

UCLA STAT 13 Introduction to Statistical Methods for the Life and Health Sciences

- **Instructor:** Ivo Dinov,
Asst. Prof. In Statistics and Neurology
 - **Teaching Assistants:** Sovia Lau, Jason Cheng
UCLA Statistics
- University of California, Los Angeles, Fall 2003
http://www.stat.ucla.edu/~dinov/courses_students.html

STAT 13, UCLA, Ivo Dinov Slide 1

Chapter 2: Tools for Exploring Univariate Data

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

STAT 13, UCLA, Ivo Dinov Slide 2

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.

Slide 3 STAT 13, UCLA, Ivo Dinov

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).

STAT 13, UCLA, Ivo Dinov

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*

Slide 5 STAT 13, UCLA, Ivo Dinov

Distinguishing between types of variable

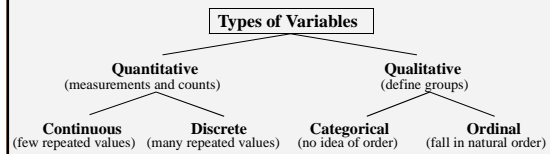


Figure 2.1.1 Tree diagram of types of variable.

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

Slide 6 STAT 13, UCLA, Ivo Dinov

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)

Slide 7 STAT 13, UCLA, Joe Dinger

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 ...

Slide 8 STAT 13, UCLA, Joe Dinger

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*.

Slide 9 STAT 13, UCLA, Joe Dinger

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.

Slide 10 STAT 13, UCLA, Joe Dinger

Different graphs of the same set of numbers

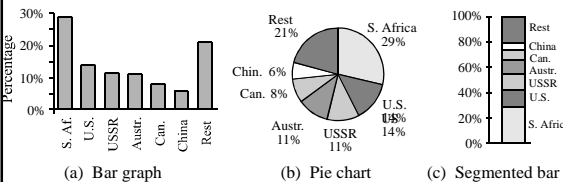


Figure 2.6.3 Percentages of the world's gold production in 1991.

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

Slide 11 STAT 13, UCLA, Joe Dinger

The dot plot

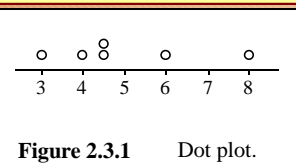


Figure 2.3.1 Dot plot.

