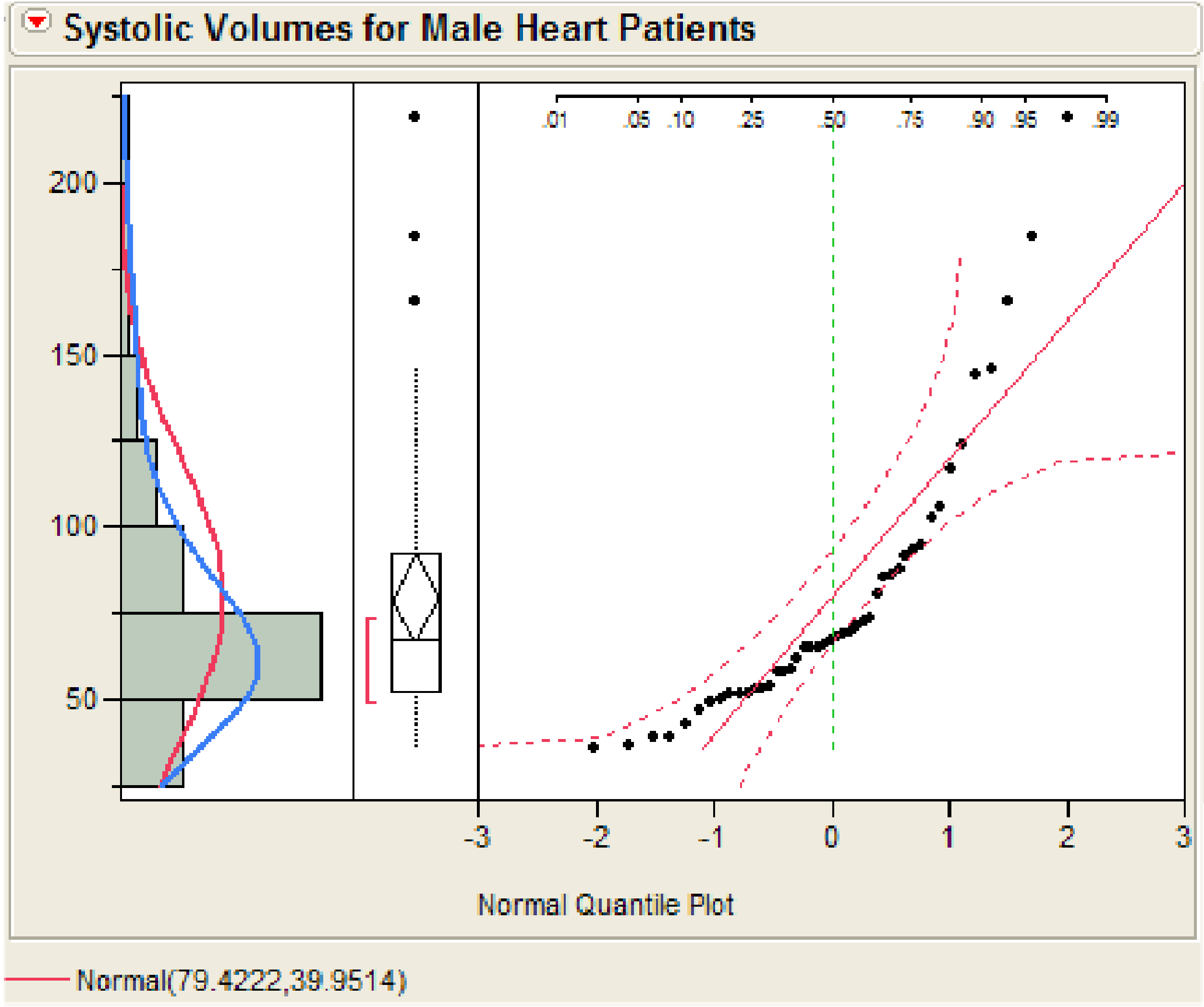


THE IDEAL PLOT:

Here is an example where the data is perfectly normal. The plot on right is a normal quantile plot with the data on the vertical axis and the expected z-scores if our data was normal on the horizontal axis.

