Conference 1

March. 15. 2019

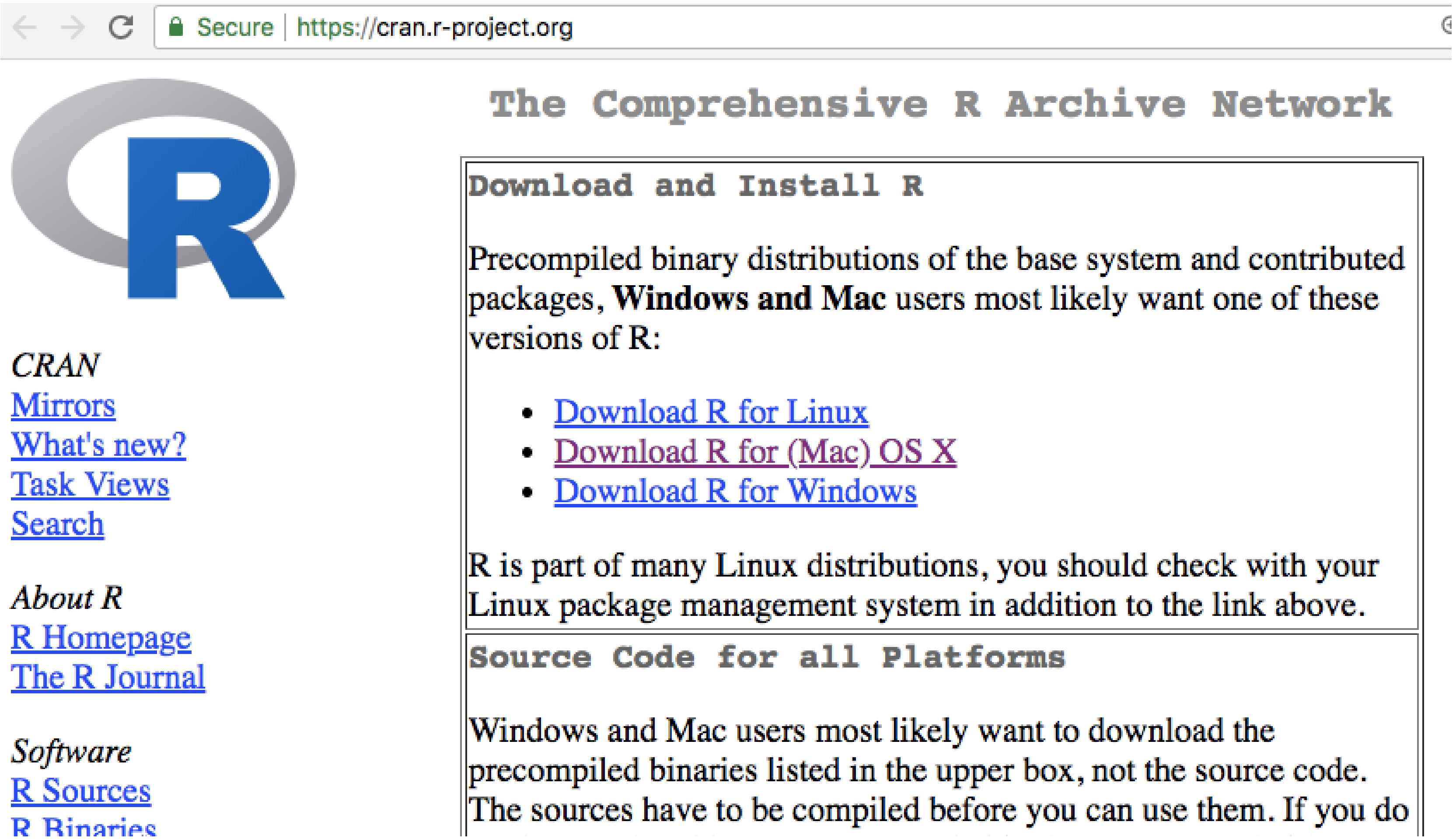
Jee-Young Moon

Outline

1. **Basics of R and Data Exploration**
2. **Event and Probability**
3. **Probability Distribution**
   a. **Binomial distribution**
   b. **Normal distribution**

I. Basics of R and Data Exploration

- Download R from https://cran.r-project.org/ (search “R software”)



- Install it

A. Data set on Maintaining Balance with Age

One day, Einstein pressed the panic button for his car by mistake and it made the people around the car jump around. He, the eldest among the people around him, was the one who jumped the most. This got him to think whether maintaining balance to an unpredictable noise gets difficult when you are older. He designed a study to recruit people of various age groups to measure the maintaining of balance to a noise. Each subject was told to memorize words on the screen while maintaining a stable upright position on a "force platform". The noise came randomly and the platform automatically measured how much each subject swayed in millimeters in both the forward/backward and the side-to-side directions.

