INTRODUCTION TO STATISTICS

Session 7/8

May 29th / June 5th 2017

Madoka Takeuchi

Review

- Multiple Regression
  - Extension of simple linear regression
  - Prediction model with many independent variables
    - $y = a_1x_1 + a_2x_2 + a_3x_3 + \ldots\ldots a$
    - $x_1, x_2, x_3, \ldots\ldots x_n$ are the independent variables
    - $a_1, a_2, a_3, \ldots a_n$ are called coefficients
- Must check that the data can actually be analyzed using multiple regression- must "pass" the eight assumptions
  - In reality, very hard to pass all the assumptions

Review: Model building/selection

- In most situations a simple model with the fewest necessary variables is favorable
  - Easier to interpret
- Choose model with the smallest variance-large R^2
- There is no right model- depends on what is of interest
  - Model selection- Needs to be justifiable

Probability

- Definition:
  - chance that something will happen or not happen
  - measure of how often a particular event will occur if something is done repeatedly
  - Study of randomness and uncertainty
- Probability quantifies the chance that a certain event will occur- "The chance of winning the lottery is 1 in 41,416,353".
  - Probability is based on the idea of a random experiment where the outcome cannot be predicted with absolute certainty before it is run

    i.e
    - coin tossing- when a coin is tossed there are two possibilities- Heads and Tails
      - The probability of getting a H is ½ and the probability of getting a T is ½
    - Throwing a dice- there are 6 possibilities (1,2,3,4,5,6)
      - The probability of getting any one is 1/6

Combinatorial Analysis: Basic principle of counting

- Definition of combinatorial analysis- theory of counting
- In probability- many problems can be solved by counting the number of ways a certain event can occur

Basic Principle of Counting

If there are "x" possible outcomes in one experiment and "y" possible outcomes in another experiment, then there are "xy" possible outcomes in the two experiments

Example (1)

1. one has 2 shirts and 4 pants
   - there are 8 (2x4) possible outfits (combinations of pants and shirts)

Example (2)

You are buying a car

There are 2 body types (sedan/ suv)

There are 5 colors (black, white, silver, red, blue)

There are 3 models (standard, sports, luxury)

How many total choices are there?

Example (3)

Q. You must choose a four-digit PIN number. Each digit can be chosen from 0 to 9. How many different possible PIN numbers can you choose?

A. 10 x 10 x 10 x 10

Q. How many different 7-place license plates are possible if the first 2 places are letters and the final 5 are numbers?

A. 26 x 26 x 10 x 10 x 10 x 10 x 10= 67600000

Permutations

- Combinations vs. Permutations
  - Combination-Order does not matter
  - Permutation- order does matter- ordered combination
- How may different arrangements of "x,y,z" are possible?
  - xyz, xzy, yxz, yzx, zxy, zyx= 6 different arrangements
  - Each arrangement is known as a permutation
  - When there are 3 objects (x,y,z) there are 6 permutations
    - The first object can be any of the three, the second object can be any of the two remaining, the third object can only be the remaining object- 3 x 2 x 1= 6
    - Suppose we have *n* objects- then we have

n* (n-1)* (n-2)*…2*1 different permutations

this is mathematically expressed as ***n!***

*i.e- 5!= 5*4*3*2*1*

**0!=1*

Example (1)

Q. How many different batting orders are there for a baseball team of 9 players?

A. 9!= 9*8*7*6*5*4*3*2*1

= 362,880

-A class consists are 7 men an 3 women. The students are ranked based on test score. No one had the same test score

Q. How many possible rankings are possible?

A. 10!= 3628800

Q. If the men and women are ranked separately, how many possible rankings are possible?

A. There are 7! possible rankings for men and 3! possible rankings for women. From the basic principle of counting, there are (7!)*(3!) possible rankings

= 5040*6

= 30240

Example (2)

Q. Suppose we have 8 cups numbered from 1-8, how different arrangement are possible?

A- 8!