When our data is approximately normal the spacing of the two will agree resulting in a plot with observations lying on the reference line in the normal quantile plot. The points should lie within the dashed lines.

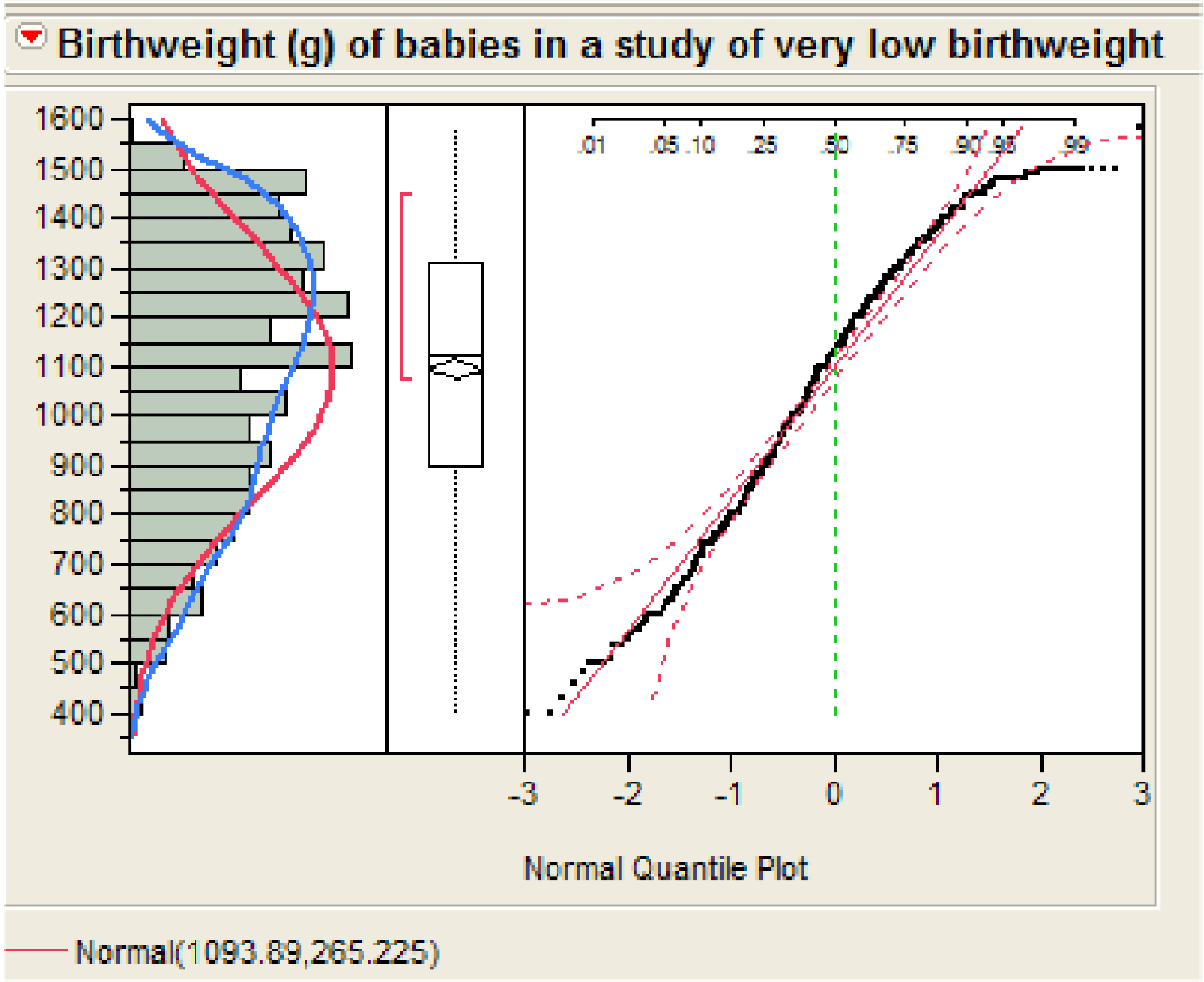
Normal Quantile Plot
(right skewness)



The systolic volumes of the male heart patients ar clearly **right skewed**.

