Lecture 2: Describing Data (Part 2)

STAT 515 — Statistical Methods I

Dr. Yen-Yi Ho

Department of Statistics, University of South Carolina

Outline

- What is Statistics
- Types of Data: Continuous, Binary, Categorical, Time-to Event
- Continuous Data
- Methods for Organizing Data
  1. Ordering Data
  2. Grouping Data
  3. Summarizing Data
  4. Graphing Data

Methods for Organizing Data

- **Ordering data**
  - Tallies
  - Stem-and-leaf displays
- **Grouping data**
  - Frequency distributions
  - Percentiles
- **Summarizing data**
  - Measures of central tendency
  - Measures of dispersion
  - Box-and-whiskers plots
- **Displaying data**
  - Tables
  - Graphs

Outline

Ordering Data — Tallies

Example: Ages of Graduate Students

Unordered Data ($n = 10$):
35, 40, 52, 27, 31, 42, 43, 28, 50, 35

Ordered Data (Order Statistics) ($n = 10$):
27, 28, 31, 35, 35, 40, 42, 43, 50, 52

Could order by hand, or **tally** into age groups:

Age Group	**Tally**
20–29	//
30–39	///
40–49	///
50–59	//

Ordering Data — Stem and Leaf Displays

An easier way: **stem-and-leaf displays**

Unordered

Stem	Leaves
2	7 8
3	5 1 5
4	0 2 3
5	2 0

Ordered

Stem	Leaves
2	7 8
3	1 5 5
4	0 2 3
5	0 2

Ordering Data — Stem and Leaf Displays (cont'd)

- Aids in sorting or ordering data.
- Retains more information than a tally.
- Consider sensible categories for the number of stems.

The example could also be shown for 10-unit groupings as:

2	7 8		2*	7 8
3	1 5 5		3*	1 5 5
4	0 2 3	or as	4*	0 2 3
5	0 2		5*	0 2

where $2^* = 20$–29, $3^* = 30$–39, etc.

Ordering Data — Stem and Leaf Display (cont'd)

For 5-year age groupings, one could display:

2*	
2.	7 8
3*	1
3.	5 5
4*	0 2 3
4.	
5*	0 2
5.	

where $2^* = 20\text{–}24$, $2. = 25\text{–}29$, etc.

Outline

Grouping Data — Frequency Distribution

Definition: Frequency Distribution Table

A list of the values or ranges of values of a variable and their frequencies of occurrence.

Age Interval	**Frequency**
20–29	2
30–39	3
40–49	3
50–59	2
Total	**10**

Guideline: choose the intervals based upon your knowledge of the range of the data, aiming for roughly $\sqrt{n}$ bins.

Grouping Data — Frequency Distributions (cont'd)

Definition: Empirical Distribution Function

A cumulative frequency distribution which describes the observed values of a variable.

Age Interval	Freq.	Cum. Freq.	Rel. Freq.	Cum. Rel. Freq.
20–29	2	2	0.2	0.2
30–39	3	5	0.3	0.5
40–49	3	8	0.3	0.8
50–59	2	10	0.2	1.0
Total	**10**		**1.0**	

Grouping Data — Percentiles

Definition: r^{th} Percentile (P_r)

The value that is greater than or equal to r percent of a sample of observations, and less than or equal to $(1 - r)$ percent of the observations.

Percentile		**Quartile**	**% of observations falling below**
P_{75}	$=$	Q_3	75%
P_{50}	$=$	Q_2	50% (median)
P_{25}	$=$	Q_1	25%

Grouping Data — Percentiles (cont'd)

Using the example ($n = 10$), a simple way to calculate quartiles:

- Q_2 = median = average of the 5th and 6th observations $\Rightarrow$ 37.5
- Q_1 = median of the lower half of data; third smallest value $\Rightarrow$ 31
- Q_3 = median of the upper half of the data; third largest value $\Rightarrow$ 43

What if $n = 11$ (odd in general)? Include the median in both the upper and lower halves of the data.