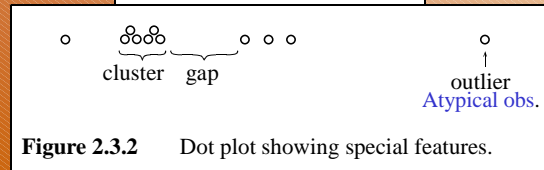
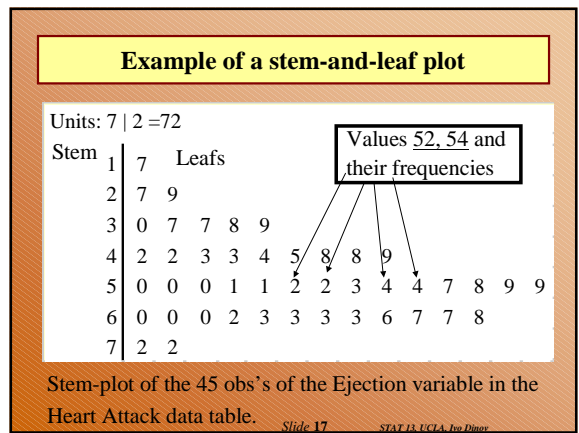
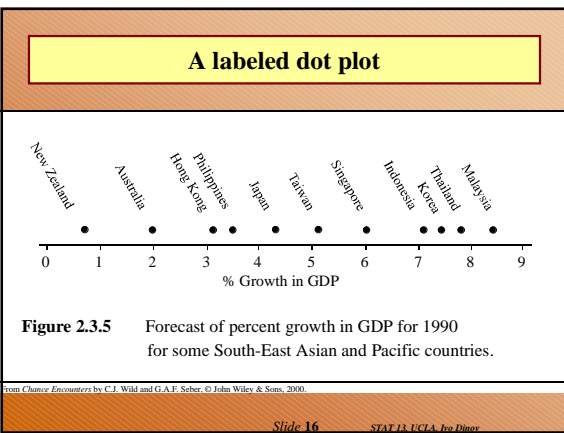
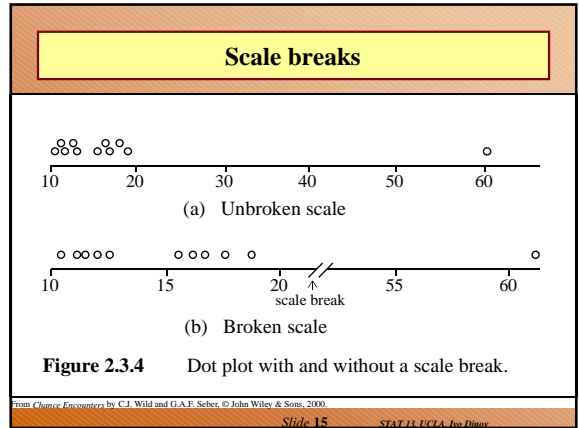
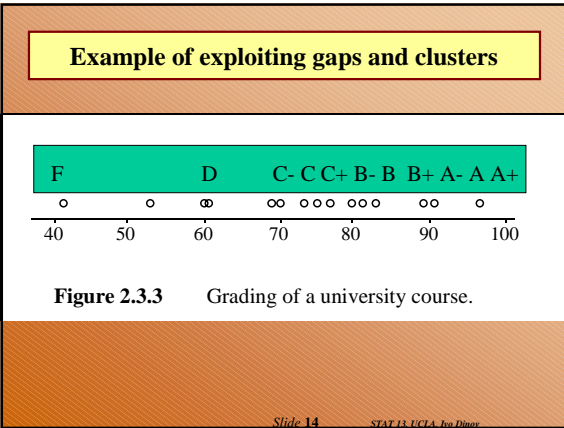


Figure 2.3.2 Dot plot showing special features.

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

Slide 13 STAT 13, UCLA, Joe Dinger



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	10.1 England & Wales

Data for 1983, 1984 or 1985 depending on the country (prior to reunification of Germany)
Source: Hutchinson (1987, page 3).