When the data is plotted vs. the expected z-scores the normal quantile plot shows right skewness by a **upward bending** curve

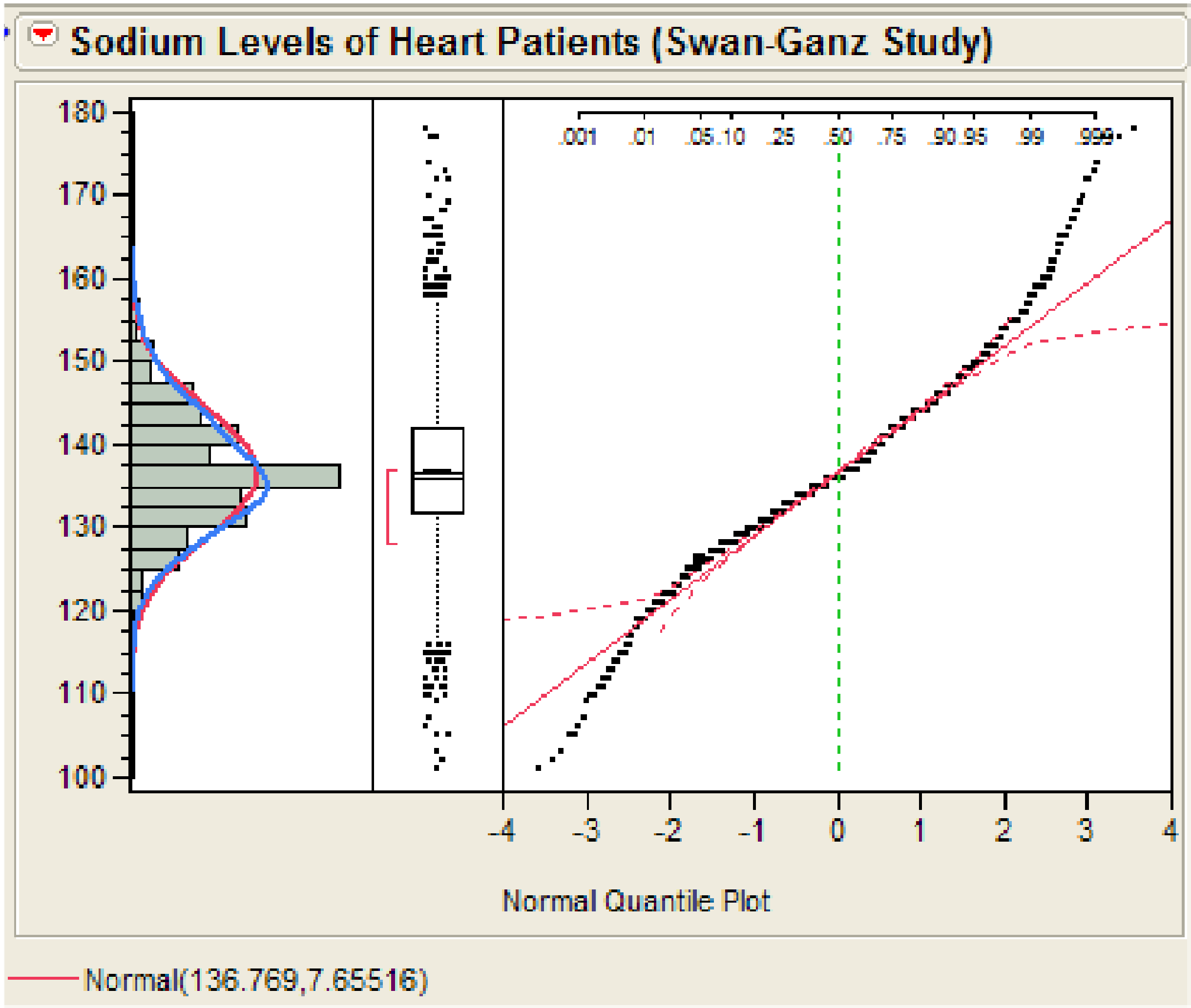
Normal Quantile Plot
(left skewness)



The distribution of birthweights from this study of very low birthweight infants is **skewed left**.

