논문작성을 위한 통계이론과 실제

한국외국어대학교
이준규

7/28-29. 2021.
전남대 영어영문학과 BK 사업단

Quantitative research

- **Quantifying** all the data (covert into numbers) > inference & interpretation

- Confirmatory (Hypothesis testing)
  - RQ = Yes/No questions
  - Related? Different?
    - e.g., H1 = The Output-focused group will outperform the input-focused group

- Research design + Inferential stats >> predetermined

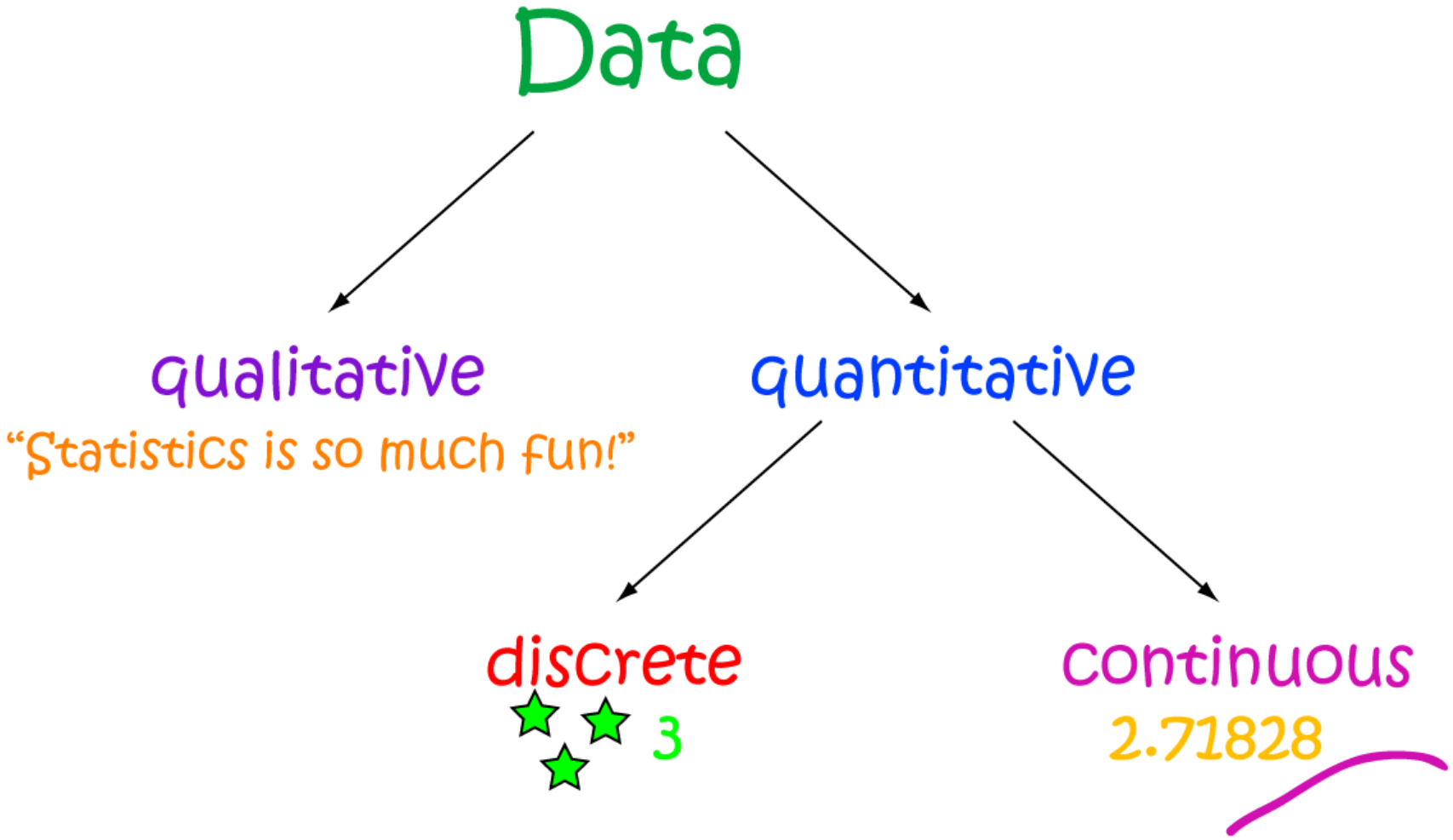
Quantifying data: creating matrix

Participants	Group	Pre-test	Post-test	WM

Quantifying data: Discrete or Continuous

Participants	Group	Pre-test	Post-test	WM
1	1	2.4	6.3	15
2	1	4.4	5.9	12
3	2	3.2	4.1	10
4	1	3.6	7.8	9
5	2	2.8	4.1	17
6	2	1.9	2.8	19
7	2	3.3	3.6	6
…	…	…	…	…
100	1	1.8	5.5	14

Data types


Data
qualitative
"Statistics is so much fun!"
quantitative
discrete
3
continuous
2.71828

Data conversion: Unidirectionality

- 남 vs 여
- 토플점수
- Discrete –X-> Continuous
- Continuous → Discrete
- Which one is more informative/useful?

Parametric vs nonparametric test

- Normal distribution

Parametric	Nonparametic
Independent t-test	Mann-Whitney test
Paired t test	Wilcoxon Signed-Rank test
One-way ANOVA	Kruskal-Wallis test
One-way repeated measure ANOVA	Friedman's test
Pearson correlation	Spearman correlation

L2 data & operationalization

- L2 data
  - **Latent data, Invisible data**

- Operationalization (working definition) & Interpretability
