자료 형태에 따른 통계 분석

분당서울대학교병원
응급의학과
김중희

(2021년도 대한재활의학회 추계학술대회 편집위원회 워크숍)

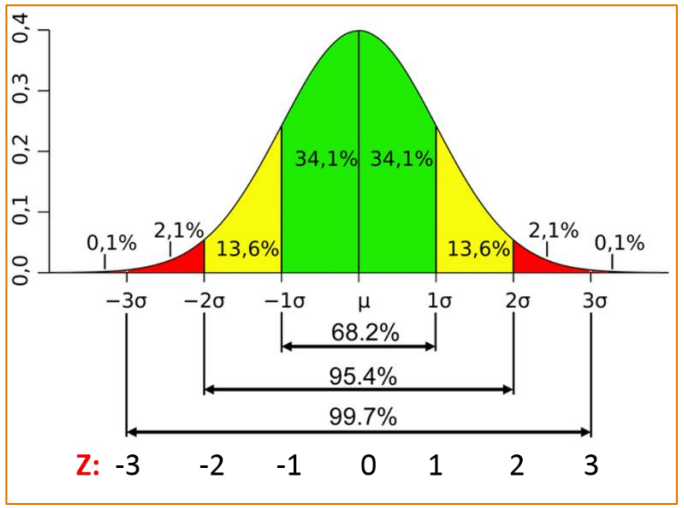
자료의 형태

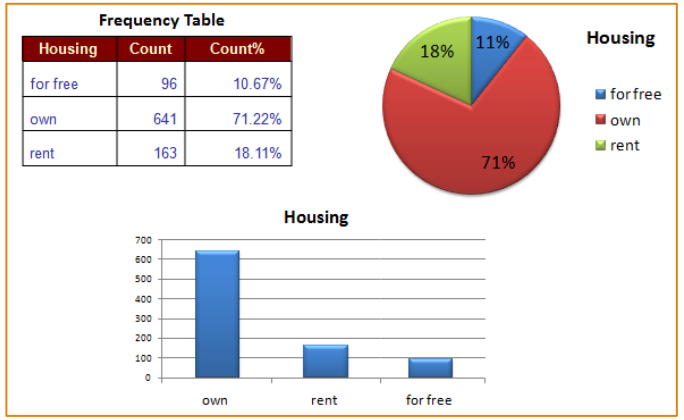
Numerical data – 나이, 키, 몸무게, 혈압 등

- Normally distributed?
- Equal variance?
- Paired?

Categorical data – 성별, 당뇨, 지역, 병기 등

- Ordered?
- Paired?





자료의 형태에 따른 통계 분석 방법

Description
- Numerical data
  - Mean and confidence interval
  - Median and interquartile range (IQR)
- Categorical data
  - Frequency and proportion

Comparison
- Numerical data
  - Parametric vs. Non-parametric test
- Categorical data

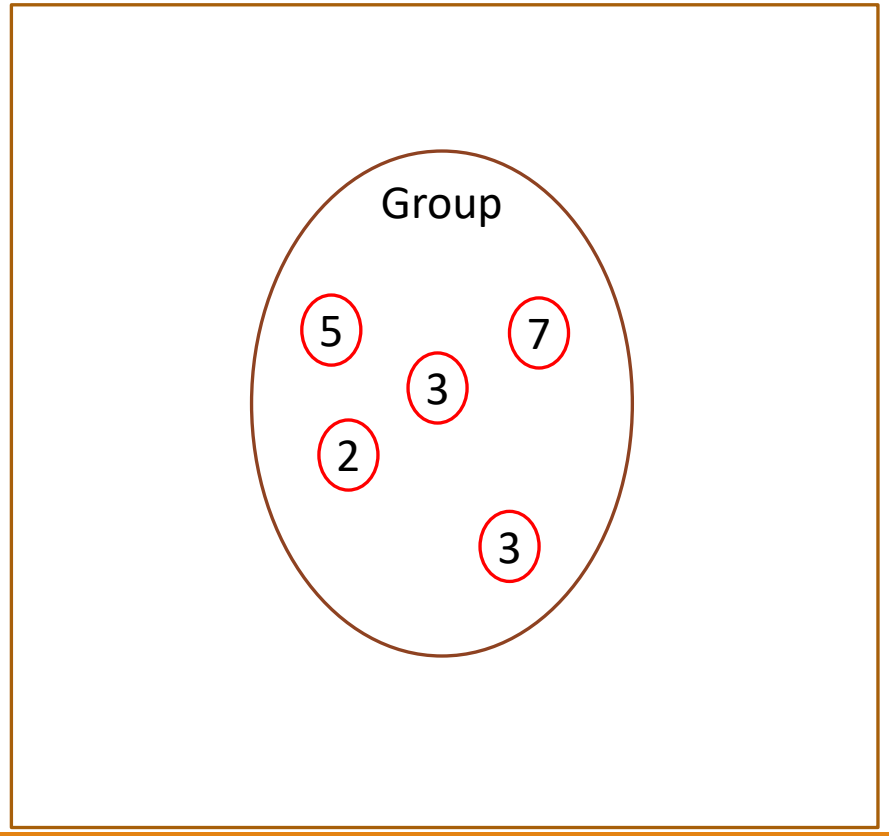
Predictive modelling (regression)
- Numerical output
  - (Multiple) Linear regression
- Binomial output
  - (Multiple) Logistic regression

	≤6 h	6-12 h	12-24 h	>24 h	
	(*N* = 15 640)	(*N* = 14 181)	(*N* = 10 304)	(*N* = 8516)	*P*
Age, categorised, N (%)					<.001
≤45 y	4251 (27.2%)	2721 (19.2%)	1802 (17.5%)	1115 (13.1%)	
46-65 y	4794 (30.7%)	4332 (30.5%)	3129 (30.4%)	2451 (28.8%)	
66-75 y	2989 (19.1%)	2932 (20.7%)	2225 (21.6%)	1976 (23.2%)	
76-85 y	2688 (17.2%)	3080 (21.7%)	2319 (22.5%)	2177 (25.6%)	
>85 y	918 (5.9%)	1116 (7.9%)	829 (8.0%)	797 (9.4%)	
Age, median (IQR)	61.0 (44.0-75.0)	66.0 (50.0-77.0)	67.0 (52.0-78.0)	69.0 (55.0-79.0)	<.001
Gender, N (%)					.052
Female	7219 (46.2%)	6551 (46.2%)	4671 (45.3%)	3795 (44.6%)	
Male	8421 (53.8%)	7630 (53.8%)	5633 (54.7%)	4721 (55.4%)	
Elixhauser comorbidities, *N* (%)					
Congestive heart failure	826 (5.3%)	928 (6.5%)	579 (5.6%)	564 (6.6%)	<.001
Cardiac arrhythmia	1405 (9.0%)	1087 (7.7%)	671 (6.5%)	655 (7.7%)	<.001
Valvular disease	271 (1.7%)	260 (1.8%)	149 (1.4%)	152 (1.8%)	.117
Pulmonary circulation disorders	173 (1.1%)	231 (1.6%)	193 (1.9%)	229 (2.7%)	<.001
Peripheral vascular disorders	548 (3.5%)	457 (3.2%)	326 (3.2%)	284 (3.3%)	.415

자료 형태에 따른 통계적 비교

One-Group *t*-test

H_0: Population mean = 3



알려진 평균은 ~인데, 이 그룹의 평균은 이를 벗어나는지?