- Research Aim: To test the association of maintaining the balance and age
  - Null hypothesis (H_0): maintaining the balance is not associated with age.
  - Alternative hypothesis (H_A): the null hypothesis is not true (maintaining the balance is associated with age).

balance.txt

	A	B	C
1	Age	FBSway	SideSway
2	Young	25	17
3	Young	21	10
4	Young	17	16
5	Young	15	22
6	Young	14	12
7	Young	14	14
8	Young	22	12
9	Young	17	18
10	MiddleAge	20	21
11	MiddleAge	23	30
12	MiddleAge	18	25
13	MiddleAge	21	12
14	MiddleAge	19	19
15	Elderly	19	14
16	Elderly	30	41
17	Elderly	20	18
18	Elderly	19	11
19	Elderly	29	16
20	Elderly	25	24
21	Elderly	21	18
22	Elderly	24	21
23	Elderly	50	37
24			

Download the dataset from the canvas (balance.txt)

Variable	Description
Age	Elderly, MiddleAge, Young
FBSway	Sway in forward/backward direction
Sidesway	Sway in side-to-side direction

Introduction to Statistics (from 1.1 Lecture note)

- Statistics is the science of data: designing collection of data, **exploring** data, and drawing conclusions from data. (Petruccelli, et al. p3)

Data exploration

Data Input

- Get and Set the working directory

> getwd() # get working directory

> setwd("/Users/moon/Desktop/Course7010/") # set working directory

✓ Alternatively, through the menu:

  ✓ In Windows: go to the *File* menu, select *Change Working Directory*, and select the appropriate folder/directory

  ✓ In Mac: go to the *Misc* menu, select *Change Working Directory*, and select the appropriate folder/directory

- Read in the data and name it as "balance"

After changing the working directory to where your downloaded data is

```
> balance <- read.delim('balance.txt', header=TRUE)
```

- View the data

```
> balance   ## all data
> head(balance) ## first few lines
> tail(balance)  ## last few lines
> balance[1:5, ]  ## first 5 lines
> balance[1:4, 1:2]  ## first 4 lines and 2 columns
```

Note:

1. R is case-sensitive.

2. A comment starts with a hash sign #

3. An input command comes after the prompt sign >

4. To know more about the command, use ?

```
> ?read.delim
```

- Check the sample size and number of variables

```
> dim(balance)   ## dimension of data
> ncol(balance)  ## number of columns in the data
> nrow(balance)  ## number of rows in the data
```

- Show the column names

```
> colnames(balance)
```

- Show the structure of the data

```
> str(balance)
> class(balance$FBSway) # class of FBSway variable
```

- Show a single variable

```
> balance$FBSway
> balance[, 'FBSway']
```

- Create a new variable

```
> x <- balance$FBSWay
> x <- c(2,5,6,1)
```

- Remove a variable

```
> rm(x)
```

Visual exploration of Data

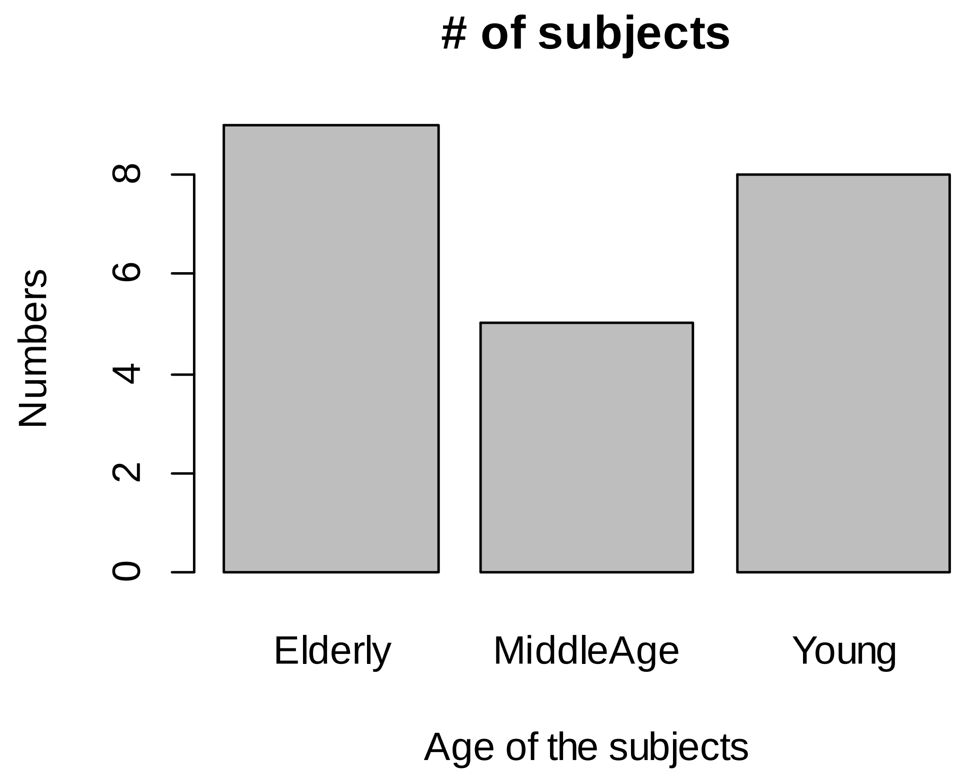
- Make a plot on the number of subjects by age group
- How many subjects of each age group

> table(balance$Age) # table() produces a contingency table of the counts at each distinct value (level) of the variable

9 elderly, 5 middle age, and 8 young age subjects were recruited.

