UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

- **Instructor:** Ivo Dinov,
  Asst. Prof. of Statistics and Neurology
- **Teaching Assistants:** Chris Barr & Ming Zheng

University of California, Los Angeles, Fall 2004
http://www.stat.ucla.edu/~dinov/courses_students.html

Chapter 2: Tools for Exploring Univariate Data

- Types of variables
- Presentation of data
- Simple plots
- Numerical summaries
- Repeated and grouped data
- Qualitative variables

TABLE 2.1.1 Data on Male Heart Attack Patients

A subset of the data collected at a Hospital is summarized in this table. Each patient has measurements recorded for a number of variables – ID, Ejection factor (ventricular output), blood systolic/diastolic pressure, etc.

- Reading the table
- Which of the measured variables (age, ejection etc.) are useful in predicting how long the patient may live.
- Are there relationships between these predictors?
- variability & noise in the observations hide the message of the data.

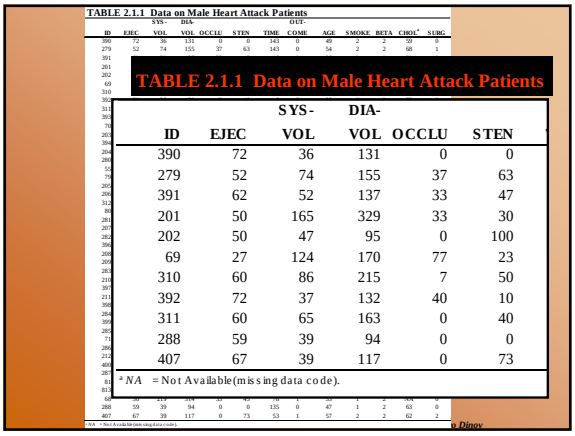


TABLE 2.1.1 Data on Male Heart Attack Patients

ID	EJEC	SYS-VOL	DIA-VOL	OCCLU	STEN
390	72	36	131	0	0
279	52	74	155	37	63
391	62	52	137	33	47
201	50	165	329	33	30
202	50	47	95	0	100
69	27	124	170	77	23
310	60	86	215	7	50
392	72	37	132	40	10
311	60	65	163	0	40
288	59	39	94	0	0
407	67	39	117	0	73

* *NA* = Not Available (missing data code).

Types of variable

- ***Quantitative*** variables are ***measurements*** and counts
  - Variables with ***few*** *repeated values* are treated as ***continuous***.
  - Variables with ***many*** *repeated values* are treated as ***discrete***
- ***Qualitative*** variables (a.k.a. factors or class-variables) describe ***group membership***

Distinguishing between types of variable

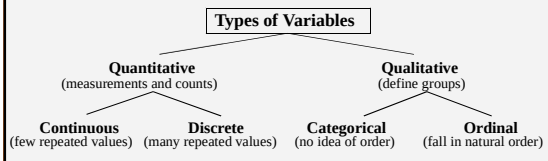


Figure 2.1.1 Tree diagram of types of variable.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Questions …

- What is the difference between quantitative and qualitative variables?
- What is the difference between a discrete variable and a continuous variable?
- Name two ways in which observations on qualitative variables can be stored on a computer. (strings/indexes)
- When would you treat a discrete random variable as though it were a continuous random variable?
  - Can you give an example? ($34.45, bill)

Storing and Reporting Numbers

- Round numbers for presentation
- Maintain complete accuracy in numbers to be used in calculations. If you need to round-off, this should be the very last operation …

Table before simplification

TABLE 2.2.1 Gold Reserves of Gold-Holding IMF Countries

Country	1970	1975	1980	1985	1990
Belgium	42.01	42.17	34.18	34.18	30.23
Canada	22.59	21.95	20.98	20.11	14.76
France	100.91	100.93	81.85	81.85	81.85
Italy	82.48	82.48	66.67	66.67	66.67
Japan	15.22	21.11	24.23	24.33	24.23
Netherlands	51.06	54.33	43.94	43.94	43.94
Switzerland	78.03	83.2	83.28	83.28	83.28
U.K.	38.52	21.03	18.84	19.03	18.94
U.S.A.	316.34	274.71	264.32	262.65	261.91