Parametric test

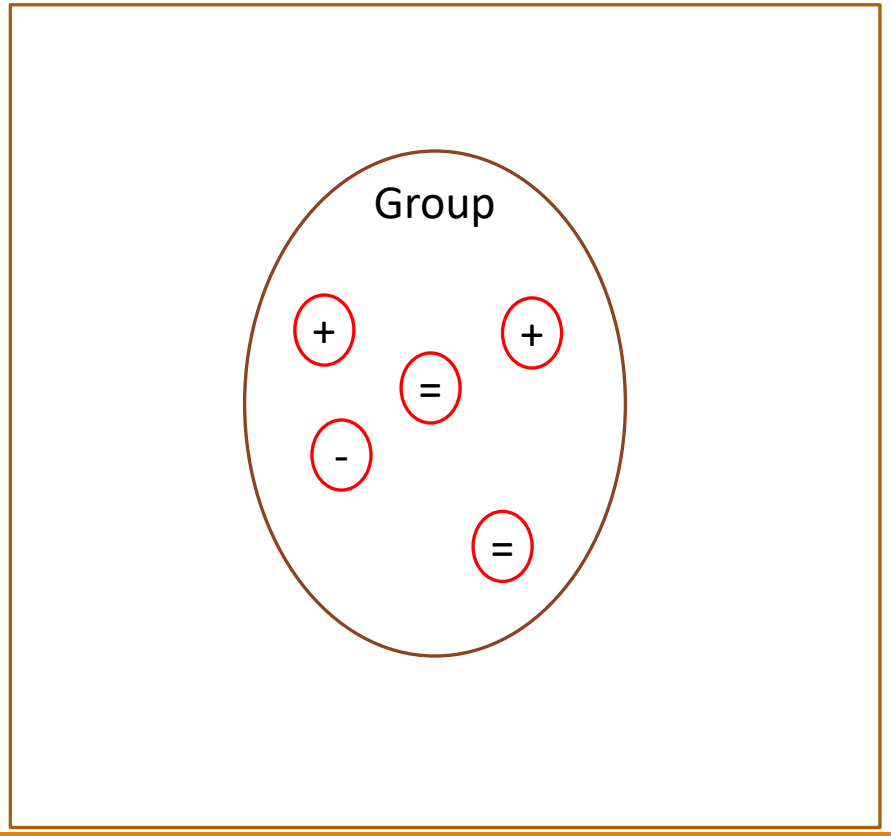
- 정규분포 가정이 필요

R example:

```
t.test(age, mu = 50)
```

Sign test

H_0: Population median = 3



알려진 평균은 ~인데, 이 그룹의 중앙값은 이를 벗어나는지?

Non-parametric test

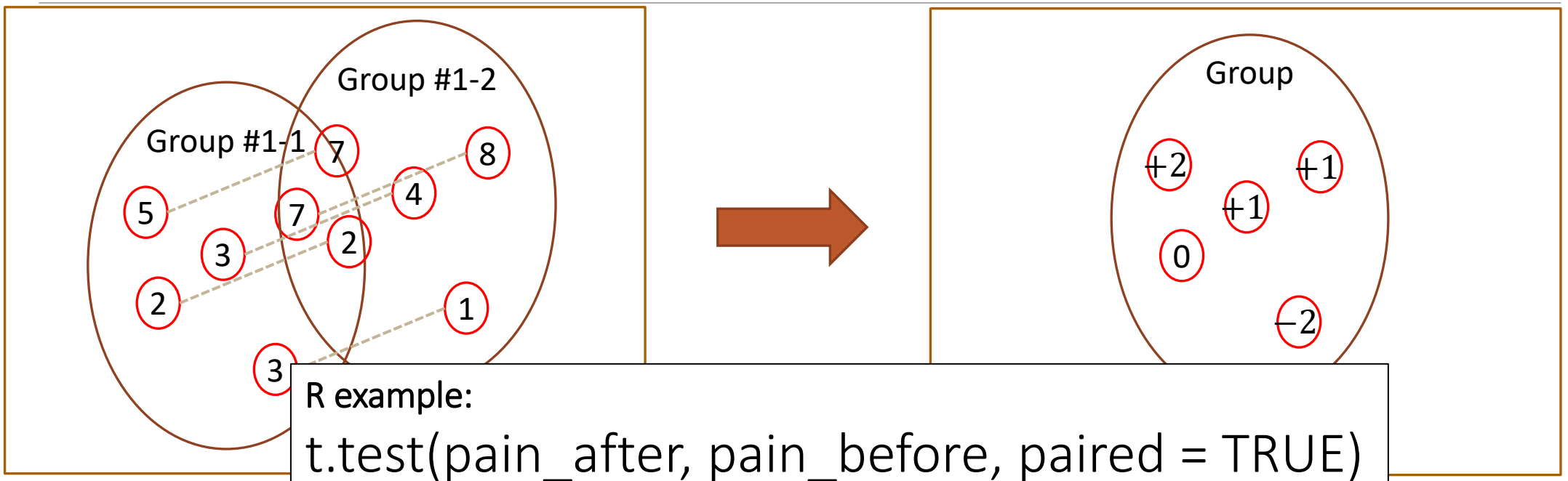
- 정규분포 가정이 필요하지 않음
- Sign (+/-)만 고려 함
- Wilcoxon signed ranks test가 더 강력함 (sign + rank 고려)