Summarizing Data — Measures of Central Tendency

- **mean (average)**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- **median** = middle observation after ordering data
- **mode** = most frequent observation

Summarizing Data — Measures of Dispersion (or Spread)

- **Range** = difference between largest and smallest values
- **Variance** = "average" of the squared differences of observations from the sample mean

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- **Standard deviation**

$$s = \sqrt{s^2}$$

Summarizing Data — Example

Example: Graduate Student Ages

27, 28, 31, 35, 35, 40, 42, 43, 50, 52

Statistic	**Value**
Mean ($\bar{x}$)	38.3 years
Mode	35 years
Median	$(35+40)/2 = 37.5$ years
Range	$52 - 27 = 25$ years
Variance (s^2)	74.7 years^2
Standard Deviation (s)	$\sqrt{74.7} = 8.6$ years

Summarizing Data — Box-and-Whiskers Plots

Graphical display using quartiles.

- Upper Hinge $= Q_3$
- Median $= Q_2$
- Lower Hinge $= Q_1$
- Interquartile Range (IQR) $= Q_3 - Q_1$ (contains the middle 50% of the observations)
- Whiskers = lines drawn to the smallest and largest observations within the calculated **fences**

Summarizing Data — Box-and-Whiskers Plots (cont'd)

What are "fences"?

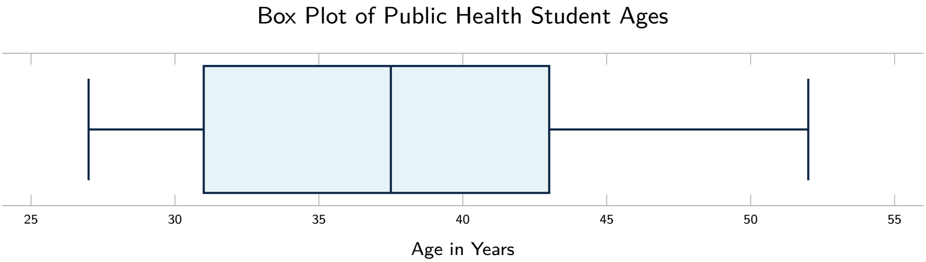
- Fences are *not* observed data points.
- Fences are calculated to provide guidelines for identifying outliers.
    - Upper Fence $=$ Upper Hinge $+1.5\times$ IQR
    - Lower Fence $=$ Lower Hinge $-1.5\times$ IQR
- Outliers are data values beyond the calculated fences (high or low).

Box-and-Whiskers Plot — Example

Back to the example of public health student ages:

- $\text{IQR} = Q_3 - Q_1 = 43 - 31 = 12$ years
- Upper Fence $=$ Upper Hinge $+1.5\times$ IQR $= 43 + 1.5(12) = 61$ years
- Lower Fence $=$ Lower Hinge $-1.5\times$ IQR $= 31 - 1.5(12) = 13$ years

Box-and-Whiskers Plot — Example (cont'd)



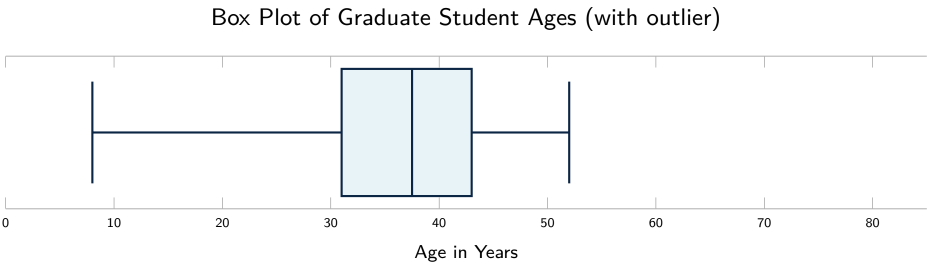
- $Q_1 = 31, \quad Q_2 = 37.5, \quad Q_3 = 43$
- $\min = 27, \quad \max = 52$

Another Box-and-Whiskers Plot Example

Suppose the data set contains two more students with ages 80 and 8 ($n = 12$).

- Upper fence $= 72$
- Lower fence $= 4$
- The boxplot would now show the outlying value of age 80 years.