  - e.g., Measuring implicit knowledge
    - retrospective verbal reports
    - direct and indirect tests
    - subjective measures
  - >> Either **discrete** or **continuous** coding + interpretation
  - (Refer to previous studies)

- (validity & reliability)

기술통계 (Decriptive statistics)

- **Description of your data**
  - Continuous data > Mean, Standard deviation, range
  - Discrete data > Frequency
  - **Meaningful description?**

추론통계 (Inferential statistics)

- 관계 (Relationship): **r**
  Correlation/Regression

- 차이/비교 (difference): **d**
  t-test, ANOVA

- **<u>중요성: 연구문제/연구설계와 직접 연관</u>**

Relationship

- Correlation
- Regression
- Factor analysis

상관분석을 이용해 답할 수 있는 연구문제

- 말하기 점수와 쓰기 점수가 관계가 있는가?
- 말하기 점수와 자신감 점수가 관계가 있는가?
- 말하기 점수와 불안감 점수가 관계가 있는가?

- **두 개의 연속형 변수 (two continuous variables)가 관계가 있는가?**

연속형 변수 (continuous variables)

- 예) 점수: 0에서 100점, 척도 (scale)에 표시할 수 있는 것)
- 가상 설문지
  - 다음 각 영역에 자신의 영어능력을 표시해 주세요 (0 = 가장 낮음, 9 = 가장 높음).

	0	1	2	3	4	5	6	7	8	9
말하기										
쓰기										
자신감										
불안감										

Data structure (Matrix)

참여자	말하기	쓰기	자신감	불안감
1	8	3	9	1
2	5	4	5	6
3	4	3	3	7
4	9	2	8	2
5	9	2	9	1
6	3	3	4	6
7	5	1	6	5
8	2	3	2	7
9	7	2	6	3
…				
100	9	9	9	0

Correlation coefficient (상관계수)

- [예시 연구문제]
  - 말하기 점수와 불안감 점수가 관계가 있는가?
  - 결과: $\boldsymbol{r}$ **= -.59,** $\boldsymbol{p}$ **= .01,** $n = 100$

- **두 개**의 연속형 변수 (two continuous variables)가 **관계**가 있느냐?