Q. Of the 8 cups we could only choose 3 cups- how many arrangements are possible?

A- 8*7*6= 336

there are 336 different ways that 3 cups could be arranged out of 8 cups

This can be written as ***n!/(n-r)!***

where ***n*** *is the number of things to choose from, and choose* ***r*** *of them*

Combinations

- We are often interested in finding the number of different groups that can be formed from the total.
  - How many different groups of 3 items can be selected from 5 items of A,B,C,D,E?
  - To answer the question→
    - There are 5 ways to select the first item, 4 ways to select the second item and 3 ways to select the 3rd item.
    - BUT every group of three, i.e ABC, will be counted six times (ABC, ACB, BAC,BCA, CAB, CBA) and in this question the order of the items is not relevant- must adjust for the number of ways the selected items can be ordered (1/r!).

$$\frac{n!}{(n-r)!} \times \frac{1}{r!} = \frac{n!}{r!(n-r)!}$$

$$\frac{n!}{r!(n-r)!} = \binom{n}{r} = \binom{n}{n-r}$$

Example (1)

Q- A group of 5 is to be formed from a group of 20 people- how many groups are possible?

A. n=20 r= 5

20!/ 5!(15!)→ 20*19*18*17*16/ 5*4*3*2*1

= 15504

Example (2)

Q. There are 5 women and 7 men and need to select 2 women and 3 men- how many combinations are possible?

A.

Among the women- n=5 r=2 so the number of combinations is 5!/2!3! =10

Among the men- n=7 r=3 so the number of combinations is 7!/3!4! = 35

Following the basic principle of counting- the number of combinations of 2 women and 3 men are 10*35= 350

Probability Terminology

- Random phenomenon- an event whose individual outcome is uncertain but if repeated, the distribution follows a regular pattern
  - Coin toss- do not know whether a single toss will be heads or tails- after multiple repetitions- 50% heads 50% tails
  - Rolling a dice- do not know what number will appear if a dice is rolled once, but after multiple replications- know that each face (1-6) has a 1/6 chance
- Outcome- the actual value after replication(s) of an experiment- result of an experiment
  - Coin tossing
    - T after one toss
    - THT after three tosses

Probability Terminology cont`d

- Sample space (S)- set of all possible outcomes of an experiment.
  - Mathematically the sample space is denoted by the symbol *S*

  i.e.
  - coin tossing- assuming that it is impossible for a coin to land on its edge, the sample space *S is* {H,T}
  - If two coins are tossed, what is the sample space?
- Event (E)- subset of outcomes contained in the sample space S

  i.e.
  - Getting a H from 1 coin toss
  - Choosing a Jack from a deck of cards

Simple Probability

- The probability of an event is a proportion/frequency
- If a sample size is finite and each of the outcomes have the same probability, then the probability of the event happening is

(the number of ways the event can happen) / (total number of outcomes)

$$P(E)= n(E)/ n(S)$$

Example

A dice is rolled, what if the probability of getting a 3?

The event of interest is getting a 3 so E={3}

The sample space is S={1,2,3,4,5,6}

The number of possible outcomes in E is 1→ n(e)=1

The number of total outcomes in S is 6→ n(s)=6

The probability of the event P(E)= 1/6

Problem

Two die are rolled- what is the probability of getting a sum bigger than 7?

What is the event of interest?

What is the Sample space?

What is n(E) and n(S)?

Problem

Event of interest = sum larger than or equal to 7

E= {(1,6), (2,5), (2,6), (3,4), (3,5),(3,6), (4,3), (4,4), (4,5), (4,6), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)}

Sample Space= {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5),(3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)}

n(E)= 21

n(S)= 36

Probabilty of the event if interest (P(E))= 21/36

Probability Rules

All probabilities are between 0 and 1

- 0 <= P(Event) <= 1

The sum of all the probabilities in the sample space is 1

- P(A) + P(B)+….=1

The probability of an event which cannot occur is 0.

- The probability of any event which is not in the sample space is zero.

The probability of an event which must occur is 1.