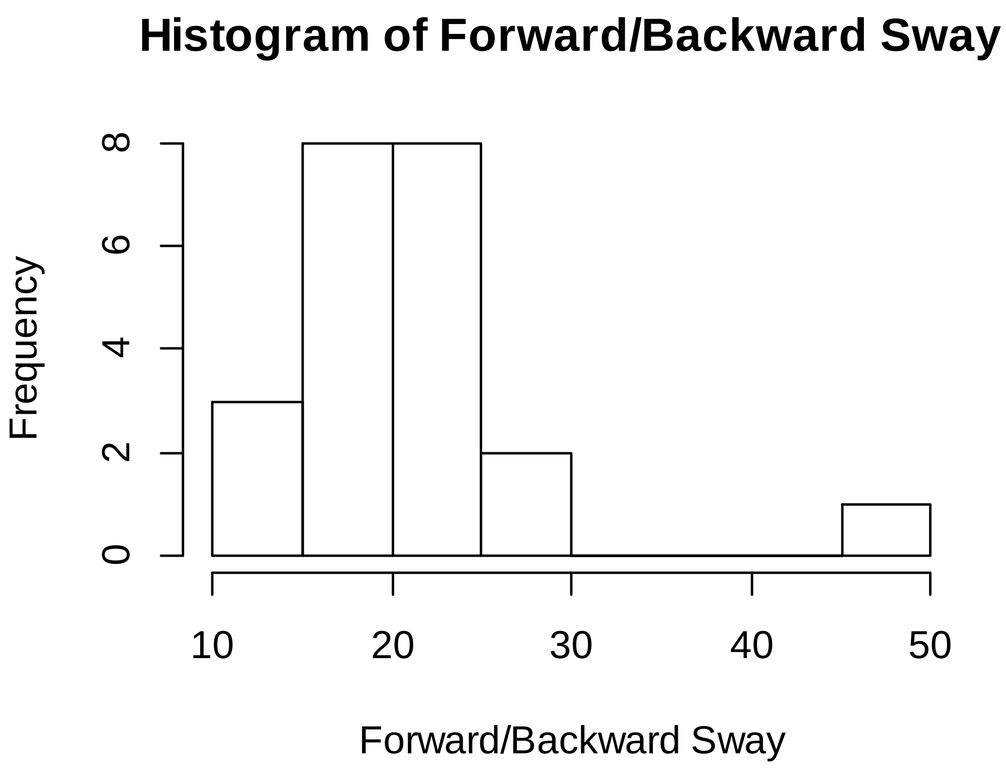
- Make a barplot

> barplot(table(balance$Age), main='# of subjects', xlab='Age of the subjects', ylab='Numbers')



- Check the distribution of forward/backward sway
- Make a histogram

```
> hist(balance$FBSway, main='Histogram of Forward/Backward Sway', xlab='Forward/Backward Sway')
```



For a single variable exploration:
Barplot for a categorical variable.
Histogram for a continuous variable.

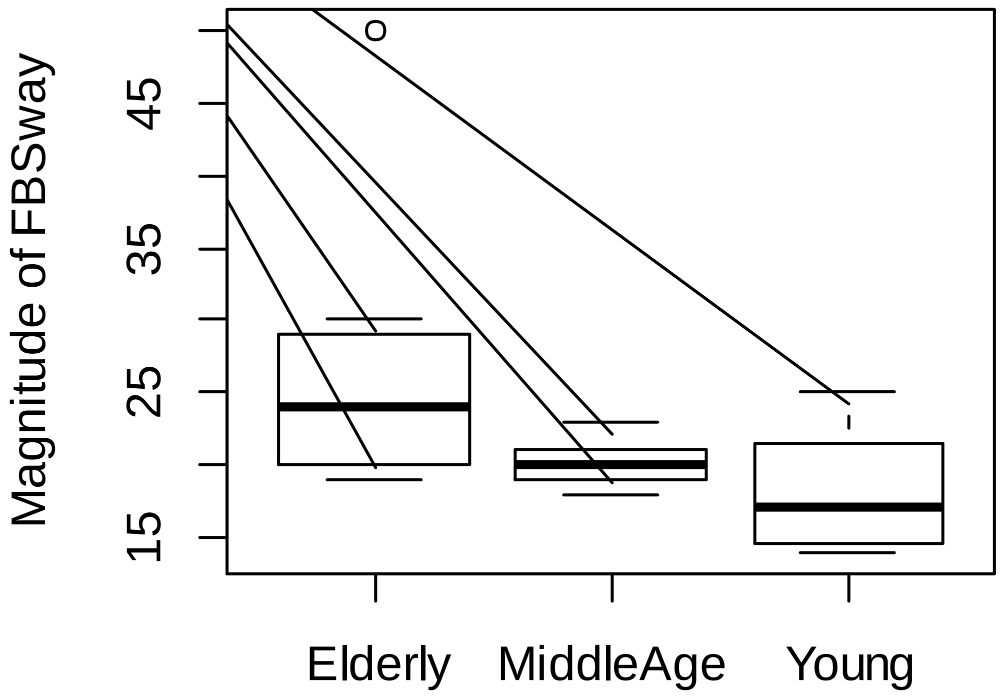
Research Aim: To test the association of maintaining the balance and age

- Q: Is Sway in forward/backward direction different across age groups?

<u>Research Aim</u>: To test the association of maintaining the balance and age

- <u>Q: Is Sway in forward/backward direction different across age groups?</u>

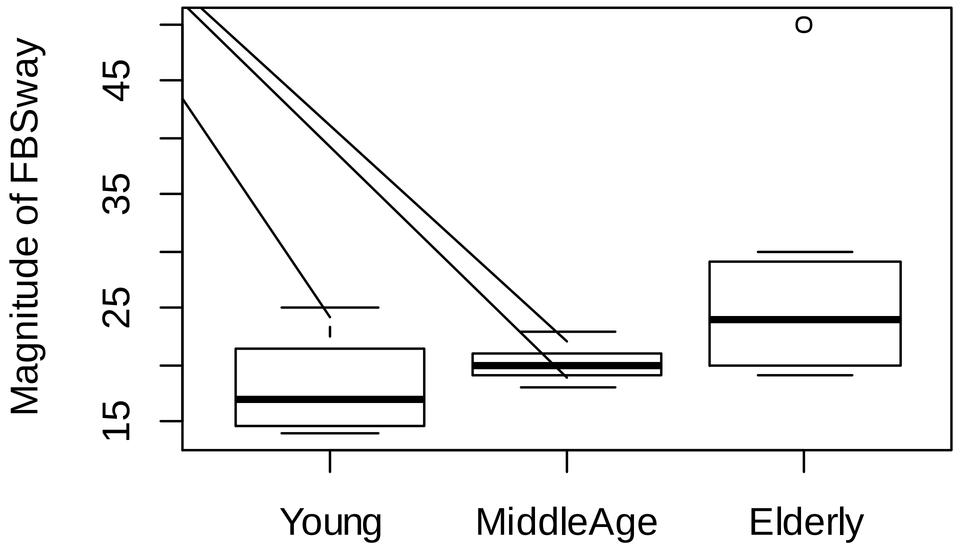
```
> boxplot(balance$FBSway~balance$Age, ylab="Magnitude of FBSway")
```



But, I want to order the groups from young, middle age, to elderly!!