R example:

```
library(BDSA)
SIGN.test(age, md = 50)
```

Paired *t*-test

H_0: Mean difference = 0



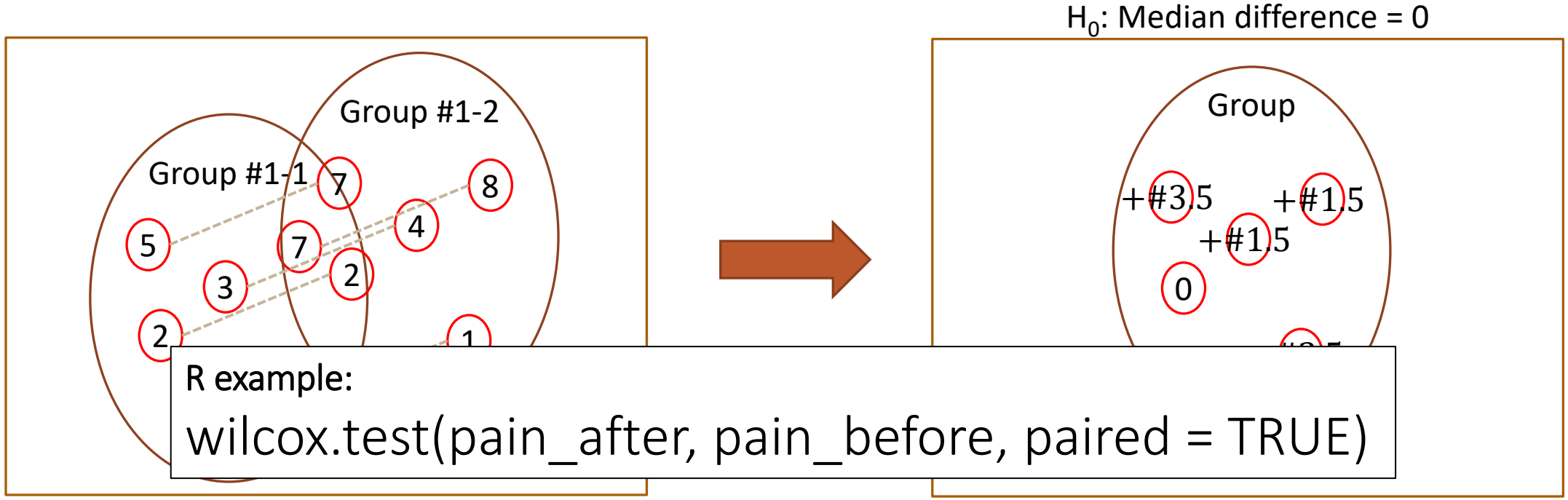
R example:

```
t.test(pain_after, pain_before, paired = TRUE)
```

같은 환자 그룹에서 치료 전후 평균이 어떻게 변했는지? 혹은 쌍을 맺은 case 및 control 사이의 평균 차이가 있는지?

각 쌍의 차이를 모아서 평균이 0과 차이 나는지 확인

Wilcoxon signed ranks test



R example:

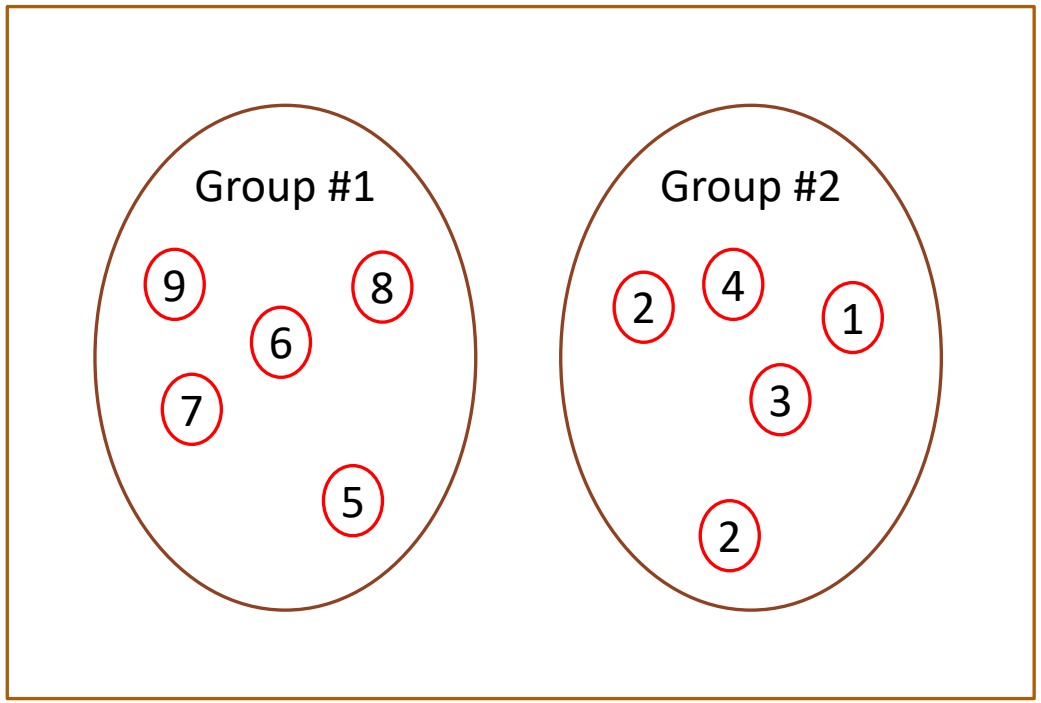
```
wilcox.test(pain_after, pain_before, paired = TRUE)
```

같은 환자 그룹에서 치료 전후 중앙값이 어떻게 변했는지? 혹은 쌍을 맺은 case 및 control 사이의 중앙값 차이가 있는지?

+rank들과 –rank들을 모아서 비교함

Unpaired (two-Group) *t*-test

H_0: Population mean is equal



독립된 두 그룹의 평균 비교

- 예: 응급실에 걸어 온 환자와 구급차 타고 온 환자의 나이가 차이 나는지?

정규분포, 등분산성 충족 필요함

R example:

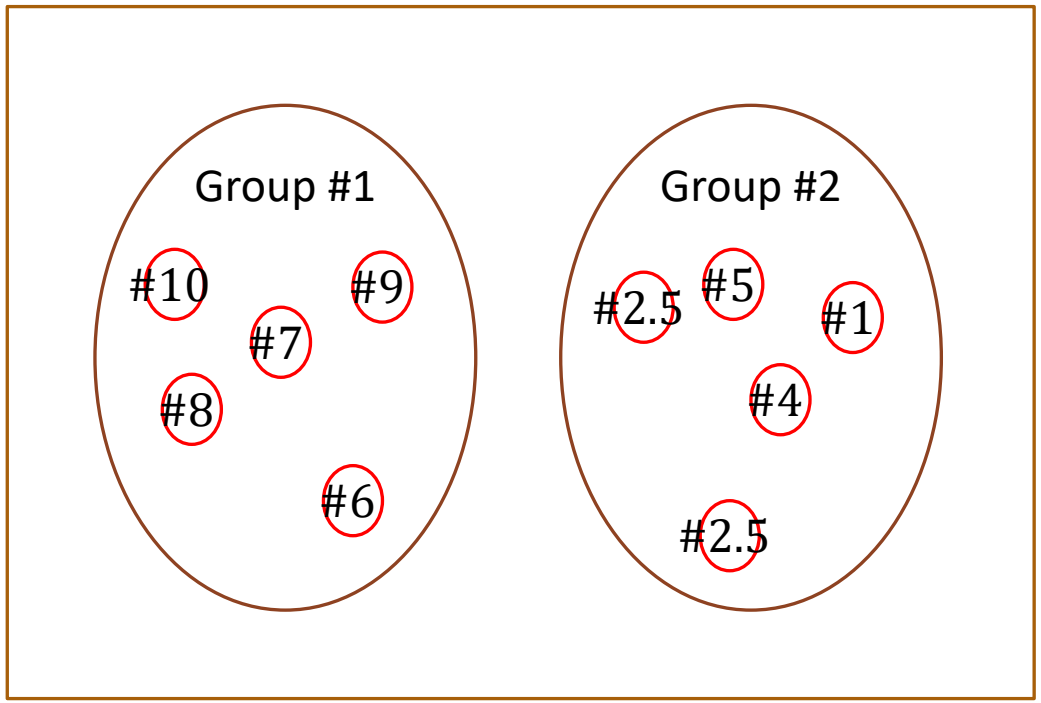
```
t.test(x=groupA, y=groupB)
```

Wilcoxon rank sum (two-Group) test

독립된 두 그룹의 중앙값 비교

Non-parametric test
- Rank 기반 비교

H_0: Population median is equal



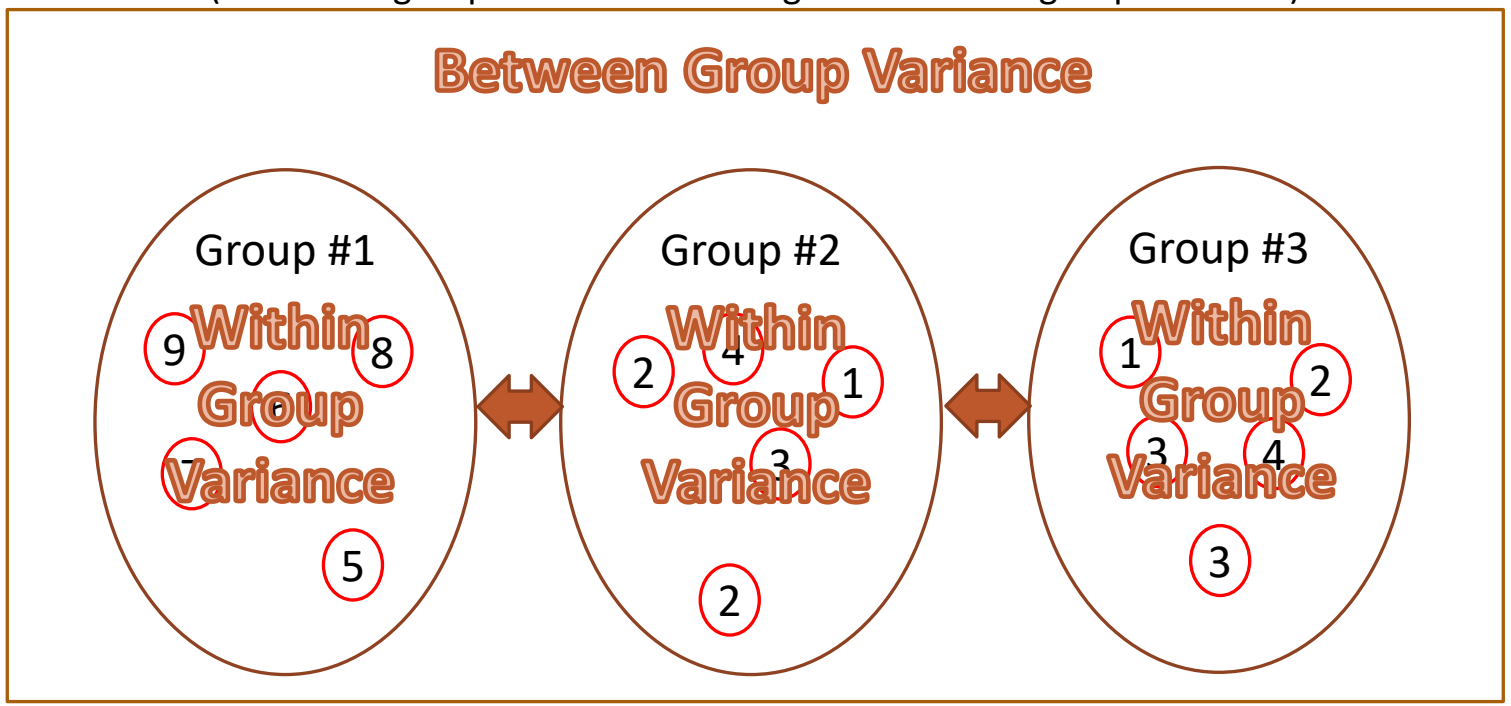
R example:

```
wilcox.test(x=groupA, y=groupB)
```

One-way analysis of variance (ANOVA)

H_0: Group means are equal

(=between group variance is not larger than within group variance)



3개 이상의 그룹을 비교 → 서로 다른 지 알려줌 (무엇이 다른 지는 알려주지 못함)

