Dis 3: Sampling (Cont'd); Probability

Related textbook chapters: 5 and 6
Ch 6 handout and solution offered by Dr. Pac can be accessed here: Handout Solution
This handout incorporates reviews with all exercises from the handout given by Dr. Pac.

1 Sampling

- Recall from discussion 1 that a sample is a subset of data taken from the population. The action of taking this subset of data to construct your sample is called **sampling**.
- Eventually, our goal is to use the sample to draw conclusion about the population (inferential statistics), so it is important that our sample resembles the population.
- Different **sampling plans** are proposed to construct the sample, weighing the benefits of the plan against the costs:
  - **Simple random sampling**: every possible sample entry has equal chance of being selected.
  - **Stratified random sampling**: separate the population into mutually exclusive sets (i.e. strata), and then draw simple random samples from each stratum.
  - **Cluster sampling**: population is first divided into groups, and then one uses simple random sampling to select groups; all observations within the selected groups thus enter the sample.

Exercise. Which sampling plan is used in each of the following examples?

1. Categorize all Econ 310 students based on their class standing (freshman, sophomore, junior, senior, above senior), and then randomly selects 30 students from each class.
   Stratified random sampling
2. Categorize all Econ 310 students based on their class standing (freshman, sophomore, junior, senior, above senior), and then randomly select 2 out of the 5 possible groups. The groups corresponding with the class standing selected are chosen as the sample.
   Cluster sampling
3. Number Econ 310 students sequentially from 1 to N. Draw 50 non-repeat random positive integers that are less than or equal to N. Select the students with the same numbers.
   Simple random sampling

- As you can already see from the exercise, factoring in the specific steps taken when sampling, some sampling plan is expected to construct a sample that more closely resembles the population than the others.

  To formally examine how far the samples are from the population, we look at two types of errors that occur:

  1. **Sampling error**: difference between the sample and the population that exists only because the observations that happen to be included in the sample.
     $\Rightarrow$ increasing the sample size reduces this error

2. **Nonsampling errors**: more serious type of error due to samples being selected improperly.
   $\Rightarrow$ increasing the sample size will NOT reduce this type of error
   Nonsampling errors can be divided into three categories:
   (a) **Errors in data acquisition**: the data is recorded wrong (due to incorrect measurement, mistake made during transcription, human errors)
   (b) **Nonresponse errors**: responses are not obtained from certain people.
   (c) **Selection bias**: some members from the target population cannot possibly be selected to be within the sample.

> Exercise. Which type of error arises from the following examples?
>
> 1. You sent out a survey to all Econ 310 students via email, but some people quickly archived your email without filling out the survey.
>   Nonsampling error – nonresponse error
> 2. You sent out a survey to all Econ 310 students via email, but some freshmen has yet to activate their UW email account, so the survey was not delivered to them.
>   Nonsampling error – selection bias
> 3. You randomly selected 30 Econ 310 students to have them answer your survey questions. All 30 of them responded, and you did not make any mistake in recording the data. However, your result derived from the sample is still quite different from the parameter in the population.
>   Sampling error
> 4. You randomly selected 30 Econ 310 students to have them answer your survey questions. All 30 of them responded, but you messed up the order of items in two columns of the data recorded.
>   Nonsampling error – errors in data acquisition

2 Probability

2.1 Sample Space, Outcome, and Event

- Say that we roll a six-sided die:
  - The roll could be 1, 2, 3, 4, 5, or 6
  - The set containing all possible rolls is defined as the **sample space**.
  - Each possible roll is an **outcome**.
    - Outcomes in a sample space should be **exhaustive** – all possible outcomes must be included in the sample space.
    - Outcomes in a sample space should be **mutually exclusive** – no two outcomes can occur at the same time.
  - An **event** is a collection of one or more simple events (i.e. outcomes) in a sample space.

Exercise. Is the following true or false? Why?

1. The set $\{\text{rolling a 1}, \text{rolling a 2}, \text{rolling a 3}, \text{rolling a 4}\}$ describes the sample space from a fair, six-sided die roll.