Units: millions of troy ounces.
Source: *The World Almanac and Book of Facts.*

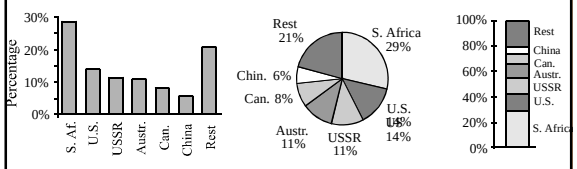
Table after simplification

TABLE 2.2.2 Simplified Table of Gold Reserves of IMF Countries

Country	1970	1975	1980	1985	1990	Average
US	320	270	260	260	260	280
Switzerland	78	83	83	83	83	82
France	100	100	82	82	82	89
Italy	82	82	67	67	67	73
Netherlands	51	54	44	44	44	47
Belgium	42	42	34	34	30	37
Japan	15	21	24	24	24	22
UK	39	21	19	19	19	23
Canada	23	22	21	20	15	20
Average	83	78	71	71	70	

Units: millions of troy ounces.

Different graphs of the same set of numbers



(a) Bar graph (b) Pie chart (c) Segmented bar

Figure 2.6.3 Percentages of the world's gold production in 1991.

n *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Questions …

- For what two purposes are tables of numbers presented? (convey information about trends in the data, detailed analysis)
- When should you round numbers, and when should you preserve full accuracy?
- How should you arrange the numbers you are most interested in comparing? (Arrange numbers you want to compare in columns, not rows. Provide written/verbal summaries/footnotes. Show row/column averages.)
- Should a table be left to tell its own story?

The dot plot

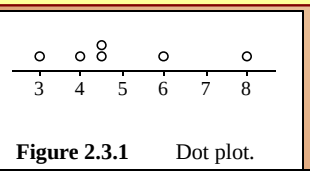


Figure 2.3.1 Dot plot.

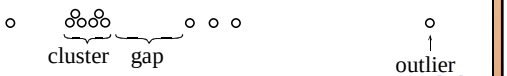


Figure 2.3.2 Dot plot showing special features.

Example of exploiting gaps and clusters

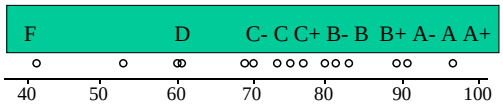


Figure 2.3.3 Grading of a university course.

Scale breaks

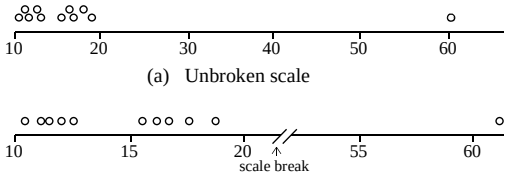


Figure 2.3.4 Dot plot with and without a scale break.

A labeled dot plot

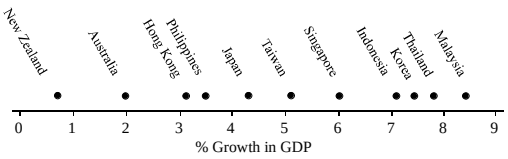


Figure 2.3.5 Forecast of percent growth in GDP for 1990 for some South-East Asian and Pacific countries.

Example of a stem-and-leaf plot

Units: 7 | 2 =72

Stem	Leafs
1	7
2	7 9
3	0 7 7 8 9
4	2 2 3 3 4 5 8 8 9
5	0 0 0 1 1 2 2 3 4 4 7 8 9 9
6	0 0 0 2 3 3 3 3 6 7 7 8
7	2 2

Values 52, 54 and their frequencies

Stem-plot of the 45 obs's of the Ejection variable in the Heart Attack data table.