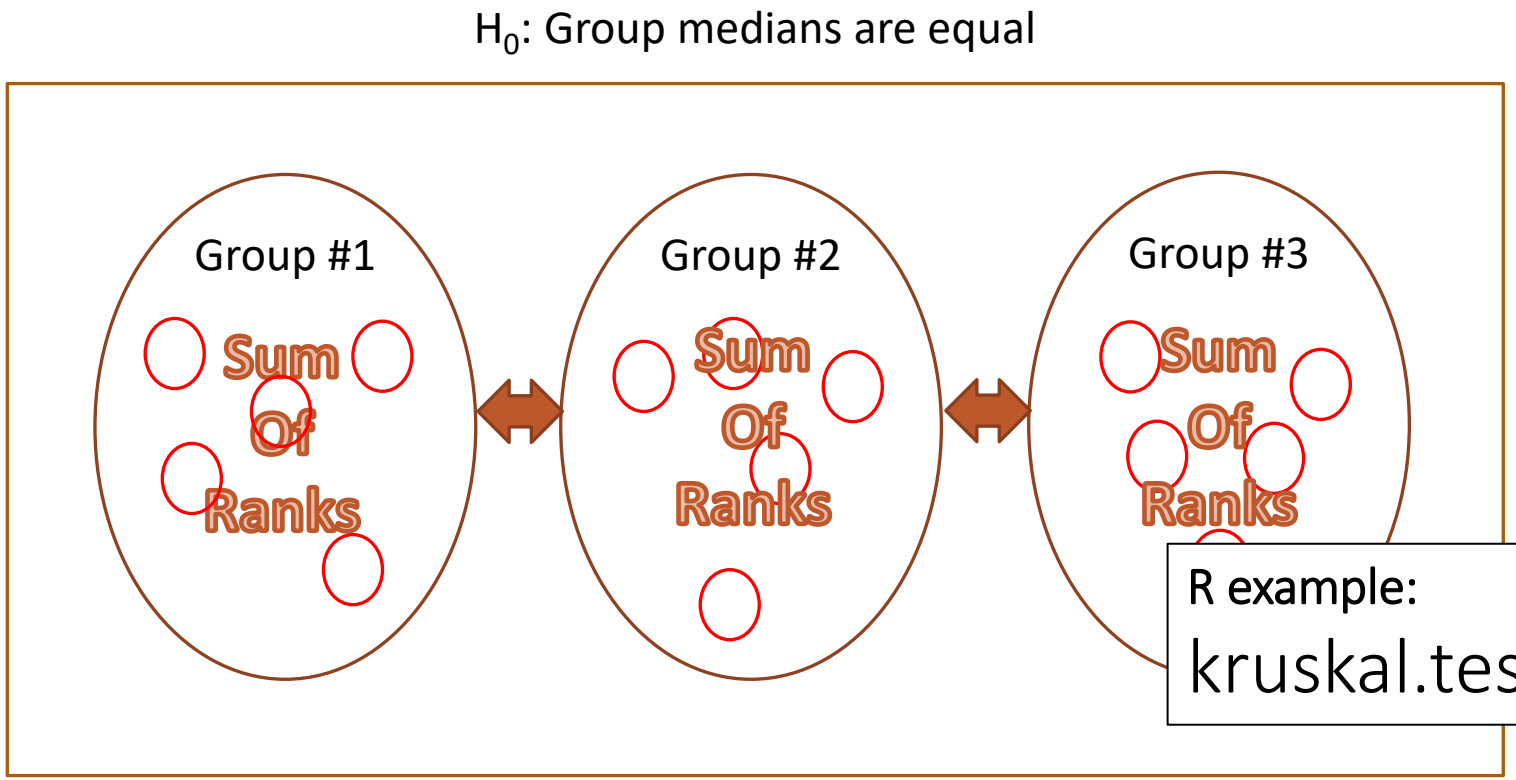
정규분포, 등분산성 충족 필요함

각 쌍 별로 post-hoc 비교

R example:

```
aov(Weight ~ Diet_group)
```

Kruskal-Wallis test



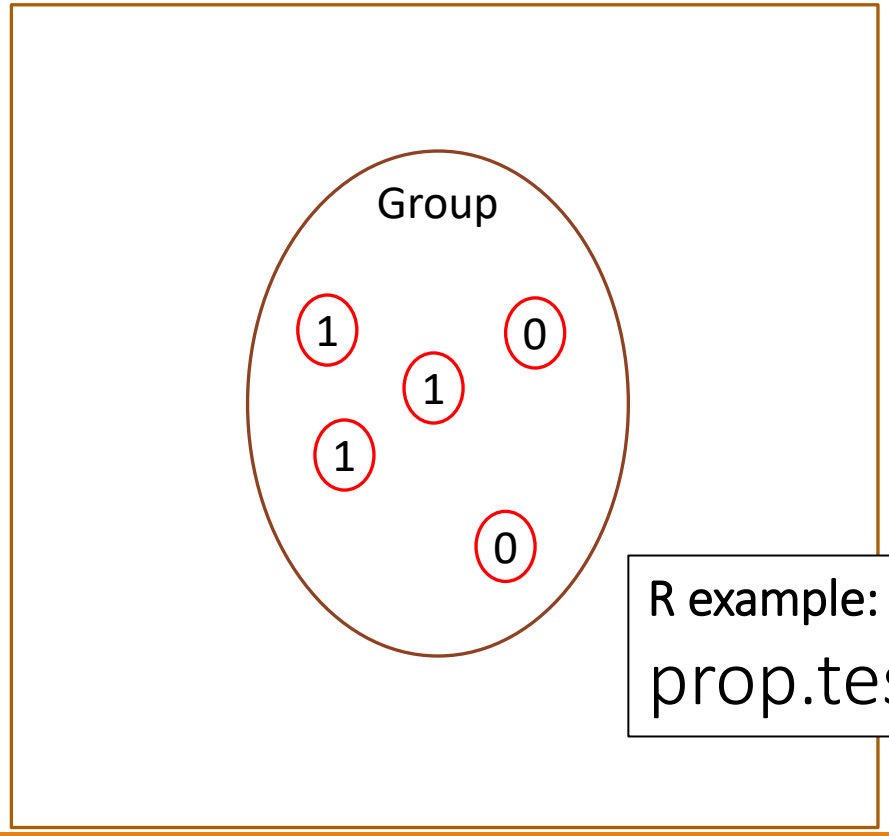
- 3개 이상의 그룹을 비교 → 중앙값이 다른 지 알려줌
- 정규분포, 등분산성 충족 필요하지 않음
- 각 그룹을 비교하기 위해 각 쌍 별로

R example:

```
kruskal.test(Weight ~ Diet_group)
```

Single proportion

H_0: Population proportion = 0.6



알려진 비율은 ~인데, 이 그룹의 평균은 이를 벗어나는지?

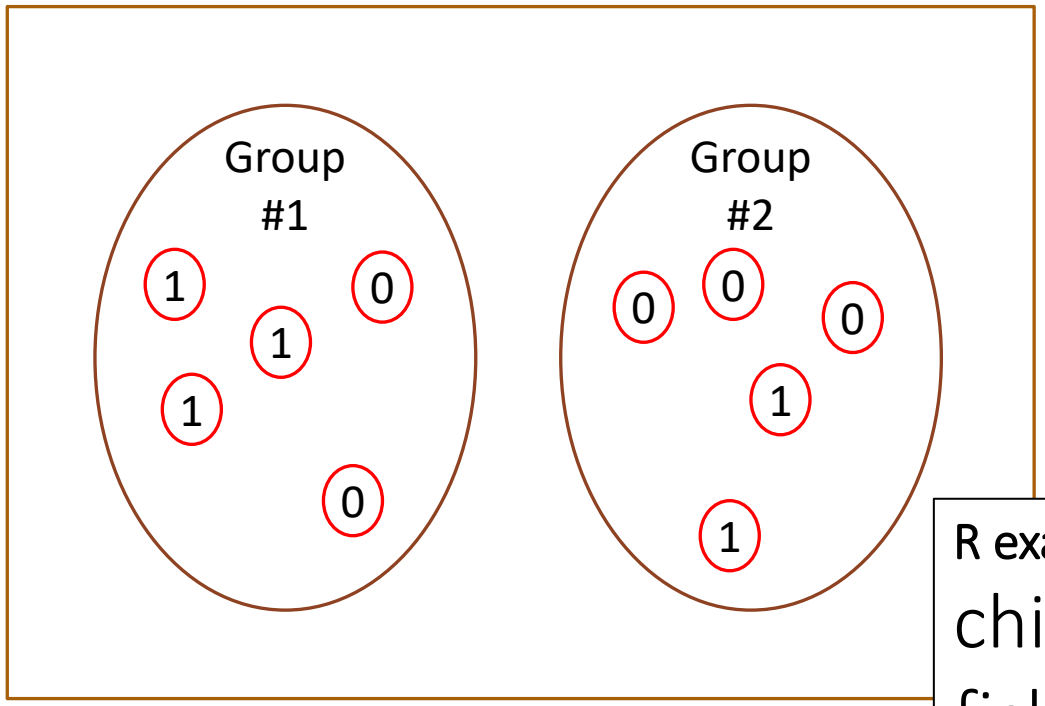
예: 연구 그룹의 고혈압 환자 비율이 우리나라 고혈압 환자의 비율과 같은가?

