UCLA STAT 110 A
Applied Statistics

•**Instructor:** **Ivo Dinov,**
Asst. Prof. In Statistics and Neurology

•**Teaching Assistants:** Helen Hu, UCLA Statistics

University of California, Los Angeles, Spring 2002
http://www.stat.ucla.edu/~dinov/

Chapter 2: Data Summaries, Plots

- 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.

Data on Male Heart Attack Patients

ID	EJEC	SYS-VOL	DIA-VOL	OCCLU	STEN	TIME	OUT-COME	AGE	SMOKE	BETA	CHOL[a]	SURG
390	72	36	131	0	0	143	0	49	2	2	59	0
279	52	74	155	37	63	143	0	54	2	2	68	1
288	59	39	94	0	0	135	0	47	1	2	63	0
407	67	39	117	0	73	53	1	57	2	2	62	2

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

[a] *NA* = Not Available (missing data code).

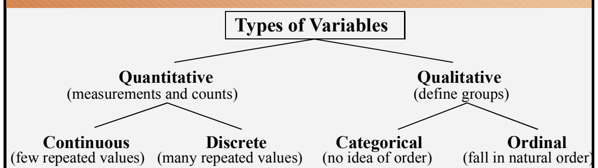
Data & Variables

- ***Variable*** is the name (label) given to the object being measured, counted, observed or recorded in any way. E.g., ID, EjectionVolume, Sys/Dia presure, etc.
- ***Data*** are the actual recording values. E.g., 120/80 (for the arterial pressure).

Types of variable

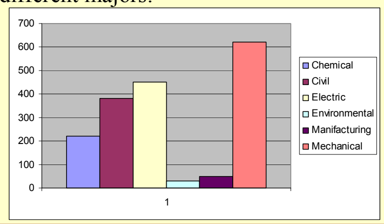
- ***Quantitative*** variables are ***measurements*** and counts
  - Variables with ***few*** *repeated values* are treated as ***continuous***.
  - Variables with ***many*** *repeated values* are treated as ***discrete***
- ***Qualitative*** variables (a.k.a. factors or class-variables) describe ***group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

Distinguishing between types of variable



Bar Chart

- List all possible categories the data is classified in!
  - Represents the frequency of occurrence of the data in each category
- Example: Number of engineering students enrolled in different majors:



220
380
450
30
50
620

Frequency Histograms

- Frequency Histograms - protocol:
  - Determine the RANGE of values **[a : b]**
  - Determine the numbers of bars (bins) to plot; **d**
  - Width of each bar: **(b-a)/d**
  - Count the frequency of your data in each bin, subinterval
  - Draw the histogram
- Example:

Frequency Histograms - Heights

- Example: Gender, Age, Heights (in inches):

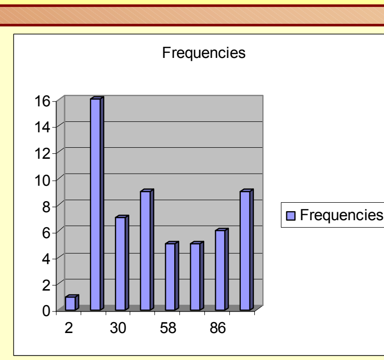
GENDER	AGE	HEIGHT	GENDER	AGE	HEIGHT	GENDER	AGE	HEIGHT
●M	35	95.6	●F	16	77.0	●F	98	128.3
●F	78	112.7	●F	5	64.5	●M	6	63.0
●M	85	122.8	●F	38	91.7	●F	59	103.5
●M	30	91.1	●M	70	107.9	●M	16	77.5
●M	5	68.6	●F	68	112.7	●F	91	127.3
●M	57	93.1	●F	90	127.2	●M	14	69.7
●F	7	63.8	●F	81	121.4	●M	35	91.0
●F	100	144.0	●M	95	130.9	●F	51	102.2
●M	45	97.5	●M	15	73.0	●M	83	126.7
●M	77	120.1	●M	72	112.5	●M	30	86.0
●F	16	70.6	●M	30	89.2	●F	91	120.8
●F	37	93.1	●F	60	108.2	●M	93	125.5
●F	41	93.8	●F	17	79.1	●M	53	102.8
●M	6	67.0	●F	37	91.0	●F	15	72.5
●F	6	65.0	●M	92	140.9	●M	31	91.2
●M	51	104.9	●F	5	58.3	●F	88	114.0
●M	25	86.3	●M	2	43.5	●M	32	97.1
●M	11	73.5	●M	41	90.4	●F	15	87.0
●M	76	108.0	●M	19	76.5	●M	17	67.5
						●M	8	68.0