Another Box-and-Whiskers Plot Example (cont'd)

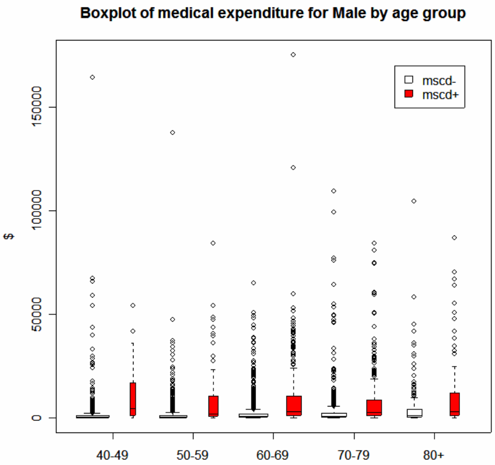


- **Outlier** at age 80 (beyond the upper fence of 72)
- Min within fences $= 8$; Max within fences $= 52$

Comparing Data — Box-and-Whiskers Plots

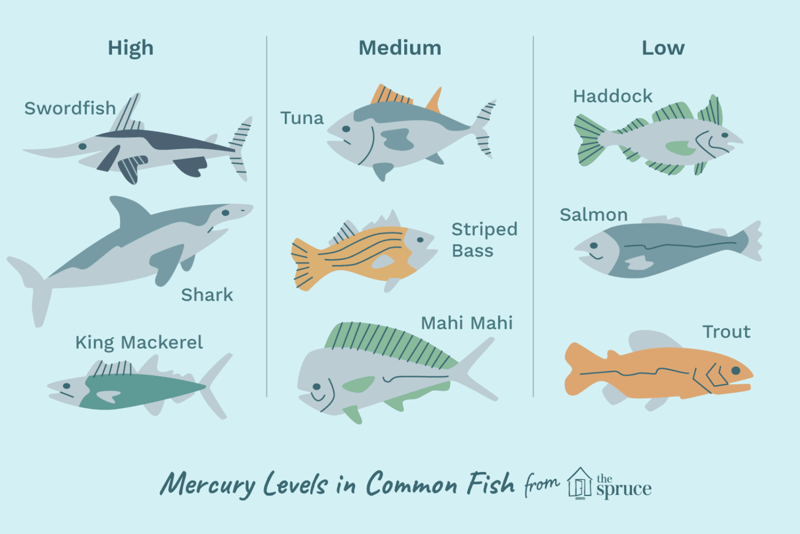
- Box plots can be used to compare different groups.
- The width of the box could be proportional to the size of the data set.

Comparing Data — Box-and-Whiskers Plots Example



mscd = major smoking-caused disease

I Like Fish, But ...



Mercury in Fish Data

The HgFish dataset in the NADA (Nondetects and Data Analysis for Environmental Data) R package contains U.S. mercury and fish tissue concentration data with multiple detection limits (0.03, 0.05, and 0.10 μg/g wet weight). It is used to evaluate how mercury levels relate to watershed land use, water chemistry, and sediment characteristics.

Mercury in Fish Data

```
###install.packages("NADA")
### install.packages("GGally")

library(NADA)
library(GGally)

data(HgFish)
head(HgFish)
```

Mercury in Fish Data

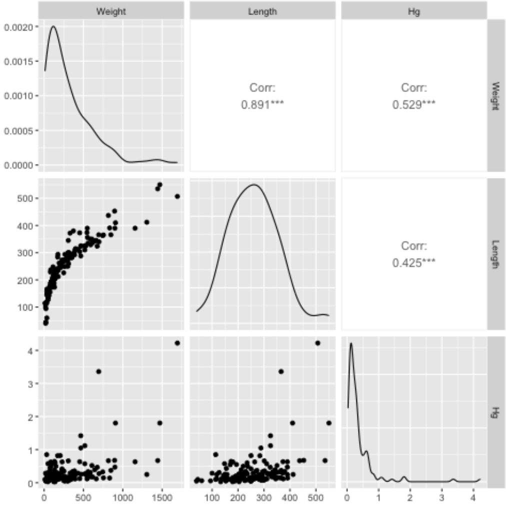
```
> head(HgFish)
  LandUse           Species Weight Length    Hg HgCen WatMeHg WatTotHg
1     Urb   LargemouthBass    311    260 0.127 FALSE   0.015     6.50
2     Urb FlatheadCatfish     191    262 0.110 FALSE   0.015     6.50
3     Urb   LargemouthBass   1305    412 0.249 FALSE   0.015     1.34
4     Urb   LargemouthBass    849    366 0.373 FALSE   0.015     1.34
5     Urb   LargemouthBass    526    324 0.249 FALSE   0.015     1.34
6     Urb   LargemouthBass    306    278 0.125 FALSE   0.015     1.34
  SedMeHg SedTotHg WatDOC SedLOI SedAVS PctWetland
1   0.281     30.6    5.8 0.0442     67          0
2   0.281     30.6    5.8 0.0442     67          0
3   0.199     45.2    6.0 0.0759   2394          1
4   0.199     45.2    6.0 0.0759   2394          1
5   0.199     45.2    6.0 0.0759   2394          1
6   0.199     45.2    6.0 0.0759   2394          1
```