   False. This list is not exhaustive, since rolling a six-sided die also includes outcomes "rolling a 5" and "rolling a 6". The actual sample space is {rolling a 1, rolling a 2, rolling a 3, rolling a 4, rolling a 5, rolling a 6}.

2. "Rolling a 3" is an event.

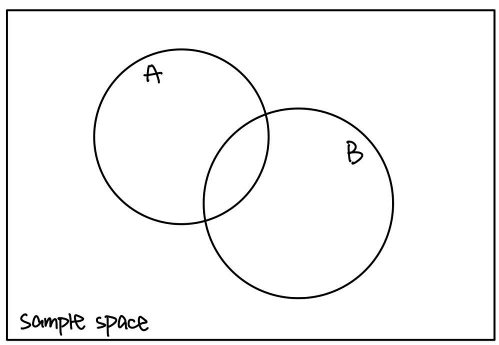
   True. "Rolling a 3" is also an outcome (also known as a simple event), and one simple event also constitutes as an event.

3. "Rolling an even number" is an event.

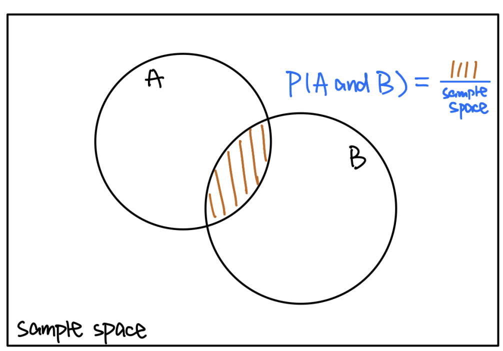
   True. "Rolling an even number" consists of simple events "Rolling a 2", "Rolling a 4", and "Rolling a 6". A collection of simple events is an event.

2.2 Joint vs. Conditional Probability

- Say that there are two events, denoted by A and B.



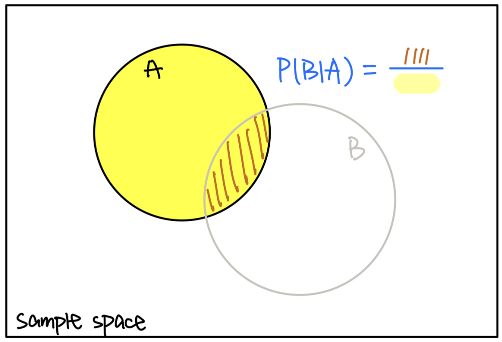
- If one is interested in the probability that event A and B occurs together, then



$$P(A \text{ and } B) = \frac{\text{Area of the line shaded space}}{\text{Area of the entire sample space (rectangle)}}$$

This is because **joint** probability looks at case where event A and B occur together with respect to all the possible outcomes (i.e. the sample space).

- Sometimes though, one is interested in the probability of B occuring conditional on A already occurs. This is depicted as



$$P(B|A) = \frac{\text{Area of the line shaded space}}{\text{Area of event } A} = \frac{P(A \text{ and } B)}{P(A)}$$

This is because **conditional** probability of $P(B|A)$ looks at case where event B occurs with respect to A already occuring.

- One way to visualize the relationship between joint and conditional probability is through the tree diagram (will see this in exercise later).
- Conditional probability gives rise to **Bayes Law**:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\underbrace{P(B|A)P(A) + P(B|A^C)P(A^C)}_{P(A \text{ and } B)+P(A^C \text{ and } B)}}$$

Go to Exercise Q3

2.3 Mutually Exclusive vs. Independence

- Two events are **mutually exclusive** if they cannot occur at the same time.

  e.g. When rolling a fair six-sided die, "Rolling a 3" and "Rolling a 4" are mutually exclusive, since you cannot roll both a 3 and a 4 at the same time.
- Two events are **independent** if one event is not impacted by the other.

  e.g. When rolling a fair six-sided die, "Rolling a 3 in the first round" and "Rolling a 3 in the second round" are independent, since the first and second roll does not influence each other.

