1 Review of the last lecture

- Statistics is not perfect! Keep in mind that to use a specific formula, you have to check conditions. More on coming soon.

- Measure of center: sample mean, sample median.

- Measure of variation: sample variance, sample standard deviation (SD).

- Sample mean is a random variable. Population mean is a fixed value.

- **Sample Variance** of n observations. Divided by \( n-1 \) is needed for unbiasedness. We will learn more detail later:
  
  Formula 1 (p64): 
  \[ s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \]
  
  Formula 2 (p67): 
  \[ s^2 = \frac{1}{n-1} \left( \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right) \]

- **Sample Standard Deviation (Sample SD)** (p66): 
  \[ s = \sqrt{\text{sample variance}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \]
Q2) (LEC3 p10) Suppose I forgot my worst and best score. THEREFORE, I added two fake numbers, 300 and 300. Complete the table.

TRUE : 93 114 115 118 120 121 122 134 139 146

TWO FAKES: 114 115 118 120 121 122 134 139 300 300

<table>
<thead>
<tr>
<th>statistic</th>
<th>TRUE DATA</th>
<th>TWO FAKES</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>122.2</td>
<td></td>
</tr>
<tr>
<td>median</td>
<td>120.5</td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>115</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>120.5</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>min</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>max</td>
<td>146</td>
<td></td>
</tr>
</tbody>
</table>

ANS:
mean = (122.2*10 -146-93 + 300+300 )/10 = 158.3

Q1: np=10 * 0.25 =2.5. Use 3rd =118
Q2 (=median): np=10 * 0.50 =5. Use (5th+6th)/2 = (121+122)/2 =121.5
Q3: np=10 * 0.75 =7.5. Use 8th =139.
min=93, max=300
We can guess importance of mean and median. They give us center information about the data. But why do we care about sample SD (or sample variance)? If you know sample SD, what can we tell about the nature of the data set? Sample SD tells you spread of data:

2 Empirical Guidelines for Symmetric Bell-shaped Data Curve(p70)

- Approximately 68% of the data lie within $\bar{x} \pm s$.
- Approximately 95% of the data lie within $\bar{x} \pm 2s$.
- Approximately 99.7% of the data lie within $\bar{x} \pm 3s$.

Simply speaking, if your data is symmetric bell shaped, then your next observation will be very likely fall in $\bar{x} \pm 3s$.

”Empirical Guideline” sounds fancy. It just means ”Based on the observed data”. Hence,

”Empirical Guidelines for Symmetric Bell-shaped Data Curve tell us that approximately 68% of the data lie within $\bar{x} \pm s.$”

is same as

”Based on the observed data for Symmetric Bell-shaped Data Curve, approximately 68% of the data lie within $\bar{x} \pm s.”
Figure 1: Symmetric Bell shape data
Figure 2: Symmetric But Non-Bell shape data. $\bar{x} \pm 2s$ contains all data, (and so does $\bar{x} \pm 3s$, of course).
Figure 3: Population mean=50, population SD=1, 5, 10 from top.
3 [4.2] Probability on an event

- **Experiment** (p138): The process of observing a phenomenon that has variation in its outcomes.

- **Sample space**: The collection of all possible distinct outcomes of the experiment. Denoted by $S$.

- **elementary outcome = simple event = element of the sample space**: Each outcome.

- **event**: the set of elementary outcomes.

  eg) Flip a fair coin once. You will get Head or Tail.
  Experiment: Flip a fair coin once.
  Sample space: $\{Head, Tail\}$.
  Event A: Observe Head
  Event B: Observe Tail
  Event C: Observe Head and Tail at the same time. (This is impossible event but still this is an event. We call it empty event)
  Event D: Observe Head or Tail. (We know this happens all the time. But still this is an event.)

- **Tree diagram** (p139): See the figure there.
• **Discrete sample space** (p140): Sample space is discrete.
  eg) Observe dice roll. (You will observe 1, 2, 3, 4, 5, 6)

• **Continuous sample space**: Sample space is continuous.
  eg) Observe an arbitral real number (You will observe -4.1, 3.14444, 200.1, etc)

• **Probability of an event**: A numerical value that represents the proportion of times the event is expected to occur when the experiment is repeated under identical conditions. The probability of event A is denoted by \( P(A) \).

  - Probability must satisfy 3 rules (p141):
    1. \( 0 \leq P(A) \leq 1 \) for all events A
    2. \( P(A) = \sum_{\text{all } e \in A} P(e) \).
    3. \( P(S) = \sum_{\text{all } e \in S} P(e) = 1 \) (Remember, \( S \) is sample space.)

• Equally likely outcomes and the uniform probability model (p142) (**discrete sample space only!**)
  \[
p(A) = \frac{m}{k} = \frac{\text{Number of elementary outcomes in } A}{\text{Number of elementary outcomes in } S}
  \]
  Remember \( S \) is sample space !.
  eg) Use special dice with value 1, 2, 30, 45 (4 faces). Then \( S = \{1, 2, 30, 45\} \)
Number of elementary outcomes = 4 (=k).

Let $A_i =$ observe number $i$. Then

\[ P(A_1) = \frac{1}{4}, \]
\[ P(A_2) = \frac{1}{4}, \]
\[ P(A_{30}) = \frac{1}{4}, \]
\[ P(A_{45}) = \frac{1}{4}, \]

Let $B = \{1 \text{ or } 45\}$ Then

\[ P(B) = \frac{2}{4} \]

because $k=4$ (no change), and $m=2$.