- 관계를 하나의 대표값으로 표시 → 상관계수 (correlation coefficient) = **r**
  - **Pearson's r** [두 변수가 정규분포를 보일때]
  - Spearman rho, Kendall's tau [두 변수가 정규분포를 보이지 않을때]

상관계수 must-know

- **Strength (강도)** → r (상관계수)
- **The r is statistically significant? → p-value**
- **Directionality (방향성): → positive (+)** or **negative (-)**

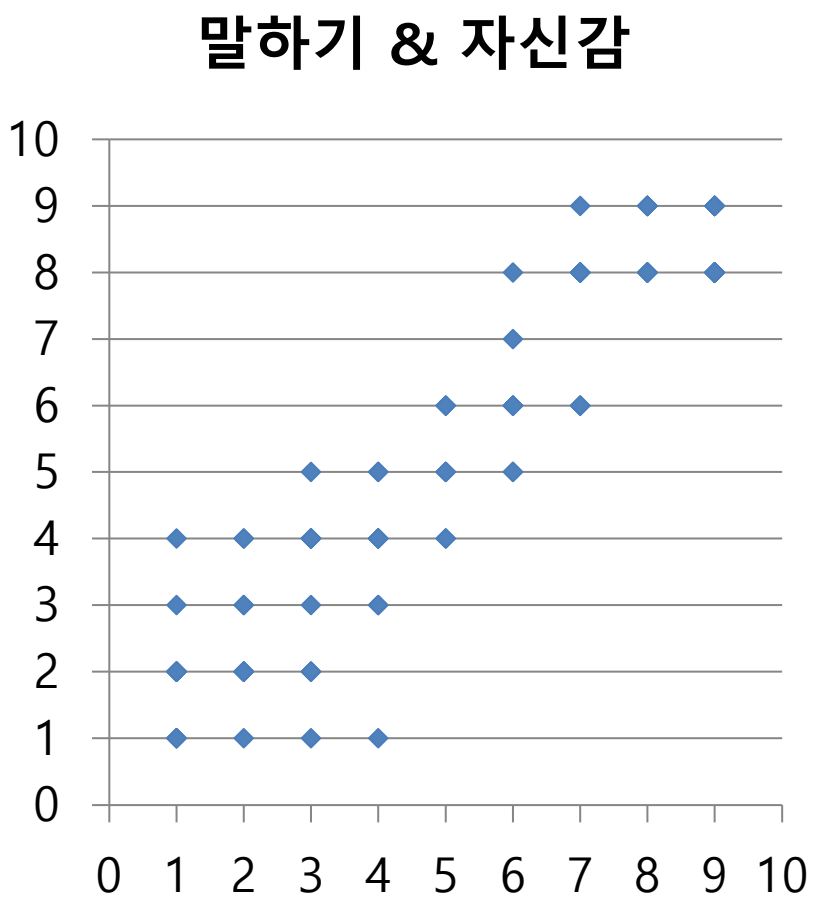
상관계수의 강도

- **Strength (강도)** → r (상관계수)
  - 상관계수 (r)는 절대값 0에서 1 사이의 숫자:
  - (예) *r* =.34, *r* =.67