R example:

```
prop.test(sum(survived), length(survived), p = 0.5)
```

Chi-squared test

H_0: Population proportion is equal



Characteristic	Group 1	Group 2	Total
Present	a	b	$a + b$
Absent	c	d	$c + d$
Total	$n_1 = a + c$	$n_2 = b + d$	$n = a + b + c + d$
Proportion with characteristic	$p_1 = \frac{a}{n_1}$	$p_2 = \frac{b}{n_2}$	$p = \frac{a+b}{n}$

두 독립된 그룹 사이의 비율을 비교함

예: 두 처치 그룹 사이의 사망율이 동일한지?

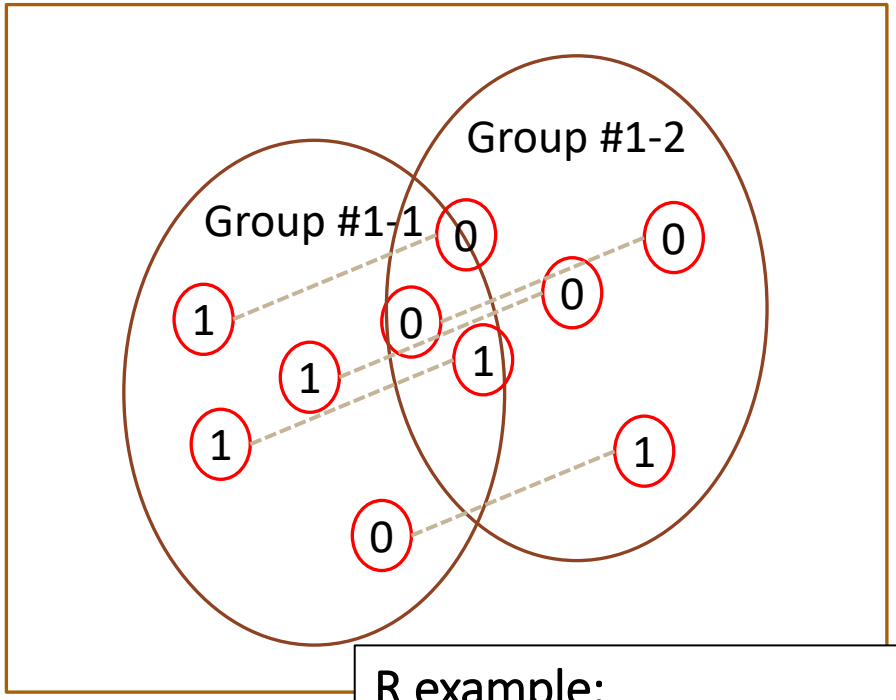
Fisher's exact test

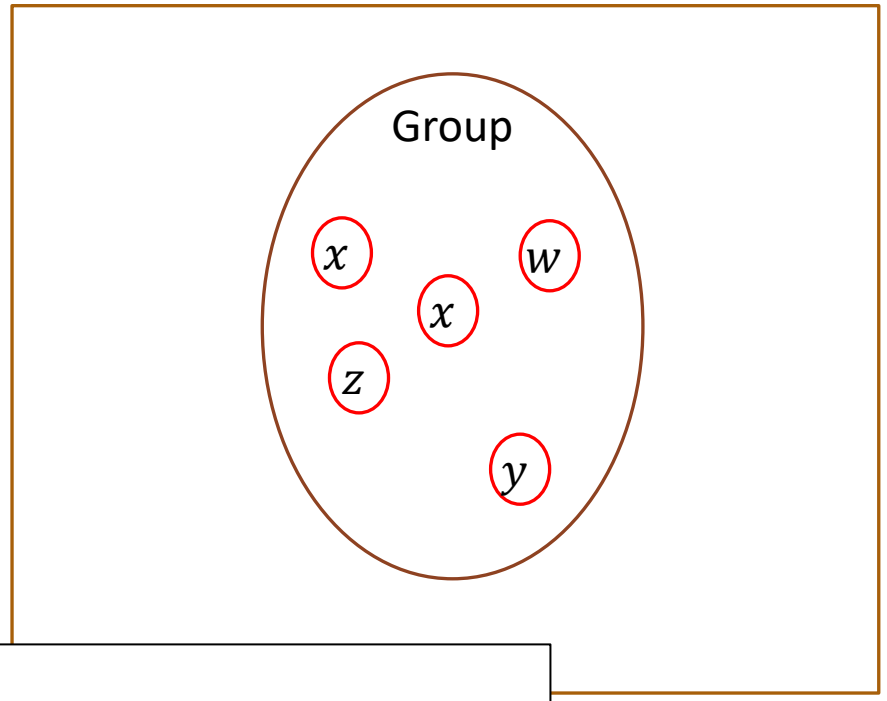
R example:

```
chisq.test(table(survived, treated))
fisher.test(table(survived, treated))
```

McNemar's test

H_0: $p_w+p_x=p_w+p_y$ and $p_y+p_z=p_x+p_z$ → $p_x=p_y$





R example:

```
mcnemar.test(table(first_test, second_test))
```

짝지어진 그룹

- 예: 같은 환자 그룹 내 MRI 이상 유무
  Sensitivity, Specificity 차이
  (주의: PPV, NPV는 안됨)

		1	
			Total no. of pairs
Present	w	x	$w + x$
Absent	y	z	$y + z$
Total	$w + y$	$x + z$	$m = w + x + y + z$

Chi-squared test (larger tables)

2 by 2 테이블 형태보다 더 큰 테이블에서도 사용할 수 있음.

Row categories	**Col 1**	**Col 2**	**Col 3**	**...**	**Col *c***	**Total**
Row 1	f_{11}	f_{12}	f_{13}	...	f_{1c}	R_1
Row 2	f_{21}	f_{22}	f_{23}	...	f_{2c}	R_2
Row 3	f_{31}	f_{32}	f_{33}	...	f_{3c}	R_3
...	...	...	...	...	...	...
Row *r*	f_{r1}	f_{r2}	f_{r3}	...	f_{rc}	R_r
Total	C_1	C_2	C_3	...	C_c	n

Chi-squared test for trend

Characteristic	Col 1	Col 2	Col 3	...	Col k	Total
Present	f_{11}	f_{12}	f_{13}	...	f_{1k}	R_1
Absent	f_{21}	f_{22}	f_{23}	...	f_{2k}	R_2
Total	C_1	C_2	C_3	...	C_k	n
Score	w_1	w_2	w_3	...	w_k	

한그룹은 binary, 다른 그룹은 ordered category 일 때 사용

- 예: 처치 여부에 따라 1년 후 중증도 그룹별 분포 차이가 있는지?