- $m \times n$ counting rule (p143): When a first task has $m$ possible outcomes and a second ask has $n$ possible outcomes, then there are $m \times n$ outcomes when performing both tasks.

eg) Use special dice with value 1, 2, 30, 45 (4 faces). Throw once. Then after that you flip a coin (H or T). Then we have $4 \times 2$ outcomes:

outcome 1) observe 1 and H
outcome 2) observe 2 and H
outcome 3) observe 30 and H
outcome 4) observe 45 and H
outcome 5) observe 1 and T
outcome 6) observe 2 and T
outcome 7) observe 30 and T
outcome 8) observe 45 and T

\[ P(\text{observe 1 and H}) = \frac{1}{8}. \]
\[ P(\text{observe 1}) = P(\text{observe 1 and H}) + P(\text{observe 1 and T}) = \frac{1}{8} + \frac{1}{8} = \frac{2}{8}. \]

Here I used the rule in p142 because 8 outcomes are \textbf{all equally likely}.  

counter example) Suppose the coin is \textit{unfair}, i.e., both side said ”T”. In this case
\[ P(\text{observe 1 and H}) = 0. \]
Because you will never see H. 8 outcomes are \textbf{NOT all equally likely}. Hence p142 rule can not be applied.

- **Probability as long-run relative frequency** (p147): We can estimate \( P(A) \) by repeating the experiment many times.

4  **[4.3] Event relations and Two laws of Probability**

Suppose you cast a 4-face dice \textit{ONCE}.

Sample space: \( S = [1, 2, 30, 45] \).
Define,
\[ A_1 \equiv \text{event that you observe } 1 \equiv [1]. \]
Note: Some people use \{1\} notation in books.
\[ A_2 = [2]. \]
\[ A_{30} = [30] \]
\[ A_{45} = [45] \]

\[ B_1 \equiv \text{event that you observe 2 or 45 } \equiv [2, 45]. \]
Note: \( B_1 \neq \text{event that you observe 2 AND 45.} \)
\[ B_2 = [1, 2, 30]. \]
\[ B_3 = [1, 2] \]
\[ B_4 = [30, 45] \]
\[ B_5 = [2 \text{ AND } 45] \ (B_5 = \text{event that you observe 2 AND 45}. \text{ We know this does not happen because by definition of this experiment, we cast a dice ONLY ONCE. We know } P(B_5) = 0). \]
• "Complement of $B_1$"
  \[ \equiv \bar{B}_1 = [1, 30]. \]

• "Union of $B_1$ and $B_3$"
  \[ \equiv B_1 \cup B_3 \]
  \[ \equiv B_1 \text{ or } B_3 \]
  \[ = [1, 2, 45] \]
  ( Not $[1, 2, 2, 45]$ )

• "Intersection of $B_1$ and $B_3$"
  \[ \equiv B_1 \cap B_3 \]
  \[ \equiv B_1 \text{ and } B_3 \]
  \[ \equiv B_1 \cap B_3 \]
  \[ = [2] \]
  eg) $B_3 \cap B_4 = empty = \phi$

• Venn diagram (p154). Let’s draw it.
  Now Let’s look at relation between those events and probability.

• Law of the Complement (p155)
  \[ P(A) = 1 - P(\bar{A}) \]

• Addition Law
  \[ P(A \cup B) = P(A) + P(B) - P(AB) \]
• **(A and B are) Mutually Exclusive event**: If A happens, B never happens. AND if B happens, A never happens.

  eg1) Cast 4-face dice, $A_1$ and $A_{45}$ are mutually exclusive event.

  eg2) Cast 4-face dice, $B_3$ and $B_4$ are mutually exclusive event.

  counter eg1) $A_2$ and $B_1$ are NOT mutually exclusive event. Because they have common value 2.

• **Special Addition Law for Mutually Exclusive Events**: If A and B are mutually, exclusive,

  \[ P(A \cup B) = P(A) + P(B). \]

  eg1) Draw Venn diagram for $A_2$ and $B_2$.

  eg2) Draw Venn diagram for $B_1$, $B_2$, and $B_3$.

• Useful set operation:

  \[ T - 1 \] : \( (A \cap B) \cap C = A \cap (B \cap C) \). So people just write \( = A \cap B \cap C \) (This is ambiguous, we don’t know which binary operation should be done first, but \[ T0 \] rule shows it does not matter. So we just write this way.)

  \[ T0 \] : \( (A \cup B) \cup C = A \cup (B \cup C) \). So people just write \( = A \cup B \cup C \)

  \[ T1 \] : \( \overline{A} = A \)
$[T2]: \overline{A \cap B} = A \cup \overline{B}$
$[T3]: \overline{A \cup B} = A \cap \overline{B}$
$[T4]: \overline{A \cap B \cap C} = A \cup \overline{B} \cup \overline{C}$
$[T5]: \overline{A \cup B \cup C} = \overline{A} \cap \overline{B} \cap \overline{C}$
$[T6]: A \cap [B \cup C] = [A \cap B] \cup [A \cap C]$.

Note: LHS (left hand side) does:

step1) $[B \cup C]$,  
step2) $A \cap [B \cup C]$.

RHS (right hand side) does:

step1) $[A \cap B]$  
step2) $[A \cap C]$  
step3) $[A \cap B] \cup [A \cap C]$.

$[T7]: A \cup [B \cap C] = [A \cup B] \cap [A \cup C]$.

eg1) $\overline{A} = ?$

eg2) $\overline{A \cap \overline{B}} = ?$
eg3) \( A \cap [B \cap C] = ? \)

eg4) What the difference between \( A \cap B \) and \( \overline{A} \cap \overline{B} \) ?

5 \textbf{ R for Ch 2 }

Now is the time to learn R. We will ask you very basic questions about R in exams, if any. For example, what is the meaning of ”getwd()” etc. (You already memorized it, right?) If you want to learn R outside of the class, you have to hire a private tutor, (maybe $20/hour) . Here we can learn it for free! All we need is basic usage.