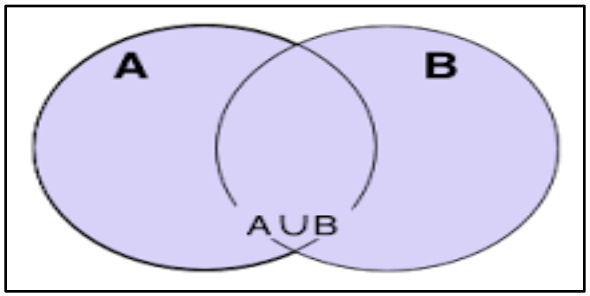
- The probability of the sample space is 1 P(S)=1

The probability of an event not occurring is one minus the probability of it occurring.

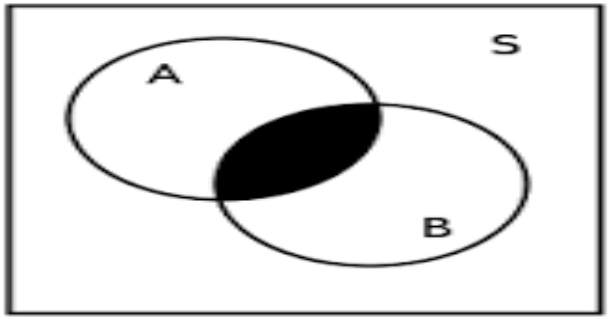
- P(E') = 1 - P(E)

Events in a sample space: set theory

1. If there exists two events A and B in the sample space, the **union** is the event that either A or B or both occurs
   - (A or B) = A ∪ B



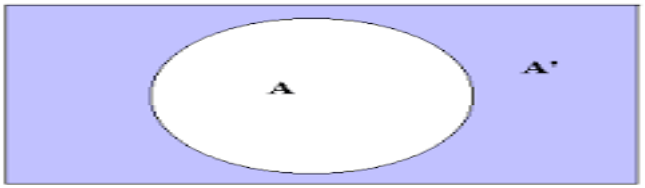
1. Intersection of event A and B is the event that both A and B occur
   - (A and B) = A ∩ B
   - AB



Events in a sample space: set theory (2)

3. The complement of an event B is the set of all outcomes in *S* that are not in B- denoted as B' or B^c

- When the event is an odd number on a dice then the complement is an even number
  - S= $A \cup A^c$
  - S= $B \cup B^c$



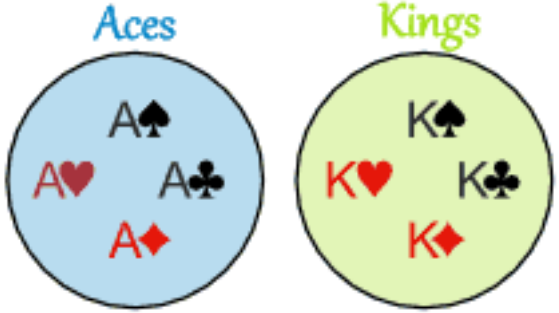
P(A) means "Probability of Event A"

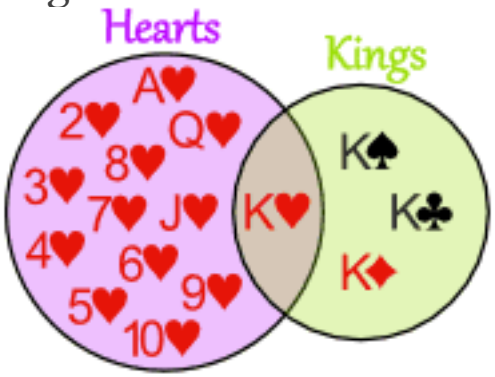
P(A') means "Probability of the complement of Event A"

P(A)+ P(A')=1

3. Events A and B are mutually exclusive/disjoint if none of the outcomes in A are in B and vice versa

- Turning left and turning right are mutually exclusive- can`t do both at the same time
- In a deck of cards- aces and kings are mutually exclusive but hearts and kings are NOT mutually exclusive since there is king of hearts





When two events (A and B) are mutually exclusive the probability that both events occur is 0