Traffic death-rates data

TABLE 2.3.1 Traffic Death-Rates (per 100,000 Population) for 30 Countries

17.4 Australia	20.1 Austria	19.9 Belgium	12.5 Bulgaria	15.8 Canada
10.1 Czechoslovakia	13.0 Denmark	11.6 Finland	20.0 France	12.0 E. Germany
13.1 W. Germany	21.1 Greece	5.4 Hong Kong	17.1 Hungary	15.3 Ireland
10.3 Israel	10.4 Japan	26.8 Kuwait	11.3 Netherlands	20.1 New Zealand
10.5 Norway	14.6 Poland	25.6 Portugal	12.6 Singapore	9.8 Sweden
15.7 Switzerland	18.6 United States	12.1 N. Ireland	12.0 Scotland	10.1England & Wales

Data for 1983, 1984 or 1985 depending on the country (prior to reunification of Germany)

Source: Hutchinson [1987, page 3].

Units: 17 | 4 = 17.4 deaths per 100,000

5	4	
6		
7		
8		
9	8	
10	1 1 3 4 5	
11	3 6	
12	0 0 1 5 6	
13	0 1	
14	6	
15	3 7 8	
16		
17	1 4	
18	6	
19	9	
20	0 1 1	
21	1	
22		
23		
24		
25	6	
26	8	

(a)

Round-off

Collapse to

12 stems

Units: 1 | 7 = 17 deaths per 100,000

0	5
0	
0	
1	0 0 0 0 0 1 1
1	2 2 2 2 3 3 3
1	5 5
1	6 6 7 7
1	9
2	0 0 0 0 1
2	
2	
2	6 7

(b)

FIGURE 2.3.7 Two stem-and-leaf plots for the traffic deaths data

TABLE 2.3.2 Coyote Lengths Data (cm)

Females											
93.0	97.0	92.0	101.6	93.0	84.5	102.5	97.8	91.0	98.0	93.5	91.7
90.2	91.5	80.0	86.4	91.4	83.5	88.0	71.0	81.3	88.5	86.5	90.0
84.0	89.5	84.0	85.0	87.0	88.0	86.5	96.0	87.0	93.5	93.5	90.0
85.0	97.0	86.0	73.7								
Males											
97.0	95.0	96.0	91.0	95.0	84.5	88.0	96.0	96.0	87.0	95.0	100.0
101.0	96.0	93.0	92.5	95.0	98.5	88.0	81.3	91.4	88.9	86.4	101.6
83.8	104.1	88.9	92.0	91.0	90.0	85.0	93.5	78.0	100.5	103.0	91.0
105.0	86.0	95.5	86.5	90.5	80.0	80.0					

Coyotes captured in Nova Scotia, Canada. Data courtesy of Dr Vera Eastwood.

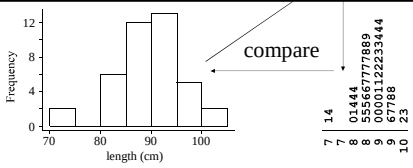
TABLE 2.3.3 Frequency Table for Female Coyote Lengths

Class Interval	Tally	Frequency	Stem-and-leaf plot
70-75	𝍸	2	7 | 1 4
75-80		0	7 |
80-85	𝍸 |	6	8 | 0 1 4 4 4
85-90	𝍸 𝍸 |||	12	8 | 5 5 5 6 6 7 7 7 7 8 8 9
90-95	𝍸 𝍸 |||	13	9 | 0 0 0 0 0 1 1 2 2 2 3 3 4 4 4
95-100	𝍸	5	9 | 6 7 7 8 8
100-105	||	2	10 | 2 3
Total		40	

Body length

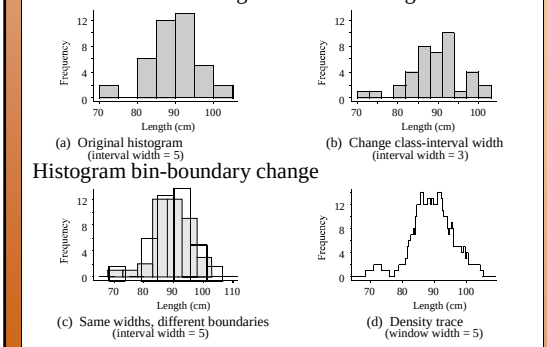
TABLE 2.3.3 Frequency Table for Female Coyote Lengths