- Re-level the age groups

```
> balance$Age <- factor(balance$Age, levels=c('Young', 'MiddleAge', 'Elderly'))
> boxplot(balance$FBSway~balance$Age, ylab="Magnitude of FBSway")
```



✓ Box plot graphically gives the location of the quartiles (first quartile, median, third quartile), and extreme values. The advantage of using box plots is that several of the characteristics of the data such as outliers, symmetry features, the range, and dispersion of the data can be easily compared between different groups. (Forthofer, 2007)

- From the visual exploration of the data using boxplot, we observe a increasing trend of forward/backward sway with increasing age.

Making a conclusion

- [Advanced] A statistical test can be done by fitting a linear regression between FBSway and ordinal age (1 to young, 2 to middle age, and 3 to elderly).

> fit <- lm(FBSway~as.numeric(Age), balance)

> summary(fit)

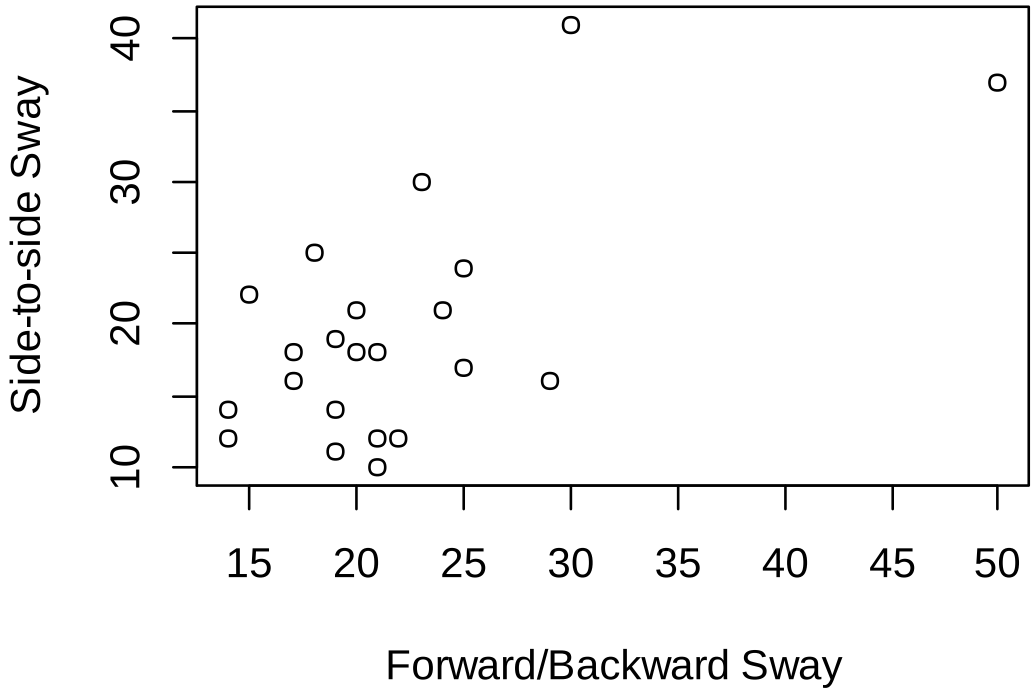
```
Call:
lm(formula = FBSway ~ as.numeric(Age), data = balance)

Residuals:
   Min      1Q  Median      3Q     Max 
-6.898 -3.635 -1.332   2.635  24.102 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)   
(Intercept)       13.504      3.650    3.70  0.00142 **
as.numeric(Age)    4.131      1.640    2.52  0.02036 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.751 on 20 degrees of freedom
Multiple R-squared:  0.241,     Adjusted R-squared:  0.203 
F-statistic: 6.349 on 1 and 20 DF,  p-value: 0.02036
```

- We reject the null hypothesis at the significance level of 0.05 and conclude that forward/backward sway is associated with age.

Get back to Data exploration

Q. How is side-to-side sway correlated with forward/backward sway?

```
> plot(balance$SideSway~balance$FBSway, xlab='Forward/Backward Sway', ylab='Side-to-side Sway')
```



With 2 variables,

1. Between continuous and categorical variable: box plot can be used
2. Between 2 continuous variables: a scatterplot can be used

B. Data set on lung function

We are interested in examining the response to inhaled ozone and sulphur dioxide among adolescents suffering from asthma. Listed in the Table are the initial measurements of forced expiratory volume in 1 second for 13 subjects involved in such a study. FEV1 is the volume of air that can be expelled from the lungs after one second of constant effort. To investigate the effect of pollutants on lung function, we might wish to know the typical value of FEV1 prior to the exposure to pollutants for the individuals in this group: **mean (SD) and median (Q1, Q3)**.