R example:

```
table <- as.table(rbind(Pr=c(384, 536, 335), Ab=c(951, 869, 438)))
prop_trend_test(table)
```

자료 형태에 따른 회귀 분석

회귀식

- 단순 선형 회귀분석 : 변수가 1개인 경우

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- 다중 선형 회귀분석 : 변수가 여러개인 경우

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

Explanatory variables (설명 변수)

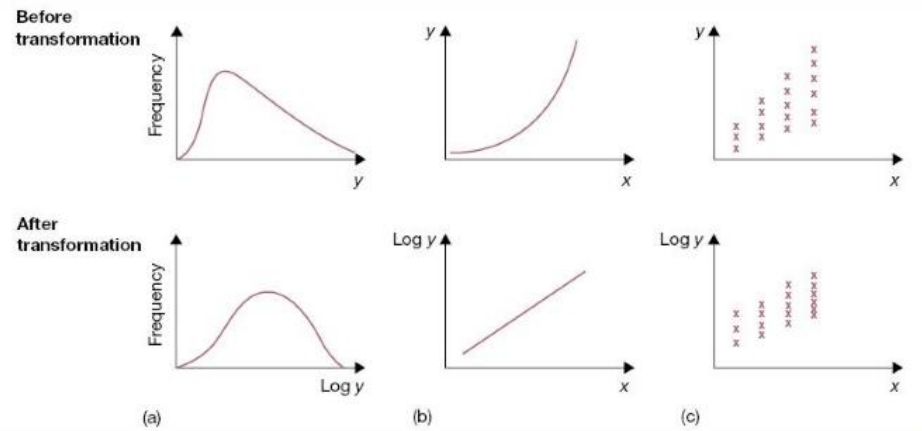
Nominal variables

Ordinal variables

- As numerical
- As categorical

Numerical variables

- Assessing the assumption of linearity
- How to deal with non-linearity
  - Binning
  - Transform
  - Polynomial regression



Outcome variables (결과 변수)

Number → Linear regression

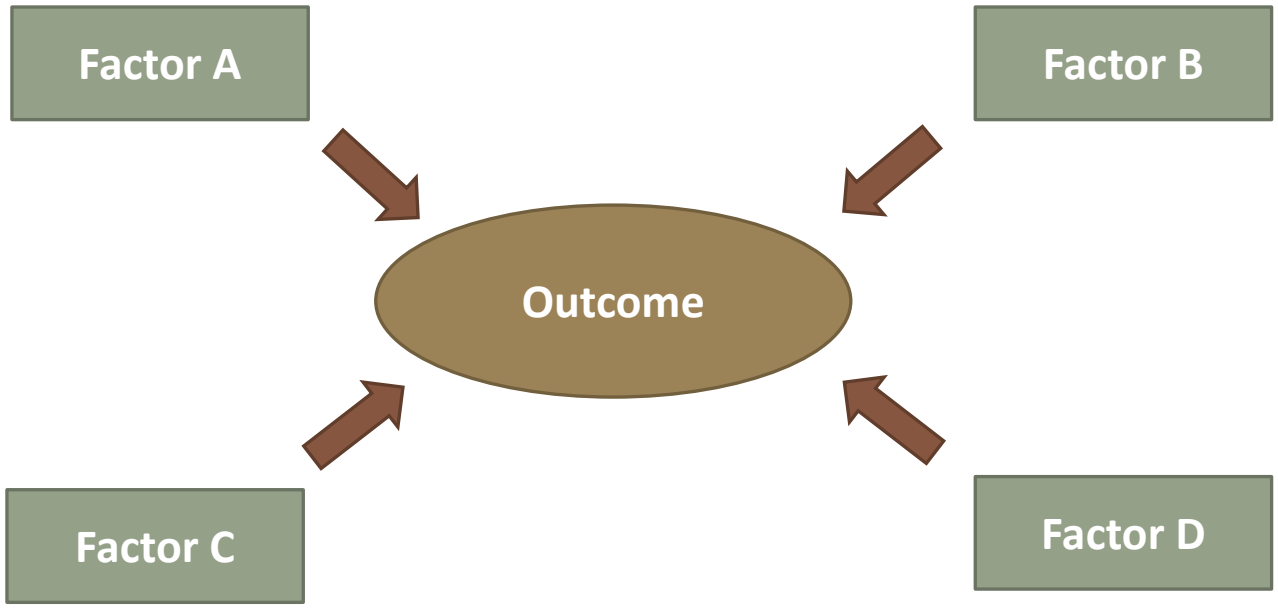
Binary outcome (Yes/No) → Logistic regression