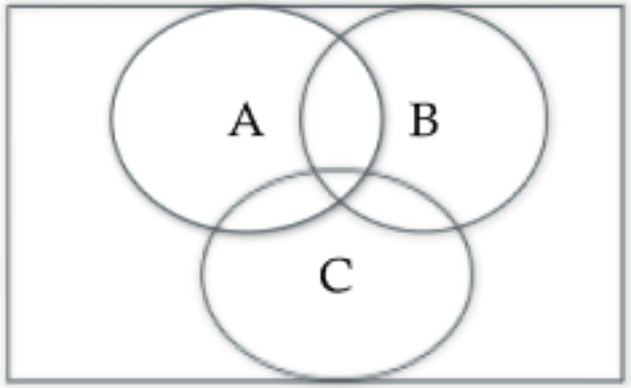
$$P(A \text{ and } B)=0 \qquad P(A \cap B)=0$$

- Only valid when the events are mutually exclusive

$$P(A \text{ or } B) = P(A) + P(B)$$

Set Algebra Axioms (1)

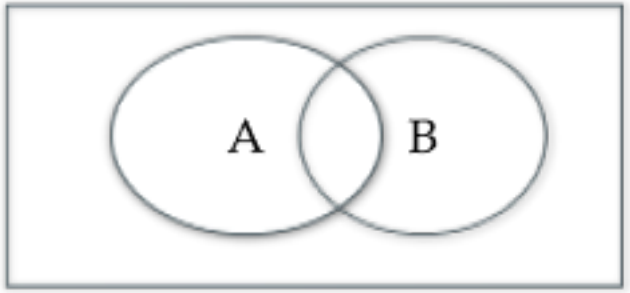
1. Operations are commutative
   - $A \cup B = B \cup A$
   - $A \cap B = B \cap A$
2. Operations are associative
   - $(A \cup B) \cup C = A \cup (B \cup C)$
   - $(A \cap B) \cap C = A \cap (B \cap C)$



3. The symbol for the empty set is $\varnothing$. The empty set contains no outcomes.
   - $A \cap \varnothing = \varnothing$
   - $A \cup \varnothing = A$
   - $A \cap A^c = \varnothing$

4. In events that are not mutually exclusive,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$



Set Algebra Axioms (2)

4. Rules involving the Sample space S.
   - $A \cap S = A$
   - $A \cup S = S$
   - $A \cup A^c = S$

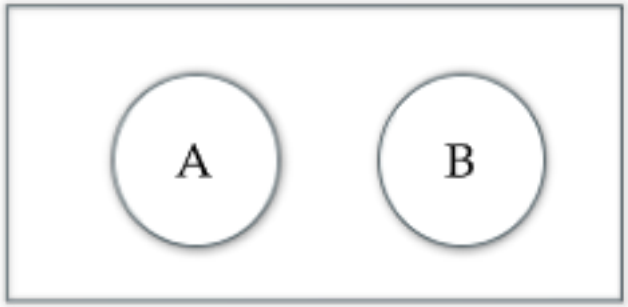
5. Rules about the complement
   - $(A^c)^c = A$
   - $\varnothing^c = S$
   - $S^c = \varnothing$

Set Algebra Theorems

Theorem 1

$$A \cap A = A$$

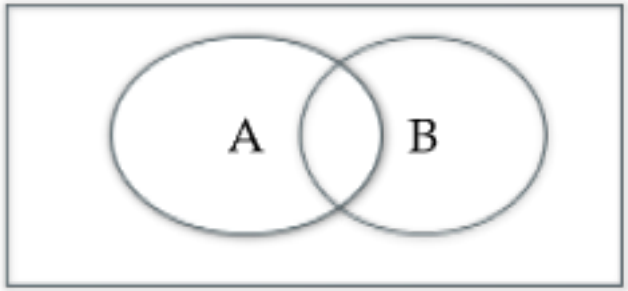
$$A \cup A = A$$



Theorem 2.

