August 21, 2018
Instructor and TA’s (weekly consultation location and hours, by appointments only)

Denis HY LEUNG SOE 5047
email: denisleung@smu.edu.sg
phone: 68280396

TA’s

Mon 1230-1530 SOE/SOSS GSR3-2 Aloysius LEE Jie Lun
aloysius.lee.2017@business.smu.edu.sg

Thu 1530-1830 SOE/SOSS GSR3-6 Amos PEE Mun Hsuen
amos.pee.2017@economics.smu.edu.sg

Mon 830-1130 SOE/SOSS GSR3-2 YANG Zhi Yao
zhiyao.yang.2015@accountancy.smu.edu.sg

Wed 1530-1830 SOE/SOSS GSR3-2 Isabel YEOW Shu Yi
isabel.yeow.2017@socsc.smu.edu.sg
Essentials

- Name card (first few weeks)
- Course webpage: http://economics.smu.edu.sg/faculty/profile/9699/Denis%20LEUNG (NOT eLearn!)
- Understanding of basic Calculus and Algebra – Appendix in course notes
- Readings before each class
- Projects vs Homework
- If you missed a class, it is responsibility to find out what you have missed from your classmates or course webpage
- Do not copy down/memorise formulae blindly. Discard as many formulae as possible as you progress
Assessments

- Class Participation (10%)
- Projects (40%)
  - 2 projects with presentation 20% each
  - Each project’s grade includes 8% individual assessment (quizzes)
- Exam (50%)
  - Closed book but one 2-sided A-4 “cheat sheet” is allowed
University of Michigan Panel Study of Income Dynamics on 753 white married women in the US (1975-76):

<table>
<thead>
<tr>
<th>Woman</th>
<th>Workforce status (1=Yes, 0=No)</th>
<th>Hrs worked</th>
<th>#kids &lt; 6 yrs</th>
<th>Age</th>
<th>Education (yrs)</th>
<th>Hourly wage rate</th>
<th>Husband's wage rate</th>
<th>Experience (yrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1610</td>
<td>1</td>
<td>32</td>
<td>12</td>
<td>3.3540</td>
<td>4.0288</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1656</td>
<td>0</td>
<td>30</td>
<td>12</td>
<td>1.3889</td>
<td>8.4416</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1980</td>
<td>1</td>
<td>35</td>
<td>12</td>
<td>4.5455</td>
<td>3.5807</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>456</td>
<td>0</td>
<td>34</td>
<td>12</td>
<td>1.0965</td>
<td>3.5417</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1568</td>
<td>1</td>
<td>31</td>
<td>14</td>
<td>4.5918</td>
<td>10.0000</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2032</td>
<td>0</td>
<td>54</td>
<td>12</td>
<td>4.7421</td>
<td>6.7106</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1440</td>
<td>0</td>
<td>37</td>
<td>16</td>
<td>8.3333</td>
<td>3.4277</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1020</td>
<td>0</td>
<td>54</td>
<td>12</td>
<td>7.8431</td>
<td>2.5485</td>
<td>35</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>750</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>31</td>
<td>12</td>
<td>0.0000</td>
<td>4.8638</td>
<td>14</td>
</tr>
<tr>
<td>751</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>12</td>
<td>0.0000</td>
<td>1.0898</td>
<td>4</td>
</tr>
<tr>
<td>752</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>12</td>
<td>0.0000</td>
<td>12.4400</td>
<td>15</td>
</tr>
<tr>
<td>753</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>9</td>
<td>0.0000</td>
<td>6.0897</td>
<td>12</td>
</tr>
</tbody>
</table>
Sample vs. Population

(a) Data are a **sample** from a **population** that we want to study
e.g., 753 women (sample) out of all white married women in
1975-1976 (population)

(b) We are interested in some characteristics of the **population**
e.g., average wage or percentage of women who earned more
than minimum wage in the **population**

(c) We use a sample to answer our questions about the population

(d) Data = Sample
Data structure and terminologies

- Sample size - \( n \): number of units (observations) in the sample

- Variables - characteristics of the units
  
  *e.g.*, Workforce status, hrs worked, age, wage rate, *etc.*

  - Often represented by symbols, \( X, Y, Z, \text{ etc.} \)
  
  - Quantitative: numeric
    
    *eg.*, Age, wage rate, hrs worked
  
  - Qualitative (Categorical): Not quantitative (no natural ordering)
    
    *eg.*, Gender, colour, race
  
  - Discrete: countable number of values
    
    *eg.*, Gender, \# kids, \# days
  
  - Continuous: uncountably many in a range \( (a, b) \)
    