Class Interval	Tally	Frequency	Stem-and-leaf plot
70-75	𝍸	2	7 | 1 4
75-80		0	7 |
80-85	𝍸 |	6	8 | 0 1 4 4 4
85-90	𝍸 𝍸 |||	12	8 | 5 5 5 6 6 7 7 7 7 8 8 9
90-95	𝍸 𝍸 |||	13	9 | 0 0 0 0 0 1 1 2 2 2 3 3 4 4 4
95-100	𝍸	5	9 | 6 7 7 8 8
100-105	||	2	10 | 2 3
Total		40	



(a) Histogram (b) Stem-and-leaf plot rotated

Figure 2.3.8 Histogram of the female coyote-lengths data.



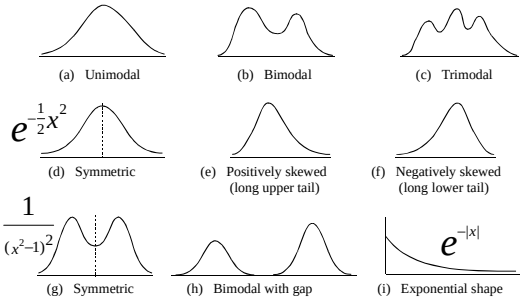
(a) Original histogram (interval width = 5)
(b) Change class-interval width (interval width = 3)
(c) Same widths, different boundaries (interval width = 5)
(d) Density trace (window width = 5)

Figure 2.3.9 Histograms and density trace of female coyote-lengths data.

Questions …

- What advantages does a stem-and-leaf plot have over a histogram? (S&L Plots return info on individual values, quick to produce by hand, provide data sorting mechanisms. But, Hist's are more attractive and more understandable).
- The shape of a histogram can be quite drastically altered by choosing different class-interval boundaries. What type of plot does not have this problem? (density trace) What other factor affects the shape of a histogram? (bin-size)
- What was another reason given for plotting data on a variable, apart from interest in how the data on that variable behaves? (shows features, cluster/gaps, outliers; as well as trends)

Interpreting Stem-plots and Histograms



(a) Unimodal (b) Bimodal (c) Trimodal
(d) Symmetric (e) Positively skewed (long upper tail) (f) Negatively skewed (long lower tail)
(g) Symmetric (h) Bimodal with gap (i) Exponential shape

Interpreting Stem-plots and Histograms

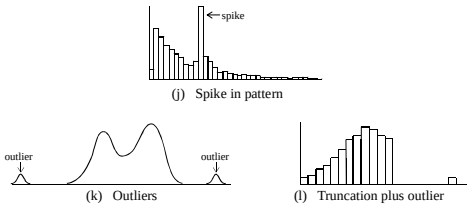
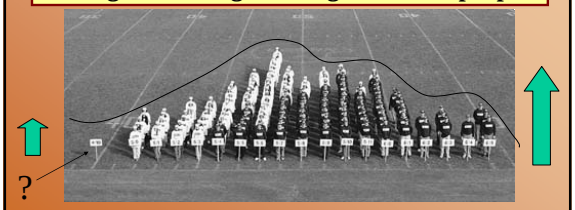


Figure 2.3.10 Features to look for in histograms and stem-and-leaf plots.

Fascinations with histograms – Histogram of heights using the actual people



Subjects are university genetics students, females in white and males in dark tops.

Questions …

- What does it mean for a histogram or stem-and-leaf plot to be bimodal? What do we suspect when we see a bimodal plot?
- What are outliers, and how do they show up in these plots? What should we try to do when we see them?
- What do we mean by symmetry and positive and negative skewness?
- What shape do we call exponential?
- Should we be suspicious of abrupt changes? Why?