When the data is plotted vs. the expected z-scores the normal quantile plot shows left skewness by a **downward bending** curve.

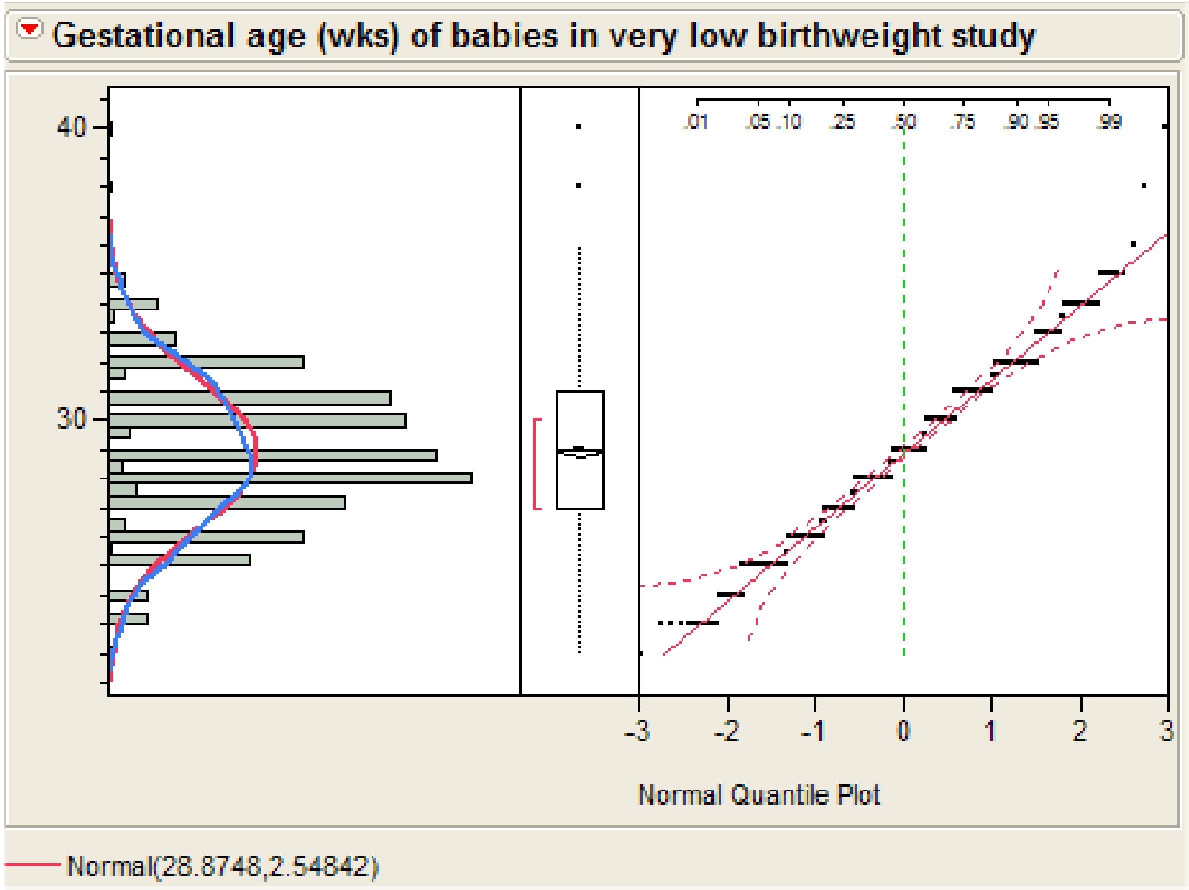
Normal Quantile Plot (leptokurtosis)



The distribution of sodium levels of patients in this right heart catheterization study has heavier tails than a normal distribution (i.e, leptokurtosis).

