
Haijun Li

lih@math.wsu.edu
Department of Mathematics and Statistics
Washington State University

Week 1
Outline

1. Course Policy
2. Ch 1 Review: Overview and Descriptive Statistics
3. Ch 1 Review: Multivariate Data Summary
4. Section 2.1: Sample Space and Events
Course Policy: Stat 360, Fall 2017

- Course information and content: https://learn.wsu.edu/webapps/login/
- My office hours: MWF 12:00-1:00pm, room 405.
- Textbook: Probability and Statistics for Engineering and the Sciences (8th or 9th edition) by J. L. Devore
- 10 weekly homeworks: Late submissions of homeworks will not be accepted. One worst HW will be dropped.
- Three exams: Exam 1 (September 22), Exam 2 (October 30), and Exam 3 (December 13). No compensation for missed exams will be considered unless prior approved arrangements have been made.
All three exams are closed book, closed notes exams, and only calculators are allowed. In Exams #2 and #3, a formula sheet (with regular size) is allowed.

Weekly homework 20 % + three exams 25 % each =95%. The remaining 5% credits will be for attendance.

Grading: 90 % = A range (A-, A), 80 % = B range (B-, B, B+), 70 % = C range (C-, C, C+), 60 % = D range (D, D+), 60 %- = F.
Probabilistic Modeling \( \iff \) Statistical Methods

- **Probabilistic Modeling**: Establish a mathematical model from physical laws that involves random variables with some unknown parameters.
- **Statistical Inference**: Estimate parameters and make decisions on product designs based on system data (so called “random sample”).

![Diagram showing physical laws, product designs, population, sample, and types of reasoning](image)

- **Physical laws** → **Types of reasoning** → **Product designs**
- **Population** → **Statistical inference** → **Sample**

"Figure 1-5 Enumerative versus analytic study."
Population \( \rightsquigarrow \) Random Sample

- **Population** = A finite set of all items under consideration.
- **Random Sample** = A small portion of the population selected at random.

---

\[ x_1, x_2, \ldots, x_n \]

**Enumerative study**

**Analytic study**

---

Future population?

---

Time
Univariate Data Summary

- **Sample of observations**: $x_1, x_2, \ldots, x_n$.
- **Sample mean**:
  \[
  \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n},
  \]
  which describes the central tendency in the data and provides a reasonable estimate of the population mean $\mu$.
- **Example**: Consider, $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, $x_8 = 13.1$.
  
  We have $\bar{x} = (12.6 + \cdots + 13.1)/8 = 13.0$. 

![Diagram of a data distribution with a mean of 13]
Deviations from Sample Mean

Figure: Deviations from $\bar{x}$
Variability in Random Sample

- **Sample variance:**

\[
S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad (= \text{total squared deviations}),
\]

\[
s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}, \quad \text{and} \quad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}.
\]

Here \( s \) is called the **sample standard deviation (SD)**. The sample variance (SD) describes the variability or spread in the data and can be used to estimate the population variance \( \sigma^2 \) (SD).

- Note that \( n - 1 \) (degrees of freedom) is used to provide a so-called “unbiased” estimate.
Table 6-1 displays the quantities needed for calculating the sample variance and sample standard deviation for the pull-off force data. These data are plotted in Fig. 6-2. The numerator of is so the sample variance is and the sample standard deviation is

Computation of $s^2$

The computation of requires calculation of, $n$ subtractions, and $n$ squaring and adding operations. If the original observations or the deviations are not integers, the deviations may be tedious to work with, and several decimals may have to be carried to ensure

$$xi - \bar{x}$$

Figure: Computation of $s$, with $\bar{x} = 13$

$$s^2 = \frac{1.60}{8-1} = 0.2286 \text{ and } s = \sqrt{0.2286} = 0.48.$$
A Shortcut Formula

Sample: $x_1, x_2, \ldots, x_n$.  
- First sample moment (= sample mean): $\hat{\mu}_1 = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.  
- Second sample moment: $\hat{\mu}_2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$.  
- Simple calculations lead to 

$$s^2 = \frac{n}{n-1} \left[ \frac{\sum_{i=1}^{n} x_i^2}{n} - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right)^2 \right] = \frac{n}{n-1} (\hat{\mu}_2 - \hat{\mu}_1^2).$$
Sample of Observations: \( x_1, x_2, \ldots, x_n \).