Table: Forced expiratory volumes in 1 second for 13 adolescents suffering from asthma

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13
FEV1	2.30	2.15	3.50	2.60	2.75	2.82	4.05	2.25	2.68	3.00	4.02	2.85	3.38

Numerical summaries

```
>fev1= c(2.30,2.15,3.50,2.60,2.75,2.82,4.05,2.25, 2.68, 3.00,4.02,2.85,3.38)

# Mean and Standard deviation
> mean(fev1)
> sd(fev1)
> var(fev1)
> sqrt(var(fev1))
```

$$Mean = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$Variance = \frac{\sum_{i=1}^{n}(x_i - mean)^2}{n-1}$$

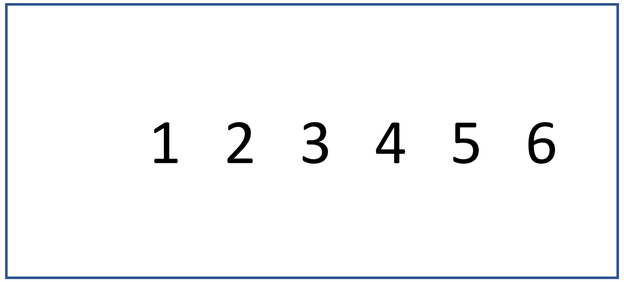
$$SD\ (standard\ deviation) = \sqrt{var}$$

```
# Median and interquartile range (IQR. Q1 and Q3)
> median(fev1)
> quantile(fev1, c(.25, .75))
```

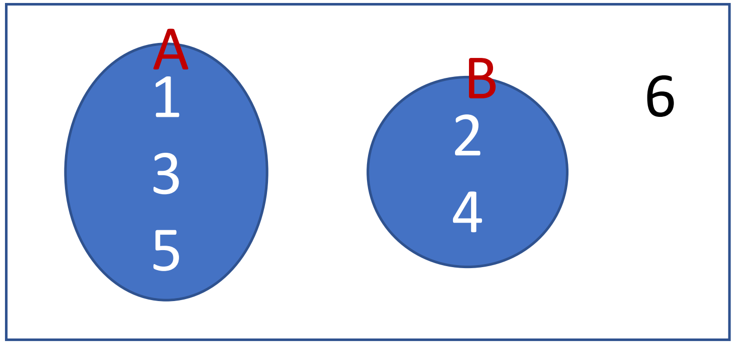
II. Event and Probability

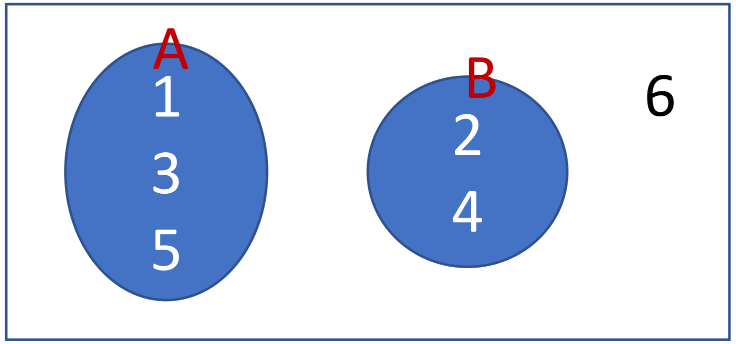
Suspect you a roll a dice once.

- A <u>sample space</u> is the set of all possible outcomes: {1,2,3,4,5,6}



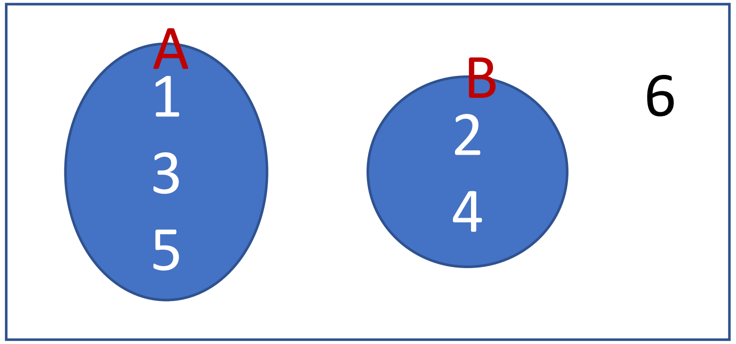
- An <u>event</u> is a set of outcomes of interest. Equivalently, it is a subset of sample space.
  - Event A: Getting an odd number. A={1,3,5}
  - Event B: Getting an even number less than 6. B={2,4}





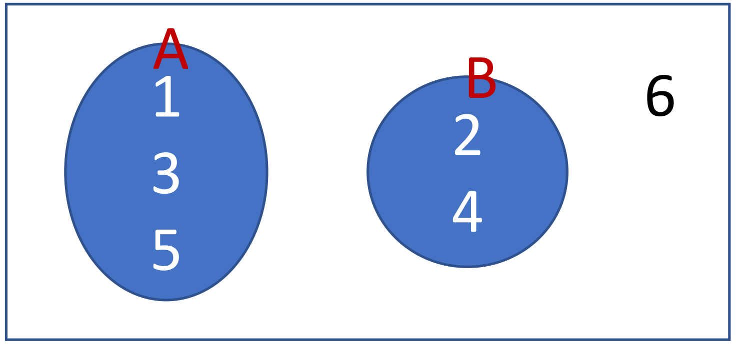
- A probability of an event: the number of outcomes in an event divided by the total number of outcomes in the sample space.
  - P(A) = 3/6= .5
  - P(B) = 2/6 = 1/3

- A complement of A is not A.
  - A^c = {2,4,6}
  - P(A^c) = 3/6=1-P(A)=1-3/6 **Complement rule**