When the data is plotted vs. the expected z-scores the normal quantile plot there is an “S-shape” which indicates kurtosis.

Normal Quantile Plot (discrete data)



Although the distribution of the gestational age data of infants in the very low birthweight study is approx. normal there is a **"staircase"** appearance in normal quantile plot.

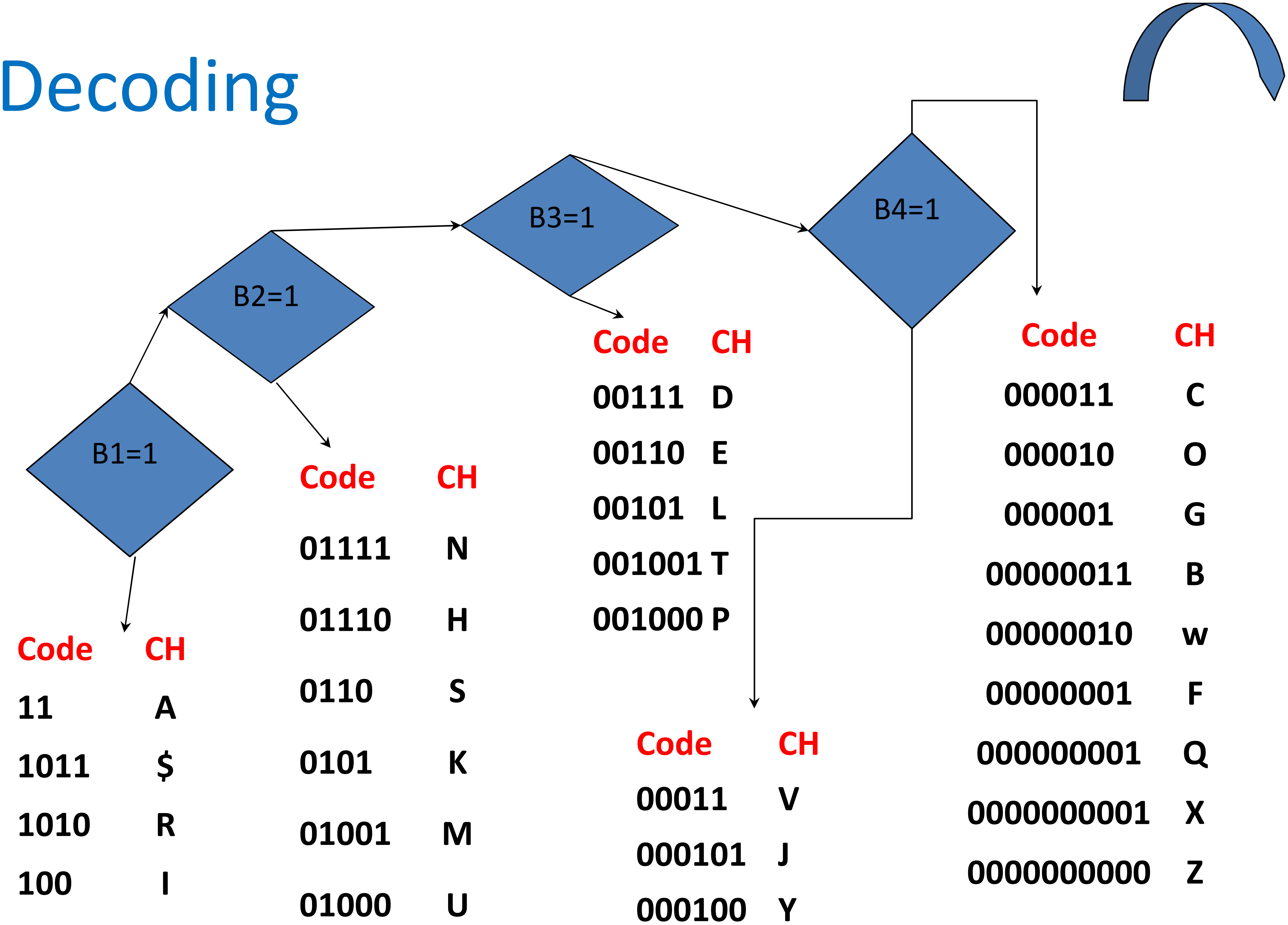
This is due to the **discrete coding** of the gestational age which was recorded to the nearest week or half week.