    *eg.*, Wage rate, age (real, not rounded), temperature

- A sample is \( n \) observations, \( X_1, \ldots, X_n \), of \( X \)
### Summarising data - Tabular summary

<table>
<thead>
<tr>
<th># kids &lt; 6 yrs</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>606</td>
<td>118</td>
<td>26</td>
<td>3</td>
</tr>
</tbody>
</table>

- Most have no kids < 6
- None have more than 3

<table>
<thead>
<tr>
<th>Yrs education</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>30</td>
<td>25</td>
<td>44</td>
<td>43</td>
<td>381</td>
<td>44</td>
<td>51</td>
<td>14</td>
<td>57</td>
<td>46</td>
</tr>
</tbody>
</table>

- Most have around 12 years (high school)
- Bimodal – two clusters, one around 12 (high school) and another around 16-17 (college)
- None below 5 or beyond 17
**Graphical summary**

- Histogram is one of the most useful graphical summaries
- Groups observations into bins
- Bin width defines grouping
- Height (area) of bin proportional to group size
Numerical summary - Measures of location

- Location = “typical” value or “average” value of a variable
- Sample mean, sample median, sample mode

\[ \bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \]

(a) Sample mean

\[ \bar{X} = \frac{1 + 1 + 4 + \ldots + 3 + 4}{10} = 2.7 \]

(b) Sample median - “middle” observation when data are ranked from lowest to highest.
   If \( n = \text{odd} \), sample median = middle value. If \( n = \text{even} \), sample median = the average of the two middle values

1, 1, 2, 2, 3, 3, 4, 4, 5

\[ \frac{2 + 3}{2} = 2.5 \]

(c) Sample mode - most frequently observed value in \( n \) observations of a variable

1, 1, 2, 2, 3, 3, 4, 4, 5;
Numerical summary - Measures of spread

- Spread describes how the value of a variable changes over \( n \) observations
- Sample range, sample variance \( (s^2) \), interquartile range (IQR)

\[ \text{eg. } X_1, \ldots, X_{10} = 1, 1, 4, 2, 5, 2, 2, 3, 3, 4 \]

(a) Sample range = largest $-$ smallest = $5 - 1 = 4$

(b) Sample variance = average distance between observations and $\bar{X}$

\[
s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{(1 - 2.7)^2 + \ldots + (4 - 2.7)^2}{10 - 1} \approx 1.79
\]

Taking square root gives standard deviation \( (s) \)

(c) IQR = \( \text{upper quartile (75-th percentile)} - \text{lower quartile (25-th percentile)} \)

\[ 1,1,2,2,2,3,3,4,4,5 \]

\[
\text{IQR} = \frac{3 + 4}{2} - \frac{1 + 2}{2} = 3.5 - 1.5 = 2
\]
Randomness

Pandemic data
Treatment outcome from \( n = 100 \) patients in a pandemic:
1 = “recovered” and 0 = “not recovered”

\[
\begin{align*}
1 & 1 1 0 0 0 1 1 1 0 0 1 0 1 0 0 1 1 1 1 \\
1 & 1 1 0 1 1 0 0 1 1 1 1 0 1 1 0 0 1 1 0 \\
0 & 1 0 1 1 0 1 0 0 1 1 1 1 0 1 0 1 1 1 0 \\
1 & 0 1 0 1 1 0 0 0 0 1 1 1 0 1 0 0 0 1 1 \\
1 & 0 1 1 1 1 0 0 1 1 1 1 0 1 0 0 1 0 1 1 \\
\end{align*}
\]

\[
\begin{array}{c|cc}
\text{Outcome} & 0 & 1 \\
\hline
n & 40 & 60 \\
\end{array}
\]

Why did some patients recovered and others not? pattern of 1 and 0’s not easy to predict – random

Probability helps to explain randomness
A fair coin has a $\frac{1}{2}$ “probability” of observing heads, what does it mean?