Frequency Histograms - Heights

- Frequency Histograms - protocol:
  - Determine the RANGE of values **[0 : 130]**
  - Determine the numbers of bars (bins) to plot; **8**
  - Width of each bar: **(130 - 0)/8 = 16.3**
  - Count the frequency of your data in each bin, subinterval
  - Draw the histogram

Frequency Histograms - Heights

- Example:



Density Histograms

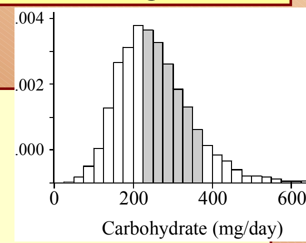
- The **height** of the histogram bar (bin):
  - f_i = frequency of the data
  - on the i-th interval,
  - w_i = width of the i-th interval
  - n – total number or data.

$$h_i = \frac{f_i}{n \times w_i}$$

- Then the **Area** of the histogram bar is: $A_i = \frac{f_i}{n}$
- And the Total area of all histogram bins is: **1 (100%)**

Density (standardized) histograms

For a standardized histogram:



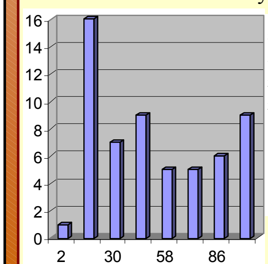
- The vertical scale is *Relative_frequency / Interval_width*
- ***Total area*** under histogram = 1
- ***Proportion*** of the data between *a* and *b* is the ***area*** under histogram between *a* and *b*

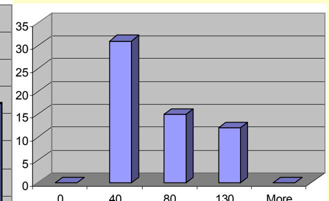
Uni- vs. Multi-modal histograms

- Number of clear humps on the frequency histogram plot determines the modality of a histogram plot.
- Note: Modality of the histogram is histogram parameter specific! Changing the width of the bins changes its appearance!

Uni- vs. Multi-modal histograms

- Number of clear humps on the frequency histogram plot determines the modality of a histogram plot.



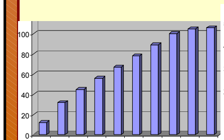


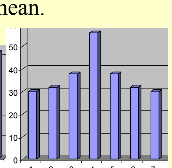
Skewness & Symmetry of histograms

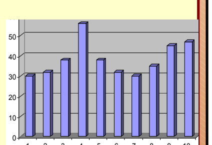
- A histogram is **symmetric** is the bars (bins) to the left of some point (mean) are approximately mirror images of those to the right of the mean.

file:///C:/Ivo.dir/UCLA_Classes/Applets.dir/HistogramApplet.html

- Histogram is **skewed** if it is not symmetric, the histogram is heavy to the left or right, or non-identical on both sides of the mean.







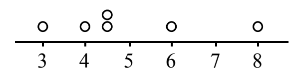
Analyzing Histogram Plots

- Modality – uni- vs. multi-modal (Why do we care?)
- Symmetry – how skewed is the histogram?
- Center of gravity for the Histogram plot – does it make sense?
- If center-of-gravity exists quantify the spread of the frequencies around this point.
- Strange patterns – gaps, atypical frequencies lying away from the center.

