Lecture 16

- Today we start Statistics (Chapter 6).
- Specifically we will start with descriptive statistics which consists of techniques for organizing and summarizing data in ways which facilitate its interpretation and analysis.

- we work with sets of real #'s which correspond to data i.e. measurements, observations...
  - ↳ eg actual volumes of water bottles at a water factory.

- In real life, it is often impractical to study a complete data set.
  - ↳ ex: it would be impractical to measure the precise volume of water in every bottle that came off the line.

- We aim to study samples i.e. some finite sets $\{x_1, \dots, x_n\}$ of real numbers chosen from a larger population.

- Goal: Given a sample $\{x_1, \dots, x_n\}$ infer info about the whole population, eg mean, variance, etc.

- Assumption: Given a sample $\{x_1, \dots, x_n\}$, we assume x_i is a value of a random var X_i, where the X_i

are independent with the same distribution

Terminology: The X_i are said to be iid (independent & identically distributed) [very common terminology]..

What kinds of information can we extract from data?

measures of Location

Given a sample $\{x_1, \dots, x_n\}$, we have:

Sample mean: $\bar{x} = \dfrac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

↳ ex: $\{2, 4, 3, 5\}$ $\bar{x} = \dfrac{2+4+3+5}{4} = 3.5$.

Sample median: - The "middle" value of your sample when the sample is arranged by size.

- If the sample has even size, the median is the average of the two middle values.

↳ ex: $\{\underbrace{1, 2, 3}, \underset{\substack{\uparrow \\ \text{middle}}}{4}, \underbrace{5, 6, 7}\}$ has median 4.

$\{\underbrace{1, 2, 3}, \underbrace{4, 5}_{\text{middle}}, \underbrace{6, 7, 8}\}$ has median $\dfrac{4+5}{2} = 4.5$

- mode: The number(s) which appears most often.

Warning the mode is not unique!!

Ex: - $\{1, 3, 1, 2, 1, 5\}$ has mode 1 since it appears 3 times.

- $\{2, 3, 2, 3, 2, 3, 5\}$
  ↳ has mode $\{2, 3\}$.
- $\{1, 2, 3, 4, 5\}$
  ↳ has mode $\{1, 2, 3, 4, 5\}$
  (since every # appears exactly once)

Measures of Variation

Given $\{x_1, \dots, x_n\}$ a sample,

Sample variance: $$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

[Note that the denominator is $n-1$, not n.
This is due to a phenomenon called "estimator bias"]

- S^2 is often tedious to compute, though there is a shortcut:

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

$$= \frac{1}{n-1} \Big(\sum_{i=1}^{n} x_i^2 + \underbrace{n\bar{x}^2 - 2\bar{x} \sum_{i=1}^{n} x_i}_{-n\bar{x}^2} \Big).$$

$$\underbrace{S^2 = \frac{1}{n-1} \Big(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \Big)}_{\text{shortcut!!}}$$

The sample standard deviation is $\sqrt{S^2} = S$.

We also have the range:

sample range $= \max_i \{x_i\} - \min_i \{x_i\}$

which is a very rough measure of spread.

<u>Example:</u> Here is a sample of data from midterm #1

We use Minitab to quickly give a description of the data

Graphs | Probability | Summary | One Sample | Two Sample | Resampling | ANOVA | Regression | Tables | Control Charts | Time Series

Make Similar

Descriptive Statistics: Sample of Test 1 scores

Statistics

Variable	N	N*	Mean	~~SE Mean~~	StDev $\sqrt{\sigma^2}$	Minimum	Q1	Median	Q3	Maximum
Sample of Test 1 scores	20	0	75.000	~~3.311~~	14.808	33.330	66.670	75.000	88.890	94.440

range $= 94.440 - 33.33 = 61.11$

mode $= 88.89$.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
	Sample of Test 1 scores											
1	88.89											
2	94.44											
3	88.89											
4	88.89											
5	77.78											
6	61.11											
7	72.22											
8	72.22											
9	77.78											
10	66.67											
11	66.67											
12	88.89											
13	55.56											
14	72.22											
15	66.67											
16	77.78											
17	33.33											
18	94.44											
19	83.33											
20	72.22											
21												
22												

Stem and Leaf Plots

Stem and leaf plots give a way to visualize data sets that aren't too big, and such that for each $x_i \in \{x_1, \dots, x_n\}$, x_i consists of two or more digits. To construct a stem-and-leaf diagram, follow the following steps:

1) divide each number x_i into two parts:
   - a stem consisting of one or more of the leading digits.
   - a leaf; consisting of the remaining digits.
2) list the stem values in a vertical column.
3) record the leaf for each observation beside its stem
4) write units for stems & leaves.

Ex: Suppose our data set is

$\{9.10,\ 9.20,\ 9.30,\ 10.40,\ 10.20,\ 9.10,\ 8.60,\ 8.00\}$.

Then our stem-~~and~~-leaf diagram becomes.

Stem	Leafs.
8	00, 60
9	10, 10, 20, 30
10	20, 40

We can also use minitab to construct stem & leaf plots. Here is one generated by the last data set after rounding:

Graphs Probability Summary One Sample Two Sample Resampling ANOVA Regression Tables Control Charts Time Series

Make Similar

Stem-and-Leaf Display: Sample of Test 1 scores rounded

Stem-and-leaf of Sample of Test 1 scores rounded, N = 20
Leaf Unit = 1

1	3	3
1	4	
2	5	6
6	6	1777
(7)	7	2222888
7	8	39999
2	9	44

count stem = 10's leafs = 1's

The "count" column in minitab indicates the median with a parentheses (7).

The counts are cumulative above and below the median, so, for example, there is 1 value starting with a 3, 1 values starting with 3 or 4, 2 values starting with 3, 4, or 5, etc.

Then 7 values starting with 8 or 9, 2 values starting with 9.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
	Sample of Test 1 scores rounded											
1	89											
2	94											
3	89											
4	89											
5	78											
6	61											
7	72											
8	72											
9	78											
10	67											
11	67											
12	89											
13	56											
14	72											
15	67											
16	78											
17	33											
18	94											
19	83											
20	72											

we can also refine our stem-and leaf plots by dividing stems. For example

Stem	Leaf
8 L	00,
8 U	60,
9 L	10, 10
9 U	20, 30
10 L	20
10 U	40

(U = upper
L = lower)