  – This is why if A and B are independent, $P(B|A) = P(B)$, and $P(A|B) = P(A)$, since conditional on event A does not give us any new information regarding event B, and conditional on event B does not give us any new information regarding event A.
- When looking at joint probability between A and B,

- If A and B are mutually exclusive, $P(A \text{ and } B) = 0$
- If A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$

Exercise. Are the following two events mutually exclusive, independent, both, or neither?

1. "Weather today in Madison at 11:05am is sunny" vs. "Weather today in Madison at 11:05am is rainy"

   Mutually exclusive. Since the weather today in Madison at 11:05am can only be sunny or rainy (or something else), but not both at the same time. Additionally, when the weather in Madison today at 11:05am is sunny, it then cannot be that the weather is rainy at this point, so the two events impact each other, so they are not independent.

2. "Roll a (fair six-sided) die and get an even number" vs. "Roll a (fair six-sided) die and get 4"

   Neither mutually exclusive nor independent. It is possible that a die roll is both even and is exactly 4, so they are not mutually exclusive. Clearly, when a die roll is 4, then it is absolutely an even number. The two events impact each other, so they are not independent.

3. "Flip a coin and get tail" vs. "Roll a (fair six-sided) die and get 4"

   Independent. Flip a coin does not interact with rolling a die. On the other hand, both events can occur at the same time, so they are not mutually exclusive.

4. "Flip a coin and get tail" vs. "Roll a (fair six-sided) die and get 10"

   Both mutually exclusive and independent. It is not possible to have both events occur at the same time (mainly because it's fully impossible to get a 10 from rolling a fair six-sided die), so they are mutually exclusive. On the other hand, the two events do not impact each other (flipping a coin has no impact on the result of rolling a die), so they are also independent.

Go to Exercise Q2

3 Exercises

1. You go to Vegas to play craps; luckily, Econ 310 has prepared you to solve the following problems:

   (a) Assuming you rolled a single die and got an even number, what's the probability the number on that die is a two?

   First, since there are six equally likely outcomes in the sample space, we know that:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$$

   By the addition rule (note that these events are mutually exclusive), we also know that:

$$P(\text{even}) = P(2) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

   Finally, note that $P(\text{even}|2) = 1$, since a die roll of two ensures the roll must be even. Given this, applying Bayes' Law, we have:

$$P(2|\text{even}) = \frac{P(2 \text{ and even})}{P(\text{even})} = \frac{P(\text{even}|2)P(2)}{P(\text{even})} = \frac{1 \times \frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

(b) Rolling two dice, what's the probability of two sixes? What's the probability of no sixes?

Since there are 36 equally likely outcomes in the sample space:

$$P(\text{six on both sides}) = P(6,6) = \frac{1}{36}$$

By the addition rule:

$$\begin{aligned} P(\text{at least one six}) &= P(\text{six on die one}) + P(\text{six on die two}) - P(\text{six on both dice}) \\ &= \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36} \end{aligned}$$

By the complement rule:

$$P(\text{no sixes}) = 1 - P(\text{at least one six}) = 1 - \frac{11}{36} = \frac{25}{36}$$

(c) Rolling two dice, what's the probability both dice are even? What's the probability either die is even?

Since rolling an even number on the first die and the second die are independent events, we can apply the simplified version of the multiplication rule:

$$\begin{aligned} P(\text{even on Die 1 and even on Die 2}) &= P(\text{even on Die 1}) \times P(\text{even on Die 2}) \\ &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \end{aligned}$$

By the same logic, the probability of both dice coming up odds is also $\frac{1}{4}$. So, we can conclude that:

$$P(\text{even on Die 1 or even on Die 2}) = 1 - P(\text{odd on Die 1 and odd on Die 2}) = 1 - \frac{1}{4} = \frac{3}{4}$$

The last equation is derived from the fact that the opposite (i.e. complement) event of either die being even is that both dice are odd.

