UCLA STAT 10
Introduction to Statistical Reasoning

- **Instructor**: **Ivo Dinov**, Asst. Prof. in Statistics and Neurology
- **Teaching Assistants:** , Yan Xiong and Will Anderson UCLA Statistics

University of California, Los Angeles, Winter 2002

http://www.stat.ucla.edu/~dinov/

Chapter 4

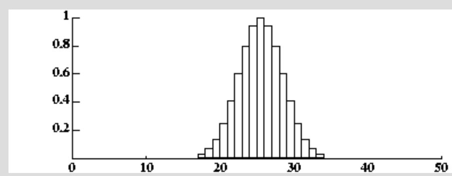
Numerical Summaries – Mean and Standard Deviation

Data representations

- **The histogram of observed data summarizes a large amount of information describing the process we have observed. Often more concise representations are needed.**
  - **Measures of central tendency – average, median, mode.**
  - **Measures of variability – Standard deviation (standard error, root-mean-square), range and quartile and inter-quartile range**
  - **Inter-quartile range**
  - **Energy of the data (sum-squared)**
  - **Etc.**

The average

- If we have to summarize a histogram, or any bar-plot for that matter, in only a few words what would these be?

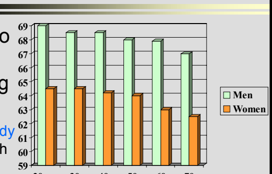


The average

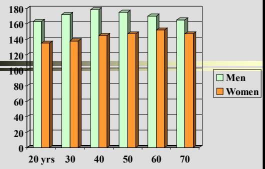
- The average of a list of numbers is their sum divided by how many there are.
  - Example: {9, 1, 2, 2, 0},
    - Average = (9+1+2+2+0)/5 = 14/5 = 2.8
  - In general, $\{a_1, a_2, a_3, \ldots, a_N\}$,
    - Average = $(a_1+a_2+a_3+\ldots+a_N)/N$.

Cross-sectional vs. Longitudinal Studies

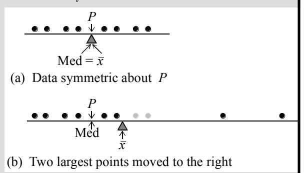
- The avg. height of men appears to decrease with age. Should we conclude the avg. person's getting shorter with time?
  - No, because this is a cross-sectional study – different subjects are compared to each other at <u>one point in time</u>.
  - In longitudinal studies – subjects/units are followed <u>over time and compared with themselves</u>.
  - Note that the people on the 20-30 yrs range are completely different from the folks in the 60-70 yrs of age. There's evidence that with time men may be getting taller – an effect which is heavily confound with the effects of aging.

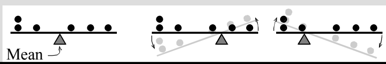


Average vs. Median



- Avg. weight for women 146 lb. Should we expect 50% below and 50% above the average?
  - No, in fact 41% are above and 59% are below the avg.
  - The histogram balances when supported on the average.
  - The median of a histogram is the value in the middle with 50% of the observations above and 50% below the median.



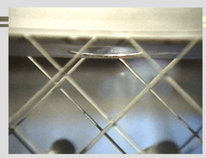


Root Mean Square (R.M.S.)

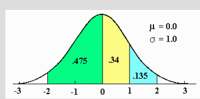
- Consider {0, 5, -8, 7, -3}, the mean is: 0.2. But it's also the mean of {0.1, 0.3, 0, 0.4, 0.2}. Of course, the 2 sequences of 5 numbers are very very different (e.g., size, sign, integer vs. double, etc.) So, the mean does not really represent **all** the info about the data!
- R.M.S. ({a_1, a_2, a_3, …, a_n}) is: $R.M.S. = \sqrt{\frac{1}{N}\sum_{k=1}^{N} a_k^2}$
- Example R.M.S.{0, 5, -8, 7, -3} = 5.4, where as
- R.M.S.{0.1, 0.3, 0, 0.4, 0.2} = 0.24494897.

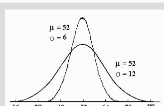
Standard Deviation (SD)

Normal Generation Movie, Quincunx



- The standard deviation is a measure of the spread of the data around its average. Most numbers in the data will be within *1 SD away* from the average, and very few will be *2 SD's*, or more, away from the average.
- With the women's height example we saw, 6,566 women ages 18-74 were surveyed, avg. height was 63.5 *in* and the SD was 2.5 *in*.
- *Rule of thumb* for data spreading:
  - Roughly 68% of all numbers from a list are within 1 SD of the average, and the other ~32% will be farther away. About 95% of the values will be within 2 SD's away from the average.