  Yes! Try to establish the reason, the jump may have to be rectified!

Descriptive statistics from computer programs like STATA

STATA Output

Standard deviation

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
age	45	50.133	51.000	50.366	6.092	0.908
Variable	**Minimum**	**Maximum**	**Q1**	**Q3**		
age	36.000	59.000	46.500	56.000		

Lower quartile *Upper quartile*

.summarize

Descriptive statistics …

- The sample mean is denoted by $\bar{x}$.

The ***sample mean*** = $\dfrac{\textbf{\textit{Sum}} \text{ of the observations}}{\textbf{\textit{Number}} \text{ of observations}}$

The sample mean is where the dot plot balances

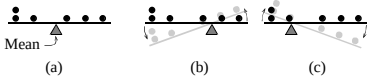


Figure 2.4.1 Mechanical construction representing a dot plot: (a) shows a balanced rod while (b) and (c) show unbalanced rods.

The sample median

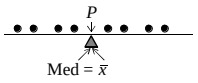
For n observations, $\{x_1, x_2, x_3, \ldots, x_n\}$. Suppose we order the observations min-to-max to get

$$\{x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}\}.$$

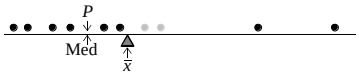
Then the ***sample median*** is the [(n+1)/2]-st largest Observation $X\left(\frac{n+1}{2}\right)$

If $\frac{n+1}{2}$ is not a whole number, the median is the average of the two observations on either side.

Effect of outliers on the mean and median



(a) Data symmetric about *P*

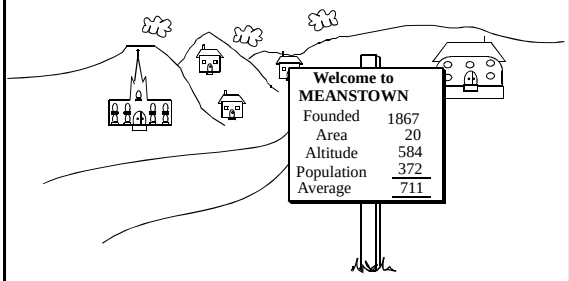


(b) Two largest points moved to the right

Figure 2.4.2 The mean and the median.

[Grey disks in (b) are the ``ghosts" of the points that were moved.]

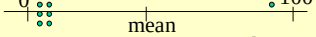
Beware of inappropriate averaging



Suggested by a 1977 cartoon in *The New Yorker* magazine by Dana Fradon.

Questions …

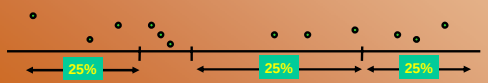
- How is the sample mean related to the dot plot?
- If the index $(n+1)/2$ is not a whole number (e.g., 23.5), how do we obtain the sample median?
- Why is the sample median usually preferred to the sample mean for skewed data? Why is it preferred for "dirty" data?



- Under what circumstances may quoting a single center (be it mean or median) not make sense?(multi-modal)
- What can we say about the sample mean of a qualitative variable? (meaningless)

Quartiles

The first quartile (Q_1) is the median of all the observations whose *position* is strictly below the *position* of the median, and the third quartile (Q_3) is the median of those above.



median

Five number summary

The five-number summery = (Min, Q_1, Med, Q_3, Max)

Inter-quartile Range

$IQR = Q_3 - Q_1$

Box plot compared to dot plot

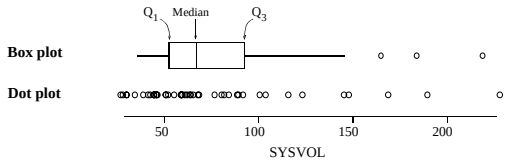


Figure 2.4.3 Box plot for SYSVOL.