Mercury in Fish Data

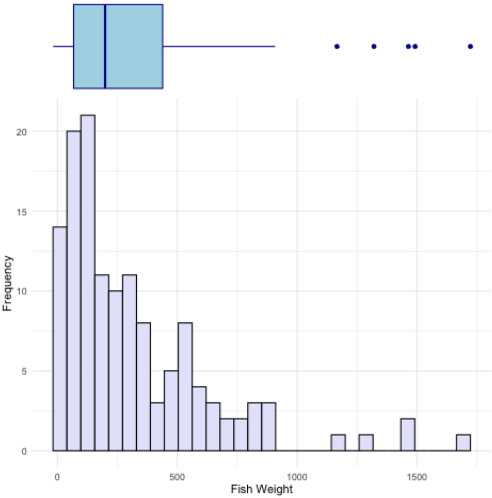
```
> str(HgFish)
'data.frame': 133 obs. of  14 variables:
 $ LandUse    : Factor w/ 5 levels "AF","Ag","Bkg",..: 5 5 5 5 5 5 5 5 3 3 ...
 $ Species    : Factor w/ 20 levels "Bowfin","BrownTrout",..: 8 7 8 8 8 8 8 8 15 15 ..
 $ Weight     : int  311 191 1305 849 526 306 341 663 40 87 ...
 $ Length     : int  260 262 412 366 324 278 282 345 150 180 ...
 $ Hg         : num  0.127 0.11 0.249 0.373 0.249 0.125 0.1 0.117 0.258 0.157 ...
 $ HgCen      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ WatMeHg    : num  0.015 0.015 0.015 0.015 0.015 0.015 0.015 0.015 NA NA ...
 $ WatTotHg   : num  6.5 6.5 1.34 1.34 1.34 1.34 1.24 1.24 NA NA ...
 $ SedMeHg    : num  0.281 0.281 0.199 0.199 0.199 0.199 0.519 0.519 0.233 0.233 ...
 $ SedTotHg   : num  30.6 30.6 45.2 45.2 45.2 45.2 55.4 55.4 5.9 5.9 ...
 $ WatDOC     : num  5.8 5.8 6 6 6 6 5.6 5.6 NA NA ...
 $ SedLOI     : num  0.0442 0.0442 0.0759 0.0759 0.0759 0.0759 0.0624 0.0624 0.0218 0.
 $ SedAVS     : num  67 67 2394 2394 2394 ...
 $ PctWetland: num  0 0 1 1 1 1 0 0 0 0 ...
```

Mercury in Fish Data

Mercury concentrations in 133 fish collected across the United States.



Box Plot and Histogram of Fish Weight



Outline

1. Methods for Organizing Data
2. Ordering Data
3. Grouping Data
4. Summarizing Data
5. Graphing Data

How to Make a Histogram

- Consider the following data collected from the 1995 Statistical Abstracts of the United States
  - For each of the 50 United States, the proportion of individuals over 65 years of age has been recorded

How to Make a Histogram

State	%	State	%	State	%	State	%
AL	12.9	IN	12.8	NE	14.1	SC	11.9
AK	4.6	IA	15.4	NV	11.3	SD	14.7
AZ	13.4	KS	13.9	NH	11.9	TN	12.7
AR	14.8	KY	12.7	NJ	13.2	TX	10.2
CA	10.6	LA	11.4	NM	11.0	UT	8.8
CO	10.1	ME	13.9	NY	13.2	VT	12.1
CT	14.2	MD	11.2	NC	12.5	VA	11.1
DE	12.7	MA	14.1	ND	14.7	WA	11.6
FL	18.4	MI	12.4	OH	13.4	WV	15.4
GA	10.1	MN	12.5	OK	13.6	WI	13.4
HI	12.1	MI	12.5	OR	13.7	WY	11.1
ID	11.6	MO	14.1	PA	15.9		
IL	12.6	MT	13.3	RI	15.6		

Source: Statistical Abstracts of the United States (1995).