(d) Rolling two dice, what's the probability they sum to seven?

There are six permutations that lead to a seven, so since there are 36 equally likely outcomes in the sample space the addition rule tells us:

$$\begin{aligned} P(\text{two dice sum to 7}) &= P(1,6) + P(2,5) + P(3,4) + P(4,3) + P(5,2) + P(6,1) \\ &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{6} \end{aligned}$$

The order between the two dice matters. For example, $P(1,6)$ represents that the first die is a 1, and the second die is a 6. This differs from getting 6 on the first die, and 1 on the second die.

2. Consider the following joint probability table:

	A	A^C
B	.15	.25
B^C	.40	.20

(a) What is $P(A \text{ and } B)$? What is $P(A)$? What is $P(B)$?
The table gives us all the joint probabilities, so we know that $P(A \text{ and } B) = 0.15$.
Using the addition rule we can calculate the marginal probability of A:

$$P(A) = P(A \text{ and } B) + P(A \text{ and } B^C) = 0.15 + 0.40 = 0.55$$

Similarly, we can find the marginal probability of B:

$$P(B) = P(A \text{ and } B) + P(A^C \text{ and } B) = 0.15 + 0.25 = 0.40$$

(b) What is $P(A \text{ or } B)$?
We use the addition rule (note that since these events are not mutually exclusive, we must subtract off the joint probability):

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.55 + 0.40 - 0.15 = 0.80$$

(c) What is $P(A|B)$? What is $P(B|A)$?
By the definition of conditional probability,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{0.15}{0.40} = 0.3750$$

Similarly,

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{0.15}{0.55} = 0.2727$$

(d) Are A and B independent?
No, since

$$P(A|B) = 0.3750 \quad \neq \quad 0.55 = P(A)$$

Or, equivalently, since

$$P(B|A) = 0.2727 \quad \neq \quad 0.40 = P(B)$$

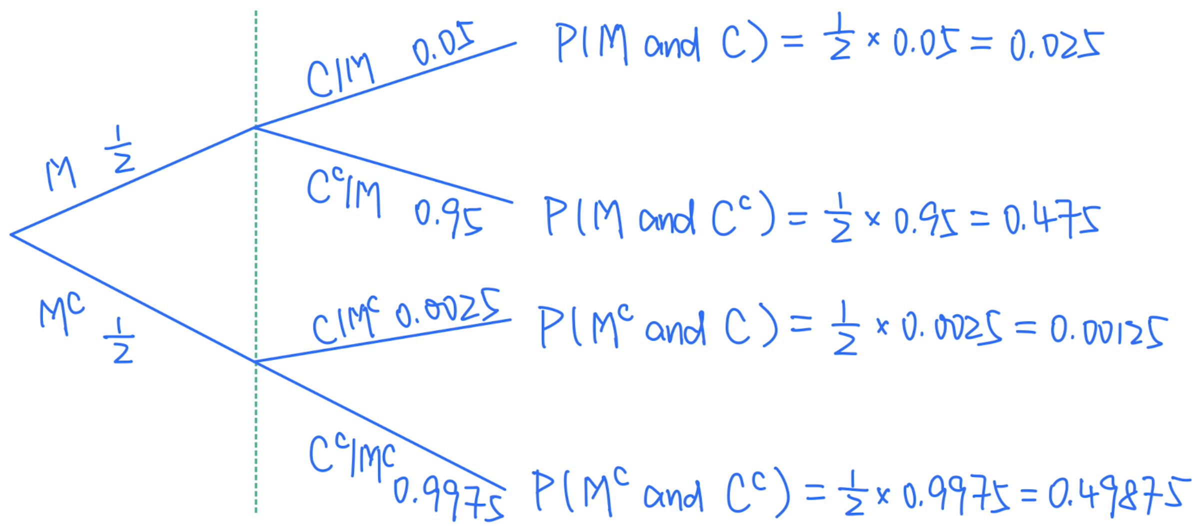
Or, equivalently, since

$$P(A \text{ and } B) = 0.15 \quad \neq \quad 0.22 = 0.55 * 0.40 = P(A) \times P(B)$$

3. Suppose exactly half the population is male. 5% of males and only 0.25% of females are color blind.

   (a) Draw a tree diagram to illustrate the sample space.