Multiple class → Multinomial logistic regression

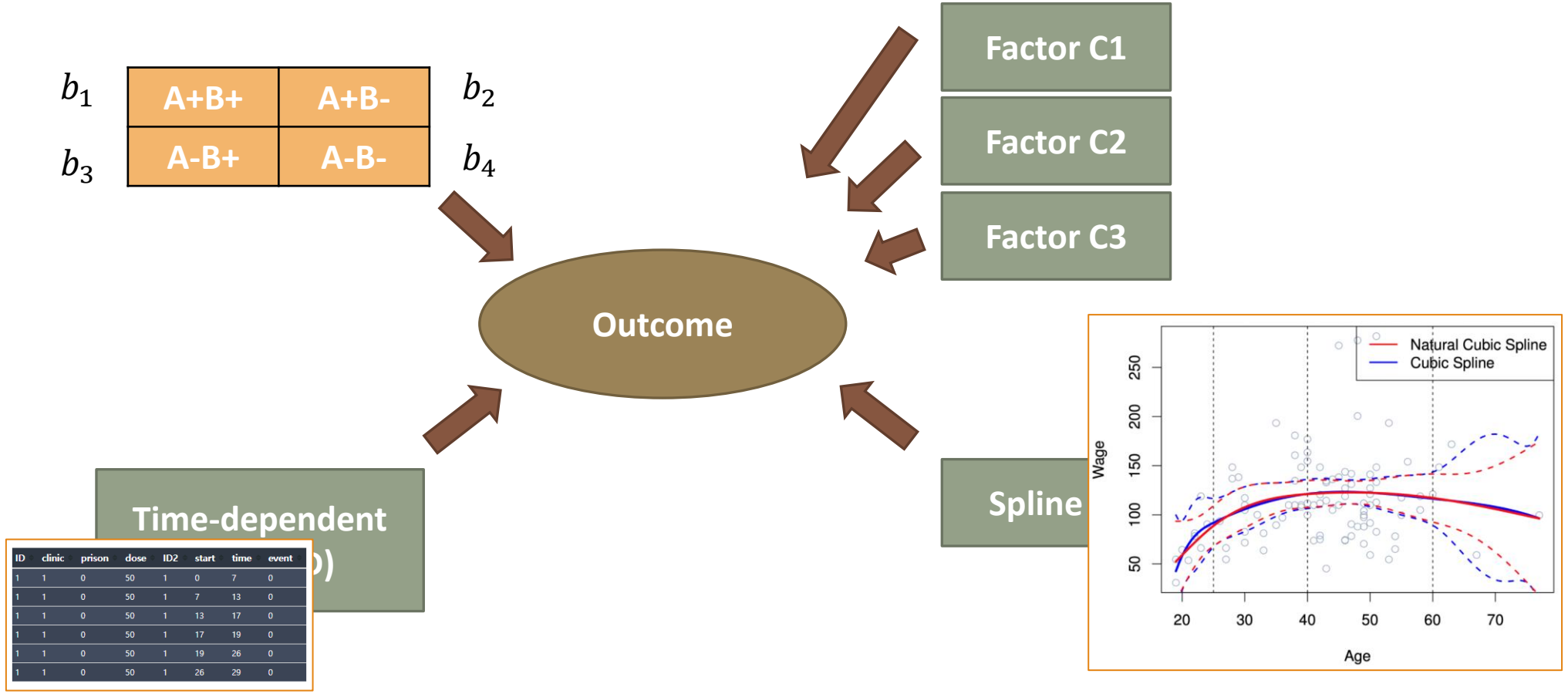
Rate → Poisson regression

Hazard → Cox regression

Model Structure in Regression

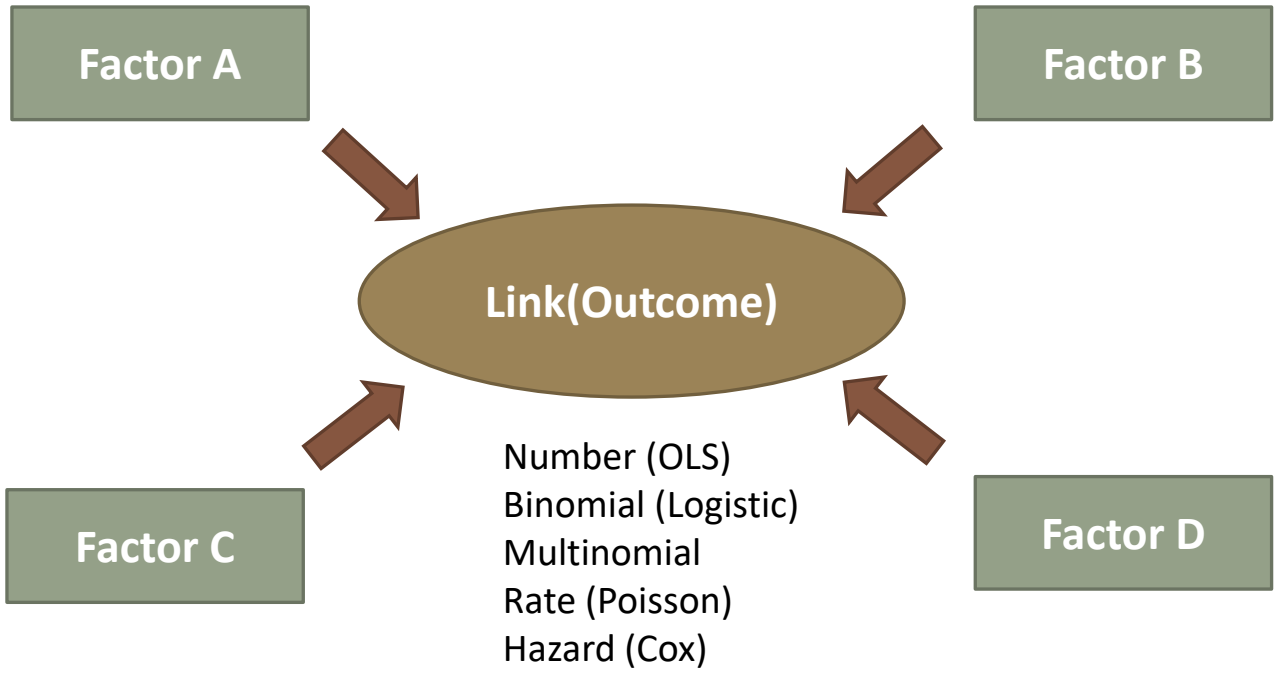


Variations in explanatory variables



ID	clinic	prison	dose	ID2	start	time	event
1	1	0	50	1	0	7	0
1	1	0	50	1	7	13	0
1	1	0	50	1	13	17	0
1	1	0	50	1	17	19	0
1	1	0	50	1	19	26	0
1	1	0	50	1	26	29	0

Variation in outcomes



Regression formula in R

```
ss_model =
  glm(septic_shock ~ age+sex+qsofa+wbc+hb+pt.inr+creatinine+albumin+crp,
      data=ed_infection,
      family = binomial)
```

Cubic Spline

```
ss_model =
  glm(septic_shock ~ bs(age, knots = 3)+
        sex+qsofa+wbc+hb+pt.inr+creatinine+albumin+crp,
      data=ed_infection,
      family = binomial)
```

Interaction

```
ss_model =
  glm(septic_shock ~ sex+age*qsofa+wbc+hb+pt.inr+creatinine+albumin+crp,
      data=ed_infection,
      family = binomial)
```

Interaction

```
ss_model =
  glm(septic_shock ~ (sex+age+qsofa)^2+wbc+hb+pt.inr+creatinine+albumin+crp,
      data=ed_infection,
      family = binomial)
```

```
              0.498338   0.783887   0.636 0.524955
age           0.036884   0.009655   3.820 0.000133 ***
qsofa         2.615147   0.497180   5.260 1.44e-07 ***
wbc          -0.002663   0.006156  -0.433 0.665275
hb            0.023619   0.036429   0.648 0.516745
pt.inr        0.154422   0.076806   2.011 0.044373 *
creatinine    0.128248   0.043895   2.922 0.003481 **
albumin      -0.572949   0.139742  -4.100 4.13e-05 ***
crp           0.028913   0.008816   3.280 0.001040 **
sexM:age     -0.011809   0.010412  -1.134 0.256717
sexM:qsofa    0.110587   0.185512   0.596 0.551099
age:qsofa    -0.017487   0.006385  -2.739 0.006166 **
```

결론

자료의 형태와,
분석의 목적에 따라,
적절한 통계 분석을 적용하자!

감사합니다!