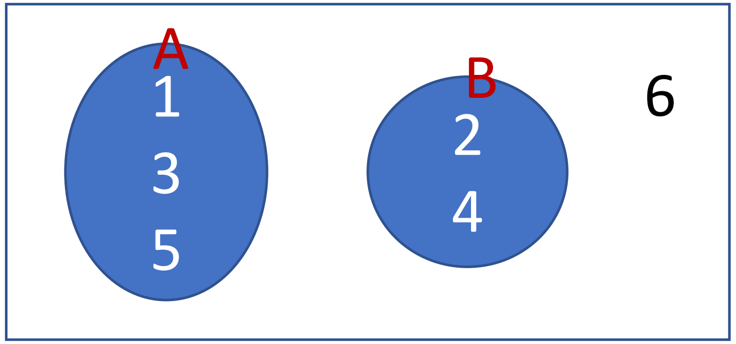
- An <u>intersection</u> is an event both A and B occur.
  - A ∩ B = nothing = ∅ : null event
  - P(A ∩ B) = 0/6 = 0

- A <u>union</u> is an event belonging to either A, B, or both.
  - A ∪ B = {1,2,3,4,5}
  - P(A ∪ B) = 5/6 = P(A)+P(B)-P(A ∩ B) = 3/6 + 2/6 – 0=5/6 **Additive rule**



- A <u>mutually exclusive</u> is that they cannot occur simultaneously (disjoint events).
  - Are A and B mutually exclusive? Yes. $P(A \cap B) = \emptyset$

- <u>Exhaustive</u>: the union of A and B covers all possible outcomes in a sample space.
  - Are A and B exhaustive? No. $P(A \cup B) \neq 1$



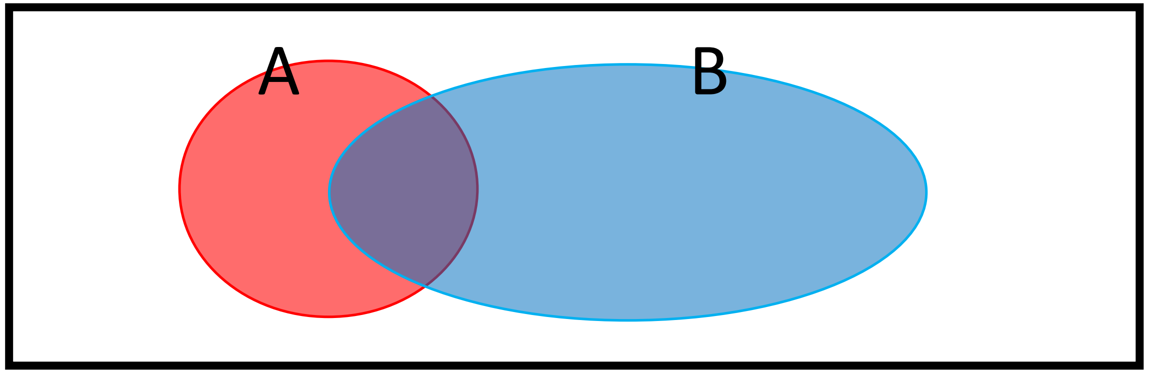
- <u>Independence</u>: Occurrence of one event does not affect the probability of occurrence of the other event. <u>$P(A \cap B)=P(A)P(B)$</u>
  - Is getting an odd number independent of getting an even number less than 6?

    => Are A and B independent?
    - $P(A \cap B)= 0$
    - $P(A)P(B) = 3/6 * 2/6 = 6/36$

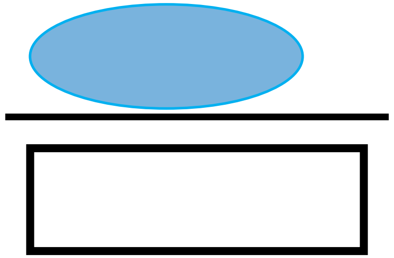
  - Suppose you are tossing a coin and tossing a dice. Is getting a head independent of getting an even number in a dice roll?
    - $P(A \cap B)= P(\{(H, 2), (H, 4), (H, 6)\}) = 3$ out of $12 = 1/4$
    - $P(A) = 1/2$, $P(B) =3/6=1/2$

Conditional probability : P(B|A)

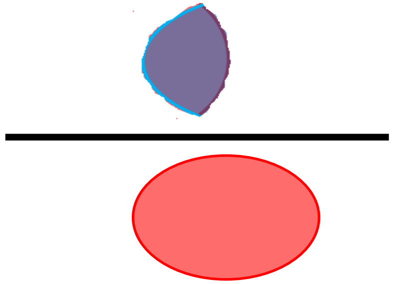
- P(B|A): A conditional probability of B given A is the probability that B occurs assuming that A has occurred. E.g. A probability of having a lung cancer if you are a smoker.
- $P(B|A) = \frac{P(B \cap A)}{P(A)}$ for P(A)>0



P(B) =



P(B|A)=



Bayes' theorem

- A Bayes' theorem is derived from conditional probability axioms.

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

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

$$= \frac{P(B|A)P(A)}{P(B \cap A) + P(B \cap A^c)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

P1. In 1992, of 4,065,014 registered births, 2,081,287 infants were boys and 1,983,727 were girls. Using this information, if a randomly selected woman were to become pregnant,

a. What is the probability that she would give birth to a boy?
   What is the probability that she would give birth to a girl?
   Are these events mutually exclusive, exhaustive, none, or both?
b. Suppose we randomly select two women from this population. What is the probability that they both give birth to boys?
c. Suppose we randomly select two women from this population. What is the probability that both children are boys, given that at least one child is a boy?