How to Make a Histogram

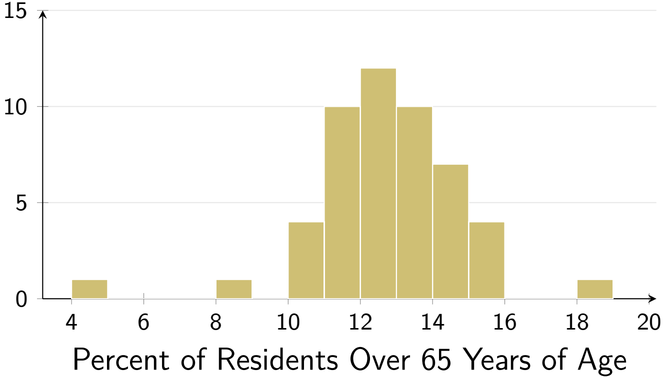
- **Break the data range into mutually exclusive, equally sized "bins":** here each is 1% wide
- **Count the number of observations in each bin**

Class	Count	Class	Count	Class	Count
$[4.0, 5.0)$	1	$[9.0, 10.0)$	0	$[14.0, 15.0)$	7
$[5.0, 6.0)$	0	$[10.0, 11.0)$	5	$[15.0, 16.0)$	4
$[6.0, 7.0)$	0	$[11.0, 12.0)$	9	$[16.0, 17.0)$	0
$[7.0, 8.0)$	0	$[12.0, 13.0)$	12	$[17.0, 18.0)$	0
$[8.0, 9.0)$	1	$[13.0, 14.0)$	10	$[18.0, 19.0)$	1

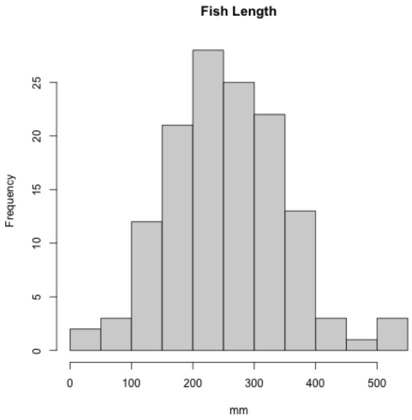
$$\bar{x} = 12.7\%; \quad s = 2.1\%$$

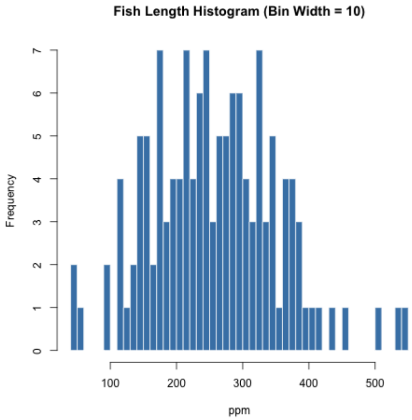
How to Make a Histogram

- **Draw the histogram**
- **Label scales**



Mercury in Fish Data (n=133)





$$\overline{x} = 258.14(mm)$$

$$sd = 95.09(mm)$$

Intervals

- **How many intervals (bins) should you have in a histogram?**
  - There is no perfect answer to this
  - Depends on sample size n
  - Rough rule of thumb: $\#$ Intervals $\approx \sqrt{n}$

n	**Number of Intervals**
10	About 3
50	About 7
100	About 10

Summarizing Data — Features to Notice

1. Skewness

- **Positively skewed (right skewed)** — more lower values, sparse higher values
  - Also: long "tail" of higher values
  - Also: mean $>$ median $>$ mode
- **Negatively skewed (left skewed)** $=$ reverse of positively skewed
- **Symmetric** $=$ not skewed in either direction