<table>
<thead>
<tr>
<th>Toss</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome</td>
<td>H</td>
<td>T</td>
<td>H</td>
<td>T</td>
<td>T</td>
<td>H</td>
<td>T</td>
<td>H</td>
<td>T</td>
<td>...</td>
</tr>
</tbody>
</table>

The long run proportion (frequency) of heads is the probability of heads.
Probability is the long run frequency of an outcome

Probability cannot predict individual outcomes

However, it can be used to predict long run trends

Probability always lies between 0 and 1, with a value closer to 1 meaning a higher frequency of occurrence

Probability is numeric in value so we can use it to:
- compare the relative chance between different outcomes (events)
- carry out calculations
**Proportion and probability**

- Toss of fair coin: $H, T, H, T, T, H, T, H, T$

<table>
<thead>
<tr>
<th>Outcome</th>
<th>$H$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Proportion</td>
<td>4/9</td>
<td>5/9</td>
</tr>
<tr>
<td>Probability</td>
<td>1/2</td>
<td>1/2</td>
</tr>
</tbody>
</table>

- Treatment outcome

<table>
<thead>
<tr>
<th>Outcome</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Proportion</td>
<td>40/100</td>
<td>60/100</td>
</tr>
<tr>
<td>Probability</td>
<td>$P(0)$</td>
<td>$P(1)$</td>
</tr>
</tbody>
</table>

- Probabilities are *population* proportions
Probability Axioms - Urn model (1)- drawing marbles from an urn (with replacement)

Draws

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>🔵</td>
<td>🔵</td>
<td>🔵</td>
<td>🔵</td>
<td>🔵</td>
<td>🔵</td>
<td>🔵</td>
<td>🔵</td>
<td>🔵</td>
<td></td>
</tr>
</tbody>
</table>

Probability

<table>
<thead>
<tr>
<th>3/5</th>
<th>2/5</th>
</tr>
</thead>
</table>
Five possible **Outcomes**: 1, 2, 3, 4, 5

Interested in **Event A**: 

\[ A = \{1, 4, 5\} \]; hence an event is a collection of outcomes

\[ P(A) = \frac{3}{5} = 0.6(60\%) = \frac{\text{Number of marbles in } A}{\text{Total number of marbles}} \]
Complementary events

- Marbles in urn: 1 2 3 4 5

- Interested in \( \bar{A} \): (Not \( A \))

\[ \bar{A} = \{ 2, 3 \} \]

\( \bar{A} \), sometimes written as \( A^C \), is called the complementary event of \( A \)

- Chance of \( \bullet \) = 1 – chance of \( \bullet \)

\[ \Rightarrow P(\bar{A}) = 1 - P(A) = 1 - \frac{3}{5} = \frac{2}{5} \]
Joint probability and disjoint events

- A **joint probability** between two events \(A\) and \(B\) is one of the following:

\[
P(A \text{ and } B) \quad P(A \cap B) \quad P(AB)
\]

- The simplest form of joint probability between \(A\) and \(B\) is when they are **disjoint** (also called **mutually exclusive**). Disjoint events cannot occur simultaneously: \(P(A \text{ and } B) = 0\)

**Example**

\[A = \{\text{blue in 1st draw}\}\]
\[B = \{\text{green in 1st draw}\}\]

\[P(A \text{ and } B) = 0\]
Partition rule

- For any $A$, a partition is any collection of disjoint subsets of $A$ that together make up $A$, e.g.,

  If $A= \{\text{all white women}\}$, a partition of $A$ is $A_1 = \{\text{all white women with wage } < 3\}$ and $A_2 = \{\text{all white women with wage } \geq 3\}$

- For any $A$, $P(A)$ can be written as the sum of the probabilities of its disjoint subsets, e.g.,

$$
P(\bullet) = P(\bullet \ \text{and odd}) + P(\bullet \ \text{and even})
= P(\{\text{1, 5}\}) + P(\{\text{4}\})
= \frac{2}{5} + \frac{1}{5}
= \frac{3}{5}
$$
Conditional probability is a useful quantification of how the assessment of chance changed due to new information: “If A happened, what is the chance of B?”

The conditional probability of “B given A” is written as $P(B|A)$.

Example Drawing marbles WITHOUT replacement

$$A = \{ \text{blue is drawn} \}$$

$$B = \{ \text{green is drawn} \}$$

$$P(B) = \frac{2}{5}$$

$$P(B|A) = \frac{2}{4}$$
The most general way of calculating joint probability is the **Multiplication Rule**

\[ P(AB) = P(B|A)P(A) = P(A|B)P(B) \]

**Example**

Marbles in urn: 1 2 3 4 5

\[
P(\text{● and ●}) = P(\text{●|●})P(\text{●}) = \left(\frac{2}{4}\right) \left(\frac{3}{5}\right) = \frac{6}{20}
\]

\[
= P(\text{●|●})P(\text{●}) = \left(\frac{3}{4}\right) \left(\frac{2}{5}\right) = \frac{6}{20}
\]

**Rearranging the multiplication rule:**

\[ P(A|B) = \frac{P(AB)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(AB)}{P(A)} \]
Independence

- A and B are **independent** means they don’t offer information about one another.