   Denote the event of being a male as M, and the event of being color blind as C. The tree diagram should look like the following:



   (b) A person is chosen at random and that person is color blind. What's the probability that the person is a male?

   Since the person chosen is color blind, being color blind (C) is the event we condition on. So the question is asking about the conditional probability $P(M|C)$.

   Using the definition of conditional probability, we have

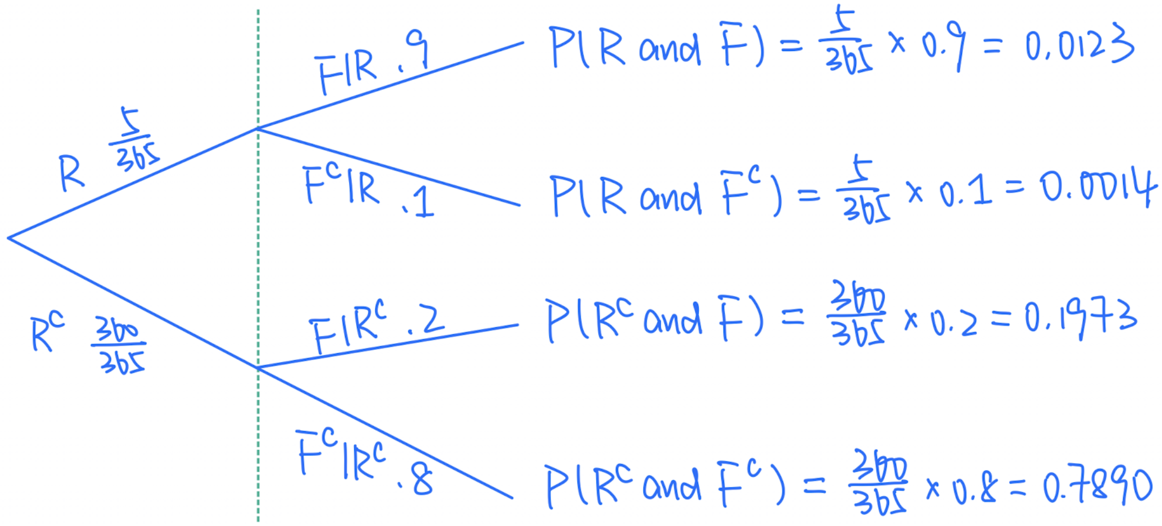
$$\begin{aligned} P(M|C) &= \frac{P(M \text{ and } C)}{P(C)} \\ &= \frac{P(M \text{ and } C)}{P(M \text{ and } C) + P(M^C \text{ and } C)} \\ &= \frac{0.025}{0.025 + 0.00125} = 0.9524 \end{aligned}$$

   Thus, when a person is color blind, the probability that the person is a male is 0.9524.

4. Marie is getting married tomorrow at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 20% of the time. Assume there are 365 days in a year.

   (a) Draw a tree diagram to illustrate the sample space.

   Denote the event of raining as R, and the event of the weatherman forecasting that it will rain as F. The tree diagram should look like the following:



(b) What is the probability it will rain tomorrow for Marie's wedding?

Since the weatherman already predicted rain for tomorrow, the probability that the question is asking should condition on the event F; that is, the question asks for $P(R|F)$.

Using the conditional probability formula, we have

$$\begin{aligned} P(R|F) &= \frac{P(R \text{ and } F)}{P(F)} \\ &= \frac{P(R \text{ and } F)}{P(R \text{ and } F) + P(R^C \text{ and } F)} \\ &= \frac{0.0123}{0.0123 + 0.1973} = 0.0587 \end{aligned}$$

Thus, given that the weatherman predicted rain for tomorrow, the probability that it will rain tomorrow is 0.0587.