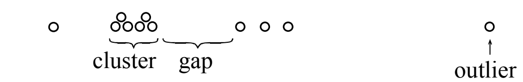
Caution: Storing and Reporting data

- 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 …

The dot plot

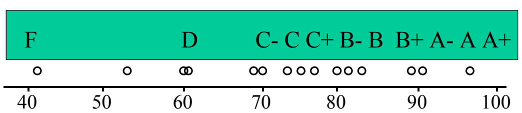


Dot plot.



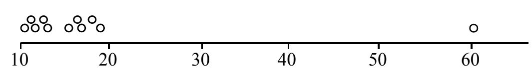
Dot plot showing special features.

Example of exploiting gaps and clusters

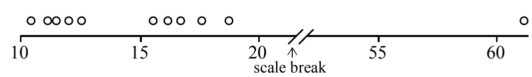


Grading of a university course.

Scale breaks



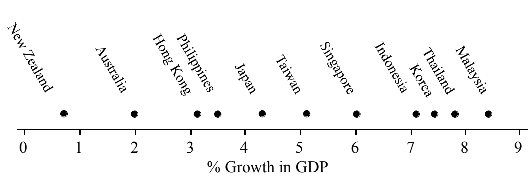
(a) Unbroken scale



(b) Broken scale

Dot plot with and without a scale break.

A labeled dot plot



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

Values 52, 54 and their frequencies

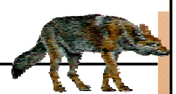
Stem	Leafs
1	7
2	7 9
3	0 7 7 8 9
4	2 2 3 3 4 5 8 8 9
5	0 0 0 1 1 2 2 3 4 4 7 8 9 9
6	0 0 0 2 3 3 3 3 6 7 7 8
7	2 2

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

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.

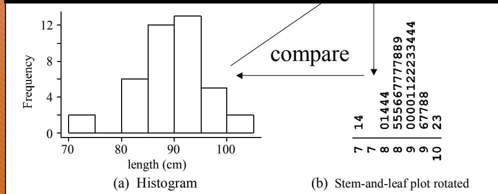


Frequency Table for Female Coyote Lengths

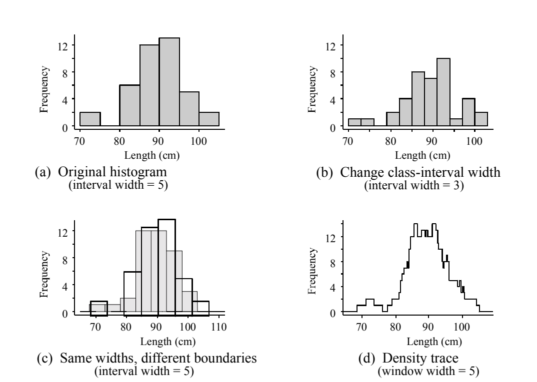
Class Interval	Tally	Frequency	Stem-and-leaf plot
70-75	~~||~~	2	7 | 1 4
75-80		0	7 |
80-85	~~||||~~ |	6	8 | 0 1 4 4 4
85-90	~~||||~~ ~~||||~~ ||	12	8 | 5 5 5 6 6 7 7 7 7 8 8 9
90-95	~~||||~~ ~~||||~~ |||	13	9 | 0 0 0 0 1 1 2 2 2 3 3 4 4 4
95-100	~~||||~~	5	9 | 6 7 7 8 8
100-105	||	2	10 | 2 3
Total	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

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	



Histogram of the female coyote-lengths data.