$$A \cap (A \cup B) = A$$

$$A \cup (A \cap B) = A$$



Theorem 3. DeMorgans Laws for sets

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup B^c$$

Conditional Probability

- probability of an event occurring given that another event has already occurred is called a conditional probability.
- The probability that B occurs, given that event A has already occurred is

$$P(B|A) = P(A \cap B) / P(A)$$

  - Can be re-written as

$$P(A \cap B) = P(B|A)\ P(A)$$

  - Since we are given that event A has occurred, we have a reduced sample space. Instead of the entire sample space S, we now have a sample space of A since we know A has occurred
  - It is the number in A and B (must be in A since A has occurred) divided by the number in A

Example

- A questionnaire was handed out to 100 people asking "Do you smoke?"
- Results-

Gender	Yes	No	Total
Female	12	28	40
Male	19	41	60
Total	31	69	100

Q1.What is the probability of a randomly selecting an individual being a male?

A1. Total males divided by the total = 60/100 = 0.60. There is no mention of smoking or not smoking so all cases are included.

Q2.What is the probability of a randomly selecting an individual being a male who smokes?

A2.The number of "Male and Smoke" divided by the total = 19/100 = 0.19

Q3.What is the probability of a randomly selecting a smoker from the males?

A3.19 males smoke out of 60 males, so 19/60 = 0.31666...

Q4.What is the probability that a randomly selected smoker is male?

A4.told that you have a smoker and asked to find the probability that the smoker is also male. There are 19 male smokers out of 31 total smokers, so 19/31 = 0.6129 (approx)

Bayes Theorem

Bayes theorem is a direct application of conditional probabilities

Bayes' Theorem is used to find the conditional probability of an event $P(A \mid B)$, when the "reverse" conditional probability $P(B \mid A)$ is the probability that is known.

$$P(A_1 \mid B) = \frac{P(A_1)\, P(B \mid A_1)}{P(A_1)\, P(B \mid A_1) + P(A_2) P(B \mid A_2)}$$

Example

- Mr X is planning on golfing tomorrow and would like to know the probability that it will rain. It recent years it has rained 50 days a year. The weather forecast predicts rain for tomorrow. When it rains the weatherman correctly forecasts rain 90% of the time. When it doesn't't rain, the weatherman incorrectly forecasts rain 10% of the time

Event A_1= it will rain tomorrow

Event A_2= it does not rain

Event B= the weatherman predicts rain

- $P(A_1)= 50/365 = .137$
- $P(A_2)= 315/365 = .863$
- $P(B|A_1)= 0.9$ (When it rains, the weatherman predicts rain 90% of the time)
- $P(B|A_2)= 0.1$ (When it does not rain, the weatherman predicts rain 10% of the time)

Mr X would like to know the probability of rain tomorrow given that the weatherman predicted rain $P(A_1|B)$

$$P(A_1 \mid B) = \{P(A_1)\, P(B \mid A_1)\} /$$

$$\{P(A_1)\, P(B \mid A_1) + P\,(A_2) P(B \mid A_2)\}$$

P(Rain | Weatherman predicts rain) =

{P(Rain) P(Weatherman predicts rain | rain)} /

{ P(Rain) P(Weatherman predicts rain | rain) + P (No Rain)P(Weatherman predicts rain | No Rain)}

$$P(A_1 \mid B) = (0.137)\ (0.9) / \ [(0.137)(0.9) \ + (.863)(0.1)]$$

$$= .1223 / \ [.1223 + .0863]$$

$$= .586$$

This means that when the weatherman predicts rain, it rains 58.6% of the time

In class Exercise

- Student A thinks she is allergic to Eggs and goes to the doctors office to get an allergy test. For people who who do have an egg allergy, the test says yes 80% of the time.
- For people who do not have allergy, the test says yes 10% of the time.
- If 1% of the population has the allergy and student A`s test says yes, what is the probability that Student A actually has the allergy?

P(Allergy)=1%
P(No Allergy)=99%
P(Yes|Allergy)=80%
P(Yes|No Allergy)=10%

Want to know P(Allergy|Yes)!

P(allergy|Yes)= {P(Allergy) P(Yes|Allergy)}/ {P(Allergy)P(Yes|Allergy) + P(no allergy) P(Yes| no Allergy)}

P(allergy|Yes)= (1%* 80%)/ {(1%*80%)+ (99%*10%)}
= (7.48%)