Code your name

CH	Code	length
$	1011	4
A	11	2
B	00000011	8
C	000011	6
D	00111	5
E	00110	5
F	00000001	8
G	000001	6
H	01110	5
I	100	3
J	000101	6
K	0101	4
L	00101	5
M	01001	5

CH	Code	length
N	01111	5
o	000010	6
P	001000	6
q	000000001	9
R	1010	4
S	0110	4
t	001001	6
U	01000	5
V	00011	5
w	00000010	8
x	0000000001	10
Y	000100	6
z	0000000000	10

Decoding



Code	CH
11	A
1011	$
1010	R
100	I

Code	CH
01111	N
01110	H
0110	S
0101	K
01001	M
01000	U

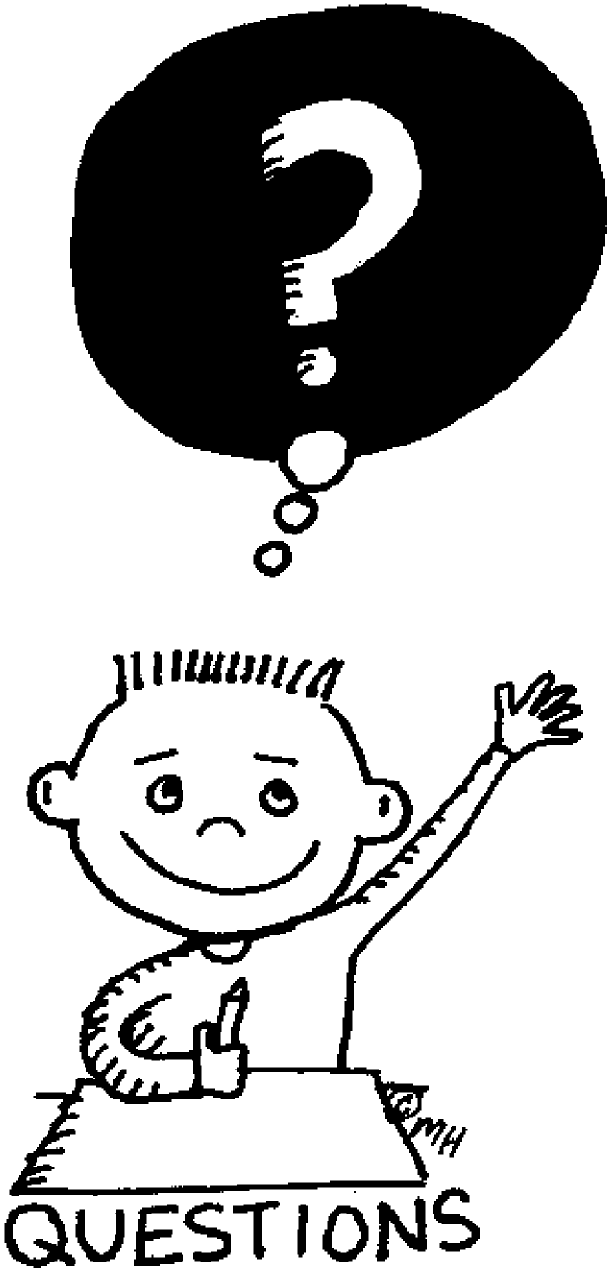
Code	CH
00111	D
00110	E
00101	L
001001	T
001000	P

Code	CH
00011	V
000101	J
000100	Y

Code	CH
000011	C
000010	O
000001	G
00000011	B
00000010	w
00000001	F
000000001	Q
0000000001	X
0000000000	Z

Dimension Reduction

- Principal components (PCA)
- Factor Analysis
- ICA
- Dimension Reduction
  - Stepwise Regression
  - Reduct



Collaborations

Thank you