Slide 18 STAT 13, UCLA, Jon Dineen

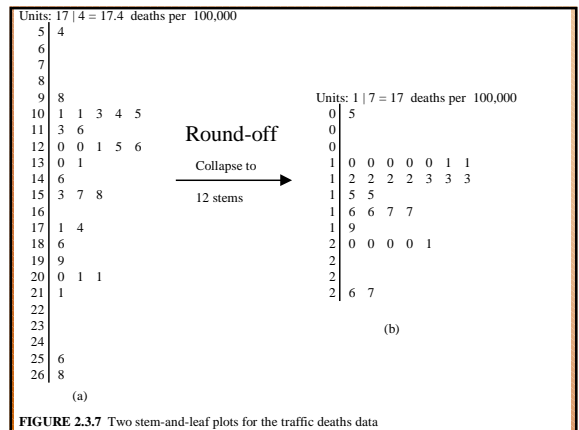


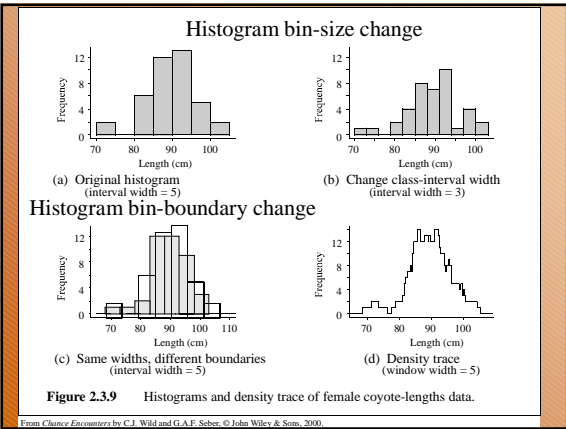
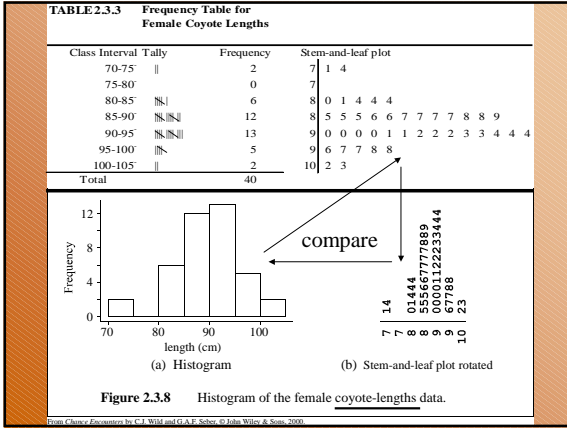
TABLE 2.3.2 Coyote Lengths Data (cm)

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						

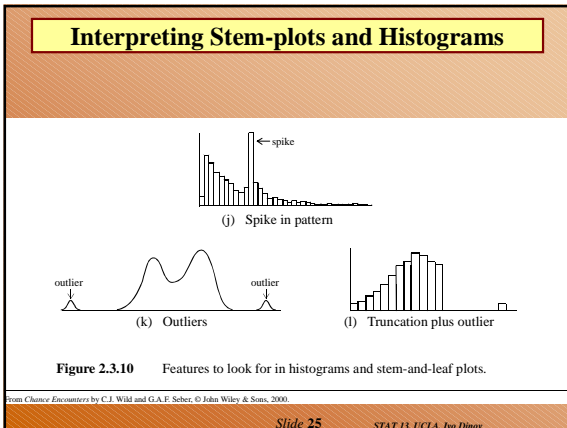
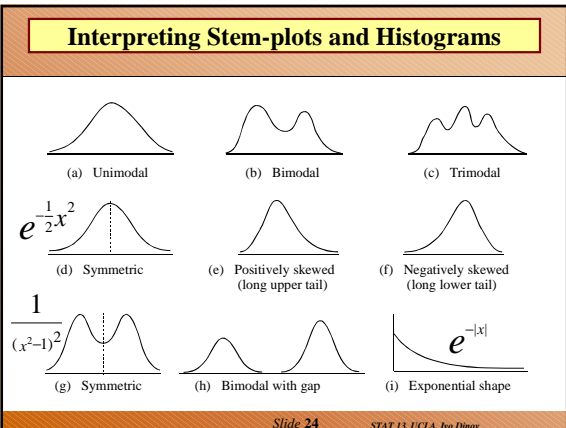
Coyotes captured in Nova Scotia, Canada. Data courtesy of Dr Vera Eastwood.

TABLE 2.3.3 Frequency Table for Female Coyote Lengths

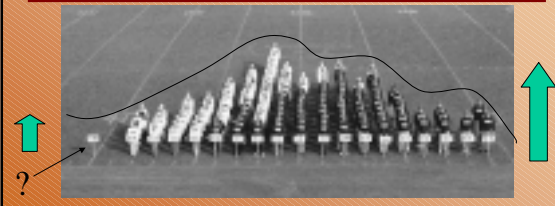
Class Interval	Tally	Frequency	Stem-and-leaf plot
70-75		2	7 1 4
75-80		0	7
80-85		6	8 0 1 4 4 4
85-90		12	8 5 5 5 6 6 7 7 7 7 8 8 9
90-95		13	9 0 0 0 0 1 1 2 2 2 3 3 4 4 4
95-100		5	9 6 7 7 8 8
100-105		2	10 2 3
Total		40	



- 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)



Fascinations with histograms – Histogram of heights using the actual people



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

Slide 26 STAT 13, UCLA, Joe DiMuro

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!

Slide 27 STAT 13, UCLA, Joe DiMuro

Descriptive statistics from computer programs like STATA

STATA Output

Descriptive Statistics		Standard deviation				
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

Slide 28 STAT 13, UCLA, Joe DiMuro

Descriptive statistics ...