Construction of a box plot

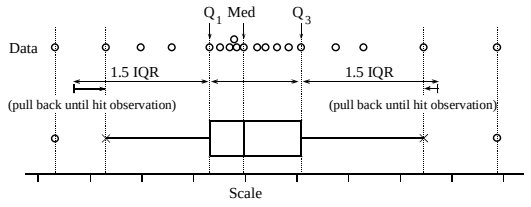


Figure 2.4.4 Construction of a box plot.

Comparing 3 plots of the same data

Stem-and-leaf of strength N = 33
Leaf Unit = 10

1	19 8
5	20 0334
5	20
10	21 00233
(8)	21 55668899
15	22 000111112
6	22 5
5	23 014
2	23
2	24
2	24
2	25 2
1	25 9

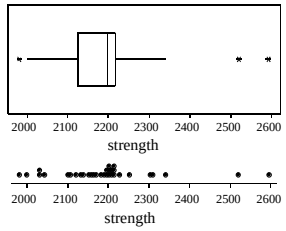


Figure 2.4.5 Three graphs of the breaking-strength data for} gear-teeth in positions 4 & 10 (Minitab output).

Frequency Table

TABLE 2.5.1 Word Lengths for the First 100 Words on a Randomly Chosen Page

3	2	2	4	4	4	3	9	9	3	6	2	3	2	3	4	6	5	3	4
2	3	4	5	2	9	5	8	3	2	4	5	2	4	1	4	2	5	2	5
3	6	9	6	3	2	3	4	4	4	2	2	4	2	3	7	4	2	6	4
2	5	9	2	3	7	11	2	3	6	4	4	7	6	6	10	4	3	5	7
7	7	5	10	3	2	3	9	4	5	5	4	4	3	5	2	5	2	4	2

Frequency Table

Value u_j	1	2	3	4	5	6	7	8	9	10	11
Frequency f_j	1	22	18	22	13	8	6	1	6	2	1

Mean from a frequency table

$\bar{x} = \frac{1}{n}$ Sum of (value $\times$ frequency of occurrence) =

$\frac{1}{n}$ (Sum of all observations)

Example: {2, 4, 2, 4, 2}
Mean = 14/5

Value	Frequency	Value x Frequency
2	3	6
4	2	8
	5	14

TABLE 2.5.2
Frequency Table for the Occurrence of Fish Species in Ocean Strata

No. of strata in which species occur (u_j)	Frequency (No. of species) (f_j)	Percentage of species $(\frac{f_j}{n} \times 100)$	Cumulative Percentage
1	117	35.5	35.5
2	61	18.5	53.9
3	37	11.2	65.2
4	24	7.3	72.4
5	23	7.0	79.4
6	12	3.6	83.0
7	14	4.2	87.3
8	10	3.0	90.3
9	9	2.7	93.0
10+	23	7.0	100.0
	n = 330	100	

Source: Haedrich and Merrett [1988]

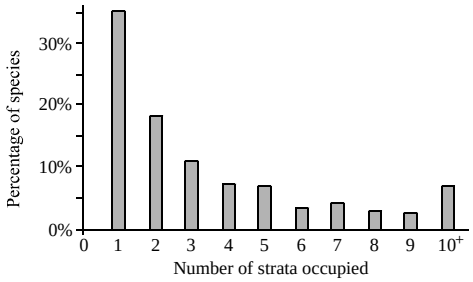
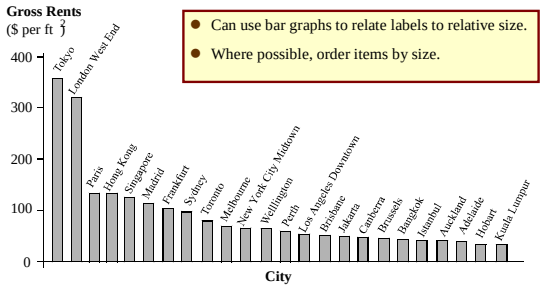


Figure 2.5.1 Bar graph for species data.

Labeled bar graphs to convey size



- Can use bar graphs to relate labels to relative size.
- Where possible, order items by size.

Figure 2.6.2 Cost of commercial rents around the world.