- If A and B are independent, conditional probability becomes unconditional:
  
  (i) \( P(A|B) = P(A) \) \quad \text{“B says nothing about A”}
  
  (ii) \( P(B|A) = P(B) \) \quad \text{“A says nothing about B”}

- Independence is NOT the same as mutually exclusive (disjoint), which is \( P(A \text{ and } B) = 0 \). In fact when A and B are disjoint, they are very dependent.
Joint probability of independence events

When \( A \) and \( B \) are independent, using the multiplication rule and, (i) and (ii) from previous slide:

\[
(iii) \quad P(AB) = P(A|B)P(B) = P(A)P(B)
\]

We can use (i), (ii) or (iii) for any two independent events \( A \) and \( B \)

Example Drawing marbles WITH replacement

\[
A = \{ \text{is drawn} \} \\
B = \{ \text{is drawn} \}
\]

\[
P(B) = \frac{2}{5} \\
P(B|A) = \frac{2}{5}
\]

\[
P(AB) = P(B|A)P(A) = P(B)P(A) = \frac{2}{5} \times \frac{3}{5} = \frac{6}{25}
\]
**Union** of events can sometimes be best visualized using a **Venn diagram** (John Venn, 1834-1923)

Example What is the probability of drawing a green or an odd number?
Union of events (2)

\[ P(\text{\ding{51} or odd}) = P(\text{\ding{51}}) + P(\text{Odd}) - P(\text{\ding{51} and odd}) \]
\[ = \frac{2}{5} + \frac{3}{5} - \frac{1}{5} \]
\[ = \frac{4}{5} \]

In general, if \( A \) and \( B \) are:

- disjoint, then \( P(A \text{ or } B) = P(A) + P(B) \)
- not disjoint, then \( P(A \text{ or } B) = P(A) + P(B) - P(AB) \)
Probability tree is useful for studying combinations of events. Branches of a tree are *conditional* probabilities.

**Example**

Drawing two marbles from urn without replacement: ● ● ● ● ● ●

- **Draw 1**
  - Probability of drawing a green marble: \( \frac{2}{5} \)
  - Probability of drawing a blue marble: \( \frac{3}{5} \)

- **Draw 2**
  - **Condition: Green marble drawn first**
    - Probability of drawing another green marble: \( \frac{1}{4} \) (since one green marble is already drawn)
    - Probability of drawing a blue marble: \( \frac{3}{4} \)
    - Probability of both events: \( \frac{2}{5} \cdot \frac{3}{4} \)

  - **Condition: Blue marble drawn first**
    - Probability of drawing another blue marble: \( \frac{2}{4} \) (since one blue marble is already drawn)
    - Probability of drawing a green marble: \( \frac{2}{4} \)
    - Probability of both events: \( \frac{3}{5} \cdot \frac{2}{4} \)
Trees are useful for visualizing $P(B|A)$ when $B$ follows from $A$ in a natural (time) order. Many problems require $P(A|B)$, **Bayes Theorem** provides an answer.

**Example**

Testing for an infectious disease.

![Bayes Theorem Diagram](image)

What is $P(D|T)$ or $P(\bar{D}|ar{T})$?
Bayes Theorem (2)

\[
P(D|T) = \frac{P(D \cap T)}{P(T)} = \frac{P(T|D)P(D)}{P(T)} = \frac{P(T|D)P(D)}{P(T \cap D) + P(T \cap \bar{D})}
\]

Partition rule

\[
= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})}
\]

Multiplication rule

\[
= \frac{9 \cdot \frac{1}{10} \cdot \frac{1}{100}}{9 \cdot \frac{1}{10} \cdot \frac{1}{100} + \frac{1}{10} \cdot \frac{99}{100}} = \frac{9}{108}
\]

In general,

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]
\( P(D \mid T) \) using tree

\[
P(D \cap T) = \frac{9}{10} \cdot \frac{9}{10} = \frac{81}{100}
\]

\[
P(D \cap \bar{T}) = \frac{1}{10} \cdot \frac{1}{10} = \frac{1}{100}
\]

\[
P(\bar{D} \cap T) = \frac{99}{100} \cdot \frac{1}{10} = \frac{99}{1000}
\]

\[
P(\bar{D} \cap \bar{T}) = \frac{99}{100} \cdot \frac{9}{10} = \frac{891}{1000}
\]

\[
P(D \mid T) = \frac{P(DT)}{P(T)} = \frac{\frac{9}{100}}{\frac{9}{1000} + \frac{99}{1000}} = \frac{9}{108}
\]