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

The *sample mean* = $\frac{\text{Sum of the observations}}{\text{Number of observations}}$

Slide 29 STAT 13, UCLA, Joe DiMuro

The sample mean is where the dot plot balances

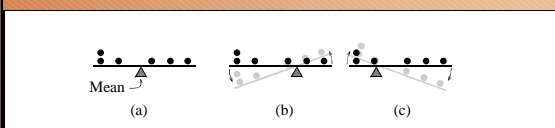


Figure 2.4.1 Mechanical construction representing a dot plot: (a) shows a balanced rod while (b) and (c) show unbalanced rods.

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

Slide 30 STAT 13, UCLA, Joe DiMuro

The sample median

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

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

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.

Slide 31 STAT 13, UCLA, Joe DiMuro

Effect of outliers on the mean and median

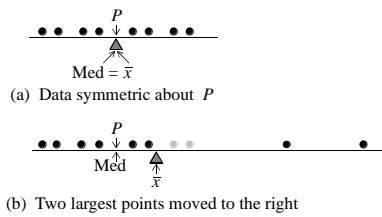
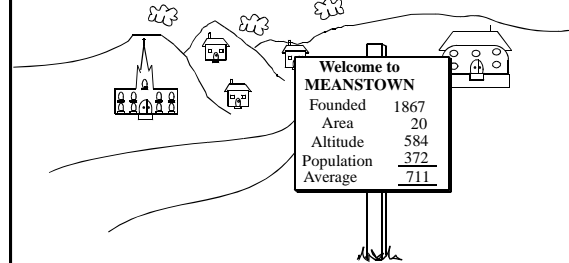


Figure 2.4.2 The mean and the median.
 [Grey disks in (b) are the "ghosts" of the points that were moved.]

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.
 Slide 32 STAT 13, UCLA, Jon Dinger

Beware of inappropriate averaging



Suggested by a 1977 cartoon in *The New Yorker* magazine by Dana Fradon.
 From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 1999.

Five number summary

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

Slide 36 STAT 13, UCLA, Jon Dinger

Inter-quartile Range

$$\text{IQR} = Q_3 - Q_1$$

Slide 37 STAT 13, UCLA, Jon Dinger

Box plot compared to dot plot

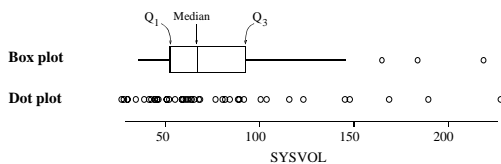


Figure 2.4.3 Box plot for SYSVOL.

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

Slide 38 STAT 13, UCLA, Jon Dinger

Construction of a box plot

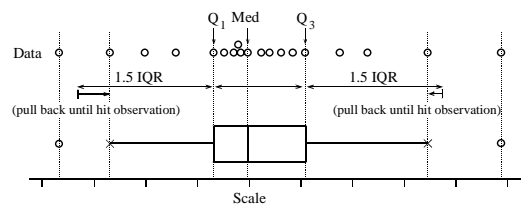


Figure 2.4.4 Construction of a box plot.

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

Slide 39 STAT 13, UCLA, Jon Dinger

Comparing 3 plots of the same data

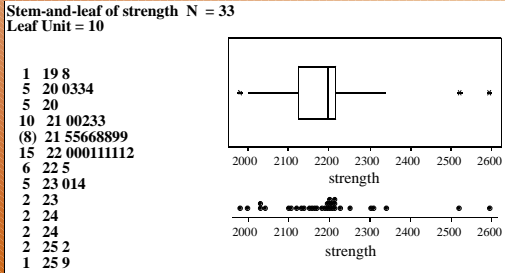


Figure 2.4.5 Three graphs of the breaking-strength data for } gear-teeth in positions 4 & 10 (Minitab output).

Slide 40 STAT 13, UCLA, Joe Dinger

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

Slide 41 STAT 13, UCLA, Joe Dinger

Mean from a frequency table

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

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

Slide 42 STAT 13, UCLA, Joe Dinger