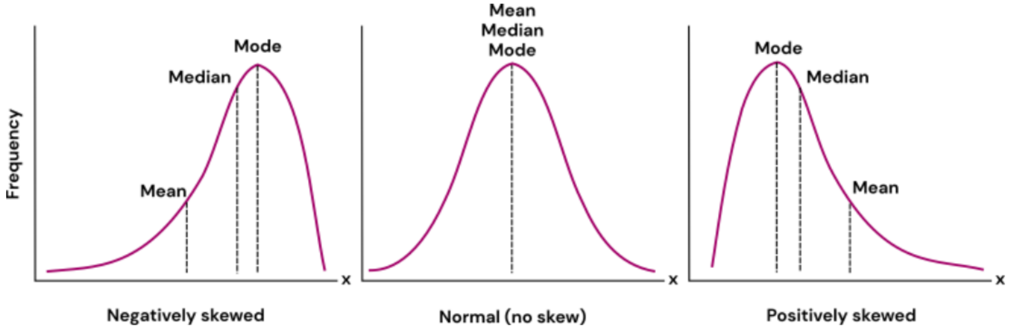
Summarizing Data — Measures of Central Tendency

- **mean (average)**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- **median** $=$ middle observation after ordering data
- **mode** $=$ most frequent observation

Location Measures



2. Outlying Values

- Values that are "far" from most values.
- **Importance**: a few outlying values can strongly influence certain statistical summary measures and analyses.

Example: Age Data

With the 80-year-old added ($n = 11$):

Data	**Mean**	**Median**
Original ($n = 10$)	38.3	37.5
With 80 added ($n = 11$)	42.1	40

Graphing on an arithmetic vs. logarithmic scale

- On an **arithmetic scale**, each increment represents change by a constant amount.
- On a **logarithmic scale**, each increment represents change by a constant multiplier.

Logarithmic Scales

Logarithm to the base 10 (common log):

$$x = 10^y \quad \text{or} \quad \log_{10}(x) = y$$

$$\log_{10}(1) = 0 \text{ since } 10^0 = 1, \qquad \log_{10}(100) = 2 \text{ since } 10^2 = 100$$

Logarithm to the base e (natural log):

$$x = e^y \quad \text{or} \quad \log_e(x) = y$$

Why Use Logarithms

- Allows plotting of numbers of different orders of magnitude on the same graph.
- May aid in data analysis and data transformations.
- May describe a biological relationship (e.g. exponential growth) more accurately.

Example of Logarithmic Graph

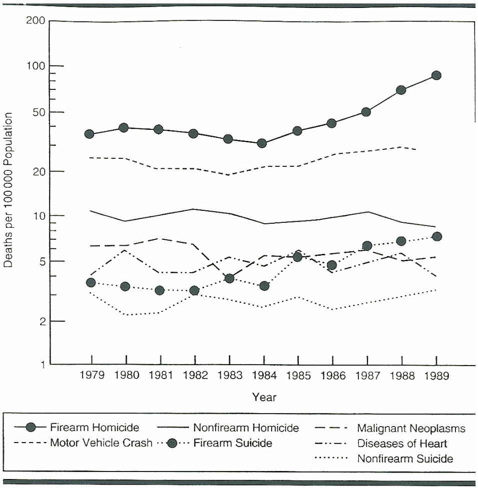
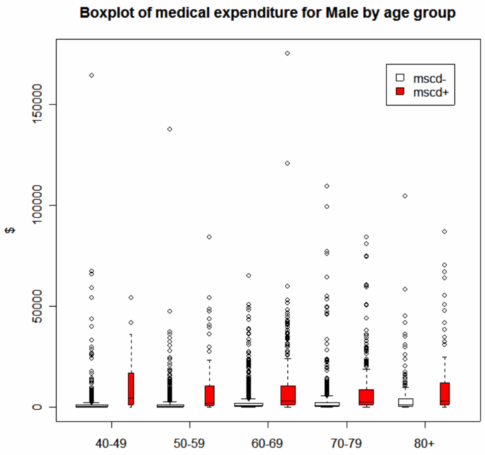


Fig 1.—Leading causes of death for black males 15 through 19 years of age, 1979 through 1989. (from th Compressed Mortality File of the National Center for Health Statistics, Centers for Disease Control).

Example of Box Plots on Original Scale



Example of Box Plots on a Logarithmic Scale