P1. In 1992, of 4,065,014 registered births, 2,081,287 infants were boys and 1,983,727 were girls. Using this information, if a randomly selected woman were to become pregnant,

a. What is the probability that she would give birth to a boy?

$$\frac{2081287}{4065014} = 0.512$$

What is the probability that she would give birth to a girl? Are these events mutually exclusive, exhaustive, none, or both?

$$\frac{1983727}{4065014} = 0.488$$

These events are mutually exclusive ($P(girl \cap boy) = 0$) and exhaustive ($P(girl \cup boy) = 1$).

P1. In 1992, of 4,065,014 registered births, 2,081,287 infants were boys and 1,983,727 were girls. Using this information, if a randomly selected woman were to become pregnant,

b. Suppose we randomly select two women from this population. What is the probability that they both give birth to boys?

Because of independence,

$$\left(\frac{2081287}{4065014}\right) \times \left(\frac{2081287}{4065014}\right) = 0.512 \times 0.512 = 0.262$$

P1. In 1992, of 4,065,014 registered births, 2,081,287 infants were boys and 1,983,727 were girls. Using this information, if a randomly selected woman were to become pregnant,

c. Suppose we randomly select two women from this population. What is the probability that both children are boys, given that at least one child is a boy?

Let A = both children are boys in randomly selected 2 women

B = at least one child is a boy in randomly selected 2 women

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

<u>Solutaion I</u>

$$P(A \cap B) = P(both\ are\ boys\ and\ at\ least\ one\ is\ a\ boy)$$

$$= P(both\ are\ boys) = P(A) = 0.512^2 = 0.262$$

$$P(B) = 1 - P(both\ are\ girls) = 1 - 0.488^2$$

$$P(A|B) = \frac{0.512^2}{1-0.488^2} = 0.355$$

P1. In 1992, of 4,065,014 registered births, 2,081,287 infants were boys and 1,983,727 were girls. Using this information, if a randomly selected woman were to become pregnant,

d. Suppose we randomly select two women from this population. What is the probability that both children are boys, given that at least one child is a boy?

Let A = both children are boys in randomly selected 2 women

B = at least one child is a boy in randomly selected 2 women

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

<u>Solution II</u>

$$\boldsymbol{P(B|A)P(A) = 1 \ * P(A) = 0.262}$$

$$P(B) = 1 - P(both\ are\ girls) = 1 - 0.488^2$$

$$P(A|B) = \frac{0.512^2}{1 - 0.488^2} = 0.355$$

P2. The world had been harmed by a widespread Z-virus, which already turned 10% of the world's population into zombies.

The scientists then invented a test kit with the sensitivity of 90% and specificity of 70%: 90% of the infected people will be tested positive while 70% of the non-infected will be tested negative.

If the test kit showed a positive result, what would be the probability that the tested subject was truly zombie?

P2. The world had been harmed by a widespread Z-virus, which already turned 10% of the world's population into zombies.

The scientists then invented a test kit with the sensitivity of 90% and specificity of 70%: 90% of the infected people will be tested positive while 70% of the non-infected will be tested negative.

If the test kit showed a positive result, what would be the probability that the tested subject was truly zombie?

Z: being a zombie

T^+: positive test result

$$P(Z|T^+) = \frac{P(T^+|Z)P(Z)}{P(T^+)} = \frac{P(T^+|Z)P(Z)}{P(T^+|Z)P(Z) + P(T^+|Z^c)P(Z^c)}$$

$$= \frac{0.9 * 0.1}{0.9 * 0.1 + 0.3 * 0.9} = \frac{0.09}{0.09 + 0.27} = \frac{0.09}{0.36} = 0.25$$

Probability Distribution

Binomial distribution

<u>What is it?</u>: It satisfies the following 4 conditions:

- A fixed number of trials (n)
- Each trial is independent of the others.
- There are only two outcomes (0, 1).
- The probability of each outcome remains constant from trial to trial.

➔ Out of **n** independent trials, the probability distribution of the number of successes (**x**).

<u>Probability distribution</u>: Binomial(n=n, p=p).

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

This is analogous to the binomial theorem formula: $1 = \left(p + (1-p)\right)^n = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x}$

Binomial distribution

Examples:

- Coin flips: the expected number of heads using n coins. Binomial(n=n, p=1/2)
- Asking 200 people if they support A vs B.

Mean: np

Standard deviation: $\sqrt{np(1-p)}$

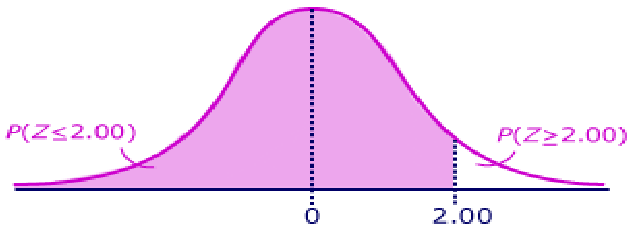
R code:

> dbinom(x=x, size=n, prob=p) ## P(X=x), the probability of obtaining x successes out of n trials when a probability of success in each trial is p.

> sum(dbinom(x=0:4, size=n, prob=p)) ## P(X$\leq x$), the probability of obtaining 0 ~ x successes out of n trials

> pbinom(x=x, size=n, prob=p) ## this gives P(X$\leq x$) as well.

Normal distribution



What is it?: Naively, a symmetric bell-shape curve.