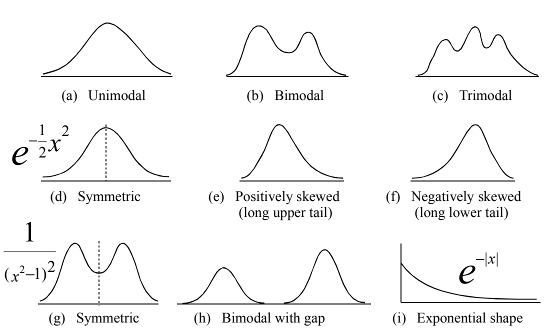


Histograms and density trace of female coyote-lengths data.

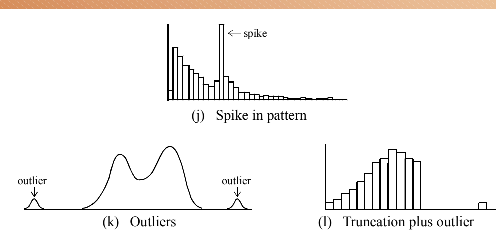
Questions …

- 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



Interpreting Stem-plots and Histograms



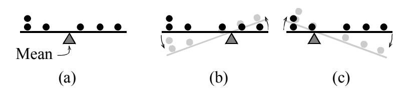
Features to look for in histograms and stem-and-leaf plots.

Descriptive statistics …

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

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

The sample mean is where the dot plot balances



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

The sample median

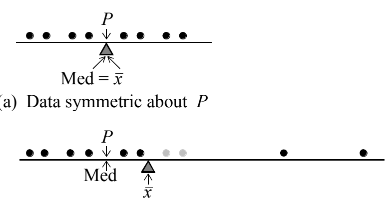
For n observations, {x1, x2, x3, …, xn}. Suppose we order the observations min-to-max to get

{x(1), x(2), x(3), …, 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

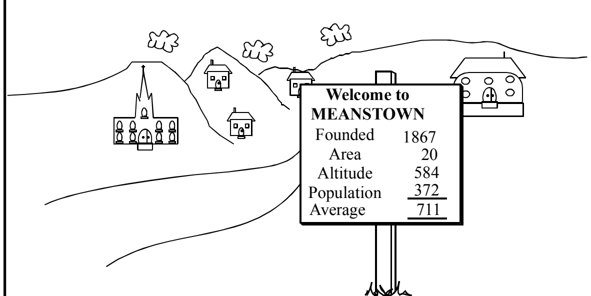


(a) Data symmetric about *P*

(b) Two largest points moved to the right

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.

Quantiles (vs. quartiles)

- The **q^{th} quantile** (100 x q^{th} percentile) is a value, in the range of our data, so that proportion of at least **q** of the data lies at or below it and a proportion of at least **(1-q)** lies at or above it.
- E.x., X={1,2,3,4,5,6,7,8,9,10}. The **20^{th}** percentile (0.2 quartile) is the value **2**, since 20% of the data is below it and 80% above it. The **70^{th}** percentile is the value 7, etc.
- We could have also selected **2.5** and **7.5** for the **20^{th}** and **70^{th}** percentile , above. There is no agreement on the exact definitions of quantiles.

Measures of variability (deviation)

- **Mean Absolute Deviation (MAD)** –

$$MAD = \frac{1}{n-1}\sum_{i=1}^{n}\left|y_i - \overline{y}\right|$$

- **Variance** –

$$Var = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

- **Standard Deviation** –

$$SD = \sqrt{Var} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

Measures of variability (deviation)

- **Example:**
- **Mean Absolute Deviation**– $MAD = \frac{1}{n-1}\sum_{i=1}^{n}\left|y_i - \overline{y}\right|$
- **Variance** – $Var = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$
- **Standard Deviation** – $SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$
- X={1, 2, 3, 4}.