- The histogram of a data set displays the frequencies of the observations.
- The horizontal axis represents the measurement scale for the data.
- Divide the range of the data into intervals (class intervals, bin) in the horizontal axis.
- Count the number of observations (bin frequency) in each class interval.
- Draw a rectangular box over the class interval.

\[
\text{Rectangle height} = \frac{\text{bin frequency}}{\text{bin width}}.
\]
Power companies need information about customer usage to obtain accurate forecasts of demands. Wisconsin Power and Light determined energy consumption (BTUs) during a particular period for a sample of $n = 90$ gas-heated homes:

| 2.97 | 4.00 | 5.20 | 5.56 | 5.94 | 5.98 | 6.35 | 6.62 | 6.72 | 6.78 |
| 6.80 | 6.85 | 6.94 | 7.15 | 7.16 | 7.23 | 7.29 | 7.62 | 7.62 | 7.69 |
| 7.73 | 7.87 | 7.93 | 8.00 | 8.26 | 8.29 | 8.37 | 8.47 | 8.54 | 8.58 |
| 10.28 | 10.30 | 10.35 | 10.36 | 10.40 | 10.49 | 10.50 | 10.64 | 10.95 | 11.09 |
| 11.12 | 11.21 | 11.29 | 11.43 | 11.62 | 11.70 | 11.70 | 12.16 | 12.19 | 12.28 |
Histogram of Energy Consumption Data

**Figure:** Histogram with 9 Bins, Bin Width = 2 BTUs
Quartiles

The histogram only provides general visual impressions (average, spread and shape) about a data set.

- Important features of a data set: center, spread, departure from symmetry, outliers.

- Sample median $\tilde{x} = \text{cutoff value that divides the data into two equal parts, half below the median and half above.}$

- The first quartile or lower fourth ($Q_1$) = value that has approximately 25% of the observations below it and approximately 75% of the observations above.

- The second quartile or sample median ($Q_2 = \tilde{x}$).

- The third quartile or upper fourth ($Q_3$) = value that has approximately 75% of the observations below it and approximately 25% of the observations above.

- Interquartile range (IQR) or fourth spread $f_s = Q_3 - Q_1$. 
Example

- Data: 34, 47, 1, 15, 57, 24, 20, 11, 19, 50, 28, 37.
- Ordered data: 1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57.

Figure: \[ \text{IQR} = f_s = Q_3 - Q_1 = 42 - 17 = 25 \]
Another Graphical Display: Box Plot

- Idea: Most data points should be within the interquartile range (IQR) with a possible spread of ± 1.5 IQR.
- A box plot displays the three quartiles, the minimum, and the maximum of the data on a rectangular box, aligned either horizontally or vertically.

![Box Plot Diagram]

Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile

First quartile Second quartile Third quartile

Whisker extends to largest data point within 1.5 interquartile ranges from third quartile

Outliers

R 1.5 IQR | IQR | 1.5 IQR

Outliers

Extreme outlier
Because there are many observations, constructing a dot diagram of these data would be relatively inefficient; more effective displays are available for large data sets. A stem-and-leaf diagram is a good way to obtain an informative visual display of a data set where each number consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps.

1. Divide each number \( x_i \) into two parts: a stem, consisting of one or more of the leading digits and a leaf, consisting of the remaining digit.
2. List the stem values in a vertical column.
3. Record the leaf for each observation beside its stem.
4. Write the units for stems and leaves on the display.

Steps for Constructing a Stem-and-Leaf Diagram

To illustrate, if the data consist of percent defective information between 0 and 100 on lots of semiconductor wafers, we can divide the value 76 into the stem 7 and the leaf 6. In general, we should choose relatively few stems in comparison with the number of observations. It is usually best to choose between 5 and 20 stems.

