• Chapter 2 Organization and Description of Data

1. Data types: Categorical and numerical
2. Describing the distribution of Data by Tables and Graphs
   (a) Frequency table
   (b) Histogram
   (c) Stem-and-leaf
   (d) Box plot. (Minimum-$Q_1$-median-$Q_2$-Maximum)
3. Measure of center: sample mean and sample median
   – effect of outliers on sample mean and sample median: outliers will draw the sample mean away from the sample median.
4. p-th percentile. Learn the definition of p-th percentile and the method to calculate the p-th percentile.
5. quartiles: $Q_1$ =25-th, $Q_2$=50-th, $Q_3$=75-th percentiles.
6. Measures of variation: sample variance, sample standard deviation, sample interquartile range = $Q_3 - Q_1$
   – effects of outliers on sample variance.

• Chapter 3 Descriptive Study of Bivariate Data

1. contingency table (two-way table) to describe the distribution of bivariate categorical data.
   – Relative frequency.
   – Relative frequency table by rows.
   – Explain the Simpson Paradox using lurking variable, when combining tables.
2. How to design experiment to make comparison. Rule to follow:
   (a) Subjects, or experimental units (e.g, patients), must be assigned to two groups all in such a manner that neither method is favored. Ideally, we like to have groups of equal size.
   (b) Use random selection.
3. Scatter diagram of bivariate numerical data.
4. The correlation coefficient – A measure of linear relation.
   – The correlation coefficient denoted by $r$, is a measure of strength of the linear relation between the $x$ and $y$ variable.
   
   $r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$
- Alternative formula to calculate $r$
- Properties of $r$
  (a) The value of $r$ always between 1 and $-1$.
  (b) The magnitude of $r$ indicates the strength of linear association.
  (c) The sign indicates the direction (negative or positive).
  (d) $r = \pm 1$ the data is on a straight line.
  (e) $r$ close to 0, the linear association is very weak.
  (f) $r$ measures the closeness of the pattern of scatter to a line. Strong curved relationship between $x$ and $y$ doesn’t imply strong linear relationship.
- Explain the spurious correlation by lurking variable.

5. Prediction of one variable from another – linear regression

\[
\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x
\]

slope $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Here $\hat{y}$ is the estimated $y$ value for given $x$ value.

- Chapter 4 Probability

1. Know how to determine the sample space, events and simple event of an experiment.

2. Probability

(a) $0 \leq P(A) \leq 1$

(b) $P(A) = \sum_{\text{all } e \in A} P(e)$

(c) $P(S) = \sum_{\text{all } e \in A} P(e) = 1$

3. Uniform probability model

(a) all simple event have identical probability $= \frac{1}{\text{size of sample space}}$

(b) Random selection.

(c) number of combinations of $x$ out of $n = \binom{n}{x}$
(d) There are 4 green balls and 6 black balls in a basket. A person randomly picks 6 balls from the basket, what is the probability that he has picked exactly 2 green balls and 4 black balls.

size of sample space: the combination of pick 6 balls out of 10 balls is \( \binom{10}{6} \)

size of event \( A = "2 \text{ green balls and 4 black balls}" \) is \( \binom{4}{2} \binom{6}{4} \)

\[
P(A) = \frac{\binom{4}{2} \binom{6}{4}}{\binom{10}{6}}
\]

(e) There 7 balls labeled as \( \{1, 2, ..., 7\} \) in a basket. Randomly pick 5 balls from the basket, what is the probability that No. 2 ball is picked.

4. Operation on sets
(a) \( A \cup \bar{A} = I \)
(b) \( A \cap \bar{A} = \emptyset \)
(c) \( A \cap (B \cup C) = AB \cup AC \)
(d) \( AB \cup A\bar{B} = I, AB \cap A\bar{B} = \emptyset \)

5. Use set operation to represent
(a) \( A \) and \( B \) both happen: \( AB \)
(b) \( A \) and \( B \) not both happen: \( \bar{A}B = A \cup B \)
(c) \( A \) or \( B \) or both happen: \( A \cup B \)
(d) Neither \( A \) nor \( B \) happen: \( A \cup B = \bar{A}\bar{B} \)
(e) \( A \) happens but \( B \) does not happen: \( A\bar{B} \)
(f) Only \( A \) or only \( B \) happens: \( A \cup B - AB \)

6. Probabilities of sets: area of sets.
(a) \( P(A \cup B) = P(A) + P(B) - P(AB) \). If \( AB = \emptyset \) (A, B are incompatible), then \( P(A \cup B) = P(A) + P(B) \)
(b) \( P(AB) = P(A) + P(B) - P(A \cup B) \)
(c) \( P(A) + P(\bar{A}) = 1 \)
(d) \( P(A) = P(AB) + P(A\bar{B}) \)
(e) Using table to calculate probability

<table>
<thead>
<tr>
<th></th>
<th>( A )</th>
<th>( \bar{A} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B )</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>( \bar{B} )</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7</td>
<td>1</td>
</tr>
</tbody>
</table>
7. Conditional Probability

(a) Given $B$ happened, the probability that $A$ also happens is

$$P(A|B) = \frac{P(AB)}{P(B)}$$

(b) $P(AB) = P(A|B)P(B)$

(c) If $A$ and $B$ are independent, $P(A|B) = P(A)$, and hence

$$P(AB) = P(A)P(B)$$

(d) $P(A) = P(AB) + P(A\bar{B}) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$

(e) Bayes Theorem

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})}$$

(f) Using tree structure to calculate conditional probability