m=2.5

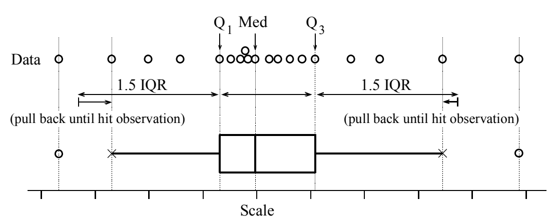
1 2 3 4

MAD=4/3=1.33

Var=5/3=1.67

SD=1.3

Construction of a box-and-whisker plot



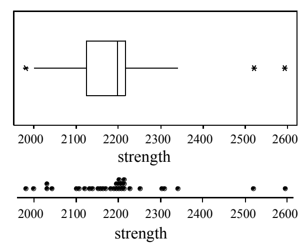
Construction of a box plot.

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

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



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

Frequency Table

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	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

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

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

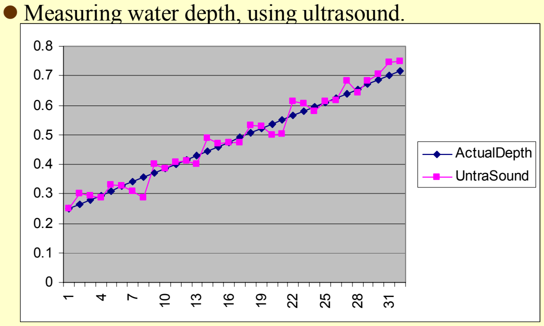
Outliers and Atypical observations

- Outliers – an extremely unrepresentative (of the process) data point
- Example: A paper mill Co. calibrates a device for measuring water depth, using ultrasound, by timing the echos. In calibrating the equipment they ran a simulation of a water tank with known dept. results:

ACTUAL	ULTRASOUND	DIFFERENCE
0.250	0.2505	0.0005
0.265	0.3011	0.0361
0.280	0.2944	0.0144
………		

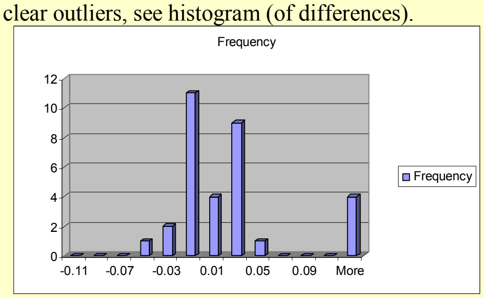
Outliers and Atypical observations

- Measuring water depth, using ultrasound.



Outliers and Atypical observations

- Measuring water depth, using ultrasound. There are 4 clear outliers, see histogram (of differences).

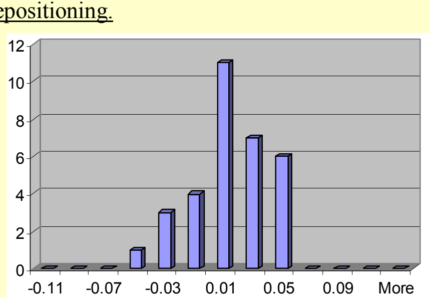


Outliers and Atypical observations

- Measuring water depth, using ultrasound. There are 4 clear outliers, see histogram (of differences).
- One engineer noticed that at depths below 0.3 ft, the radiation pattern of the ultrasound device intersected the wall of the tank, which appeared to have disturbed the measurements.
- They repositioned the ultrasound device to that the path of the sound was completely within the tank. Repeat of the measurements produced a better histogram without the initial 4 outliers.

Outliers and Atypical observations

- Measuring water depth, using ultrasound. After repositioning.



Trimmed, Winsorized means and Resistancy

- A data-driven **parameter estimate** is said to be **resistant** if it does not greatly change in the presence of outliers.
- **K-times trimmed mean**

$$\bar{y}_{tk} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} y_{(i)}$$

Order statistic

- **Winsorized k-times mean**:

$$\bar{y}_{wk} = \frac{1}{n}\left[(k+1)y_{(k+1)} + \sum_{i=k+2}^{n-k-1} y_{(i)} + (k+1)y_{(n-k)}\right]$$