**EXAMPLE 6-4**

To illustrate the construction of a stem-and-leaf diagram, consider the alloy compressive strength data in Table 6-2. We will select as stem values the numbers The resulting stem-and-leaf diagram is presented in Fig. 6-4. The last column in the diagram is a frequency count of the number of leaves associated with each stem. Inspection of this display immediately reveals that most of the compressive strengths lie between 110 and 200 psi and that a central value is somewhere between 150 and 160 psi. Furthermore, the strengths are distributed approximately symmetrically about the central value. The stem-and-leaf diagram enables us to determine quickly some important features of the data that were not immediately obvious in the original display in Table 6-2.

In some data sets, it may be desirable to provide more classes or stems. One way to do this would be to modify the original stems as follows: Divide the stem 5 (say) into two new stems, 5L and 5U. The stem 5L has leaves 0, 1, 2, 3, and 4, and stem 5U has leaves 5, 6, 7, 8, and 9. This will double the number of original stems. We could increase the number of original stems by four by defining five new stems: 5z with leaves 0 and 1, 5t (for twos and threes) with leaves 2 and 3, 5f (for fours and fives) with leaves 4 and 5, 5s (for sixes and sevens) with leaves 6 and 7, and 5e with leaves 8 and 9.
**Example: Compressive Strength Data**

- Median $Q_2 = (160 + 163)/2 = 161.5$ (between the 40th and 41st values of ordered strength).
- $Q_1 = 143.75$ (between the 20th and 21st ordered values).
- $Q_3 = 181.00$ (between the 60th and 61st ordered values).

\[
\begin{align*}
\text{IQR} &= Q_3 - Q_1 = 181 - 143.75 = 37.25
\end{align*}
\]
A sample of total nitrogen loads $n = 57$ (kg N/day) from a particular Chesapeake Bay location:

**Figure:** The nitrogen load data showing mild and extreme outliers
Figure: Left: Houses had no recorded cases of childhood cancer. Right: Houses had recorded cases of childhood cancer.
Multivariate Data

The $n$ pairs of observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. 
HFR Index VS. S&P 500 (1/1990-9/2008 $n = 500$ monthly returns, AllAboutAlpha.com)

HFRI = index designed to reflect hedge fund performance.
Sample Space

- **Sample space** $\Omega := \{\text{all possible outcomes } \omega \text{ of the underlying experiment}\}$.

**Example:** Toss a coin three times. What is the sample space?
Let $H = \text{head}$, $T = \text{tail}$. Then

$$\Omega = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}.$$ 

**Example:** Toss a coin until the first head appears.

$$\Omega = \{H, TH, TTH, TTTTH, TTTTTH, \ldots, \underbrace{T \ldots T}_{k \geq 0}H, \ldots \}.$$
Events

Definition

Any subset $E$ of a sample space $\Omega$ is called an event, and this is denoted as $E \subseteq \Omega$.

- Impossible event: Empty set $\emptyset \subseteq \Omega$.
- Sure event: Sample space $\Omega$ ($\subseteq \Omega$) itself.
Composite Events

**Example:** Toss a coin three times. Describe the following events.

1. $E_1$: three heads occur.
   
   $E_1 = \{HHH\}$.

2. $E_2$: exactly two heads occur.

   $E_2 = \{THH, HTH, HHT\}$.

3. $E_3$: at least one tail occurs.

   $E_3 = \{TTT, TTH, THT, HTT, THH, HHT, HTH\}$.

**Composite Events via Set Operations**

- **Union** $E_1 \cup E_2$: $E_1$ or $E_2$ occurs.
- **Intersection** $E_1 \cap E_2$: Both $E_1$ and $E_2$ occur.
- **Complement** $E'$: $E$ does not occur.
- **Mutually exclusive:** $E_1 \cap E_2 = \emptyset$. 
Example:

- Toss a coin until the first head appears.

\[\Omega = \{H, TH, TTH, TTTTH, \ldots, \underbrace{T \ldots T}_k H, \ldots\}.\]

- Let $A$ be the event that the odd number of tosses are needed to get the first head; that is,

\[A = \{H, TTH, TTTTH, \ldots, \} \subseteq \Omega.\]

- Let $B$ be the event that there are at most five tosses; that is,

\[B = \{H, TH, TTH, TTTTH, TTTTTH\} \subseteq \Omega.\]

- $A \cap B = \{H, TTH, TTTTH\}$.

- $A'$ describes the event that even number of tosses are needed to get the first head.

- $A$ and $A'$ are mutually exclusive.