Calculating the Standard Deviation

- SD = (almost) R.M.S. deviation from the average.
  - Let {a_1, a_2, a_3, …, a_N} are the observed values, then:

$$SD(\{a_1, a_2, \ldots, a_N\}) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(a_k-\mu)^2}$$

  - Where the average (mean) $\mu = \frac{1}{N}\sum_{k=1}^{N} a_k$
  - Example, {20, 10, 15, 15}, $\mu = \frac{1}{4}(20+10+15+15) = 15$

$$SD = \sqrt{\frac{1}{4-1}\left[(20-15)^2+(10-15)^2+(15-15)^2+(15-15)^2\right]} = \sqrt{\frac{1}{3}(25+25)} = \sqrt{\frac{50}{3}} \approx 4.1$$

Calculating the Standard Deviation

- SD = (almost) R.M.S. deviation from the average.
  - Let {a1, a2, …, a_N} are the observed values, then:

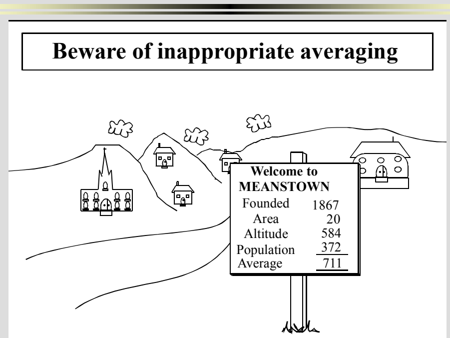
$$SD(\{a_1, a_2, \ldots, a_N\}) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(a_k-\mu)^2}$$

$$\mu = \frac{1}{N}\sum_{k=1}^{N} a_k$$

Note the difference between Our and the textbook definition of SD, see Ch. 26.

$$\widehat{SD}(\{a_1, a_2, \ldots, a_N\}) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(a_k-\mu)^2}$$

Be careful in computing various data descriptors

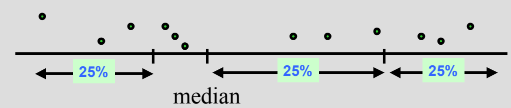


Inter-quartile Range (IQR)

We talked about this earlier
At the end of Ch. 01
→ Chapter 5

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.



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

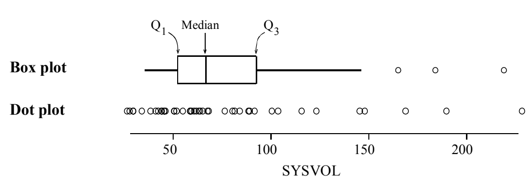


Figure 2.4.3 Box plot for SYSVOL.

Construction of a box plot

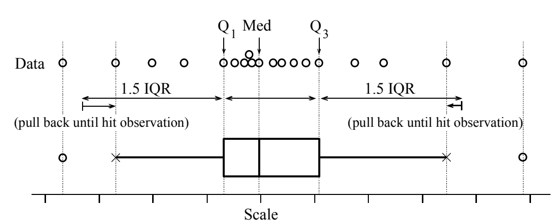
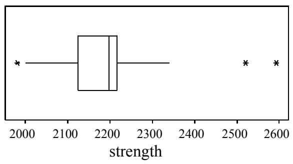


Figure 2.4.4 Construction of a box plot.

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
```



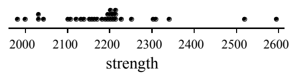


Figure 2.4.5 Three graphs of the breaking-strength data for} gear-teeth in positions 4 & 10 (Minitab output).

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	1	2	3	4	5	6	7	8	9	10	11
Frequency f	1	22	18	22	13	8	6	1	6	2	1

Mean from a frequency table

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

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

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]

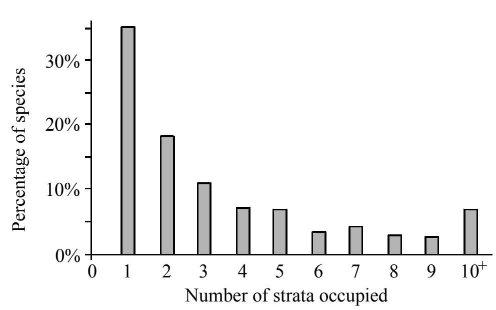


Figure 2.5.1 Bar graph for species data.