Example:

- height distribution
- systolic blood pressure distribution
- SAT test score distribution

Importance:

- Many natural phenomena follow a normal distribution, e.g., height, measurement errors
- The **central limit theorem** (CLT): With large sample size n, the sample mean (X_n) will follow a normal distribution approximately, $\overline{X_n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Notation: $N(\mu, \sigma^2)$, Mean : μ, Standard deviation: σ

Normal distribution

<u>R code</u>:

> pnorm(x, mean=mu, sd=sd) ## P(X≤ x)

> 1-pnorm(x, mean=mu, sd=sd) ## P(X>x)

> pnorm(x1, mean=mu, sd=sd) - pnorm(x2, mu=mu, sd=sd) ## P(x2<X<x1) when x2<x1

> qnorm(q, mean=mu, sd=sd) ## q=P(X < ?)

> qnorm(.975, mean=0, sd=1) ## z-score where P(Z<z-score)=.975. This satisfies P(Z>z-score)+P(Z< -1*z-score) = .95

<u>Z-transformation</u>

This transforms a normal distribution into the standard normal distribution (mean=0, SD=1).

Z-score: $\boldsymbol{z = (x - \mu)/\sigma}$

P3. A certain drug treatment cures 90% of cases of hookworm in children. Suppose that 20 children suffering from hookworm are to be treated, and that the children can be regarded as a random sample from the population. Find the probability that

a. Exactly 75% will be cured.
b. At least 18 will be cured.
c. What is the mean and standard deviation of the number of cured cases of hookworm?

P3. A certain drug treatment cures 90% of cases of hookworm in children. Suppose that 20 children suffering from hookworm are to be treated, and that the children can be regarded as a random sample from the population. Find the probability that

a. Exactly 75% will be cured.

X: a number of children cured by the treatment

X ~ Binomial(n=20, p=0.9)

This is a binomial distribution with n=20 and p=0.9.

P(X=15) = $\binom{20}{15} \times 0.9^{15} \times 0.1^{5} = \frac{20!}{(20-15)! \times 15!} \times 0.9^{15} \times 0.1^{5}$

$= 0.032$

```
> dbinom(x=15,  size=20, prob=0.9)
```

P3. A certain drug treatment cures 90% of cases of hookworm in children. Suppose that 20 children suffering from hookworm are to be treated, and that the children can be regarded as a random sample from the population. Find the probability that

b. At least 18 will be cured.

X ~ Binomial(n=20, p=0.9)

P(X≥18) = P(X=18)+P(X=19)+P(X=20)

$= \binom{20}{18} \times 0.9^{18} \times 0.1^{2} + \binom{20}{19} \times 0.9^{19} \times 0.1^{1} + \binom{20}{20} \times 0.9^{20} \times 0.1^{0}$

= 0.677

```
> dbinom(18,  size=20, prob=0.9)+ dbinom(19,  size=20, prob=0.9)+dbinom(20, size=20, prob=0.9)
```

```
> sum(dbinom(18:20, size=20, prob=0.9))
```

P3. A certain drug treatment cures 90% of cases of hookworm in children. Suppose that 20 children suffering from hookworm are to be treated, and that the children can be regarded as a random sample from the population. Find the probability that

c. What is the mean and standard deviation of the number of cured cases of hookworm?

X ~ Binomial(n=20, p=0.9)

mean = np = 20*0.9 = 18

Standard deviation = $\sqrt{np(1-p)} = \sqrt{20 \times 0.9 \times (1-0.9)} = 1.34$

P4. The serum cholesterol levels of 17-year-olds follow a normal distribution with mean 176 mg/dLi and standard deviation 30 mg/dLi. What percentage of 17-year-olds have serum cholesterol values

a. 186 or more?
b. 156 or less?
c. Between 156 and 186?
d. Find the 80^{th} percentile

P4. The serum cholesterol levels of 17-year-olds follow a normal distribution with mean 176 mg/dLi and standard deviation 30 mg/dLi. What percentage of 17-year-olds have serum cholesterol values

a. 186 or more?

Let X be the serum cholesterol level.

X ~ N(μ=176, sd=30)

P(X $\geq$ 186)=?

Z-transformation: $P(Z \geq \frac{186-176}{30})$=P(Z ≥0.33)

=1-0.6293

= 0.37

> 1-pnorm($(186-176)/30$)

> 1-pnorm(186, mean=176, sd=30)

P4. The serum cholesterol levels of 17-year-olds follow a normal distribution with mean 176 mg/dLi and standard deviation 30 mg/dLi. What percentage of 17-year-olds have serum cholesterol values

b. 156 or less?

X ~ N(μ=176, sd=30)

P(X ≤156) = ?

Z-transformation: $P(Z \leq \frac{156-176}{30}) = P(Z \leq -0.67)$

$= 1 - P(Z \leq 0.67)$

$= 1-0.7486=0.2514$

```
> pnorm(156, mean=176, sd=30)
```

P4. The serum cholesterol levels of 17-year-olds follow a normal distribution with mean 176 mg/dLi and standard deviation 30 mg/dLi. What percentage of 17-year-olds have serum cholesterol values

c. Between 156 and 186?

$X \sim N(\mu=176, sd=30)$

$P(156<X <186) = P(X<186)-P(156<X)$

$= 0.6293-0.2514=0.3779$

```
> pnorm(186, mean=176, sd=30) - pnorm(156, mean=176, sd=30)
```

P4. The serum cholesterol levels of 17-year-olds follow a normal distribution with mean 176 mg/dLi and standard deviation 30 mg/dLi. What percentage of 17-year-olds have serum cholesterol values

d. Find the 80th percentile

X ~ N(μ=176, sd=30)

In the standard normal distribution, Z-score corresponding to 80th percentile from the table, which is 0.84

$0.84 = \frac{q-176}{30}$ ➔ q=201.2 mg/dLi, which is the 80th percentile.

> qnorm(.8, mean=176, sd=30)