0 = **no** relationship

1 = perfect relationship

상관관계 강도계산

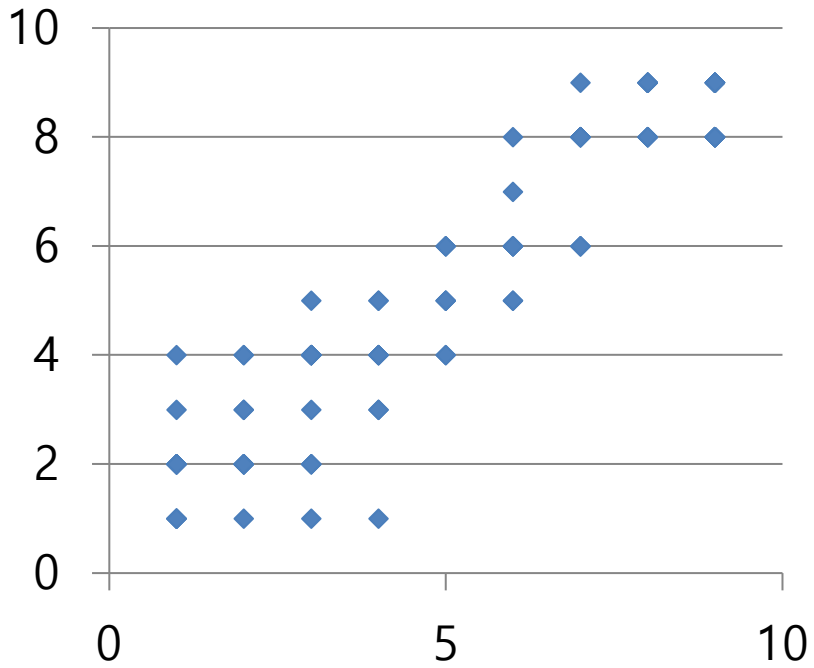


- 모든 data point를 설명할 수 있는 최단거리의 수직선 찾기
- **Y = aX + e**
  - a = r (상관계수/기울기)
  - e = error (unexplained)
- 결국
  - r이 1에 가까울 수록 통계적으로 유의미
  - 반면 r이 0에 접근할 수록 통계적으로 유임하지 않음

상관관계 강도

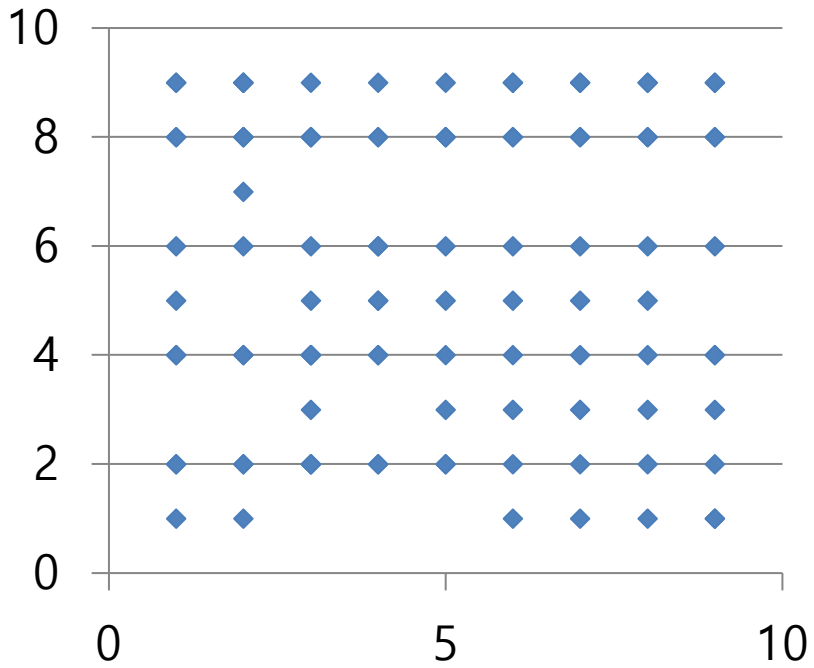
Strong relationship
r **이 1에 접근**

말하기 & 자신감



No relationship
r **이 0에 접근**

말하기 & 쓰기



상관계수가 통계적으로 유의미?

- **<u>The r is statistically significant?</u>**
  - → ***r*의 p-값이 .05보다 작으면, 통계적으로 중요/유의미**
  - →.05, .04, .03 … .01, .001 (유의미)
  - →.11, .12, .56, .67 (유의미하지 않음)

상관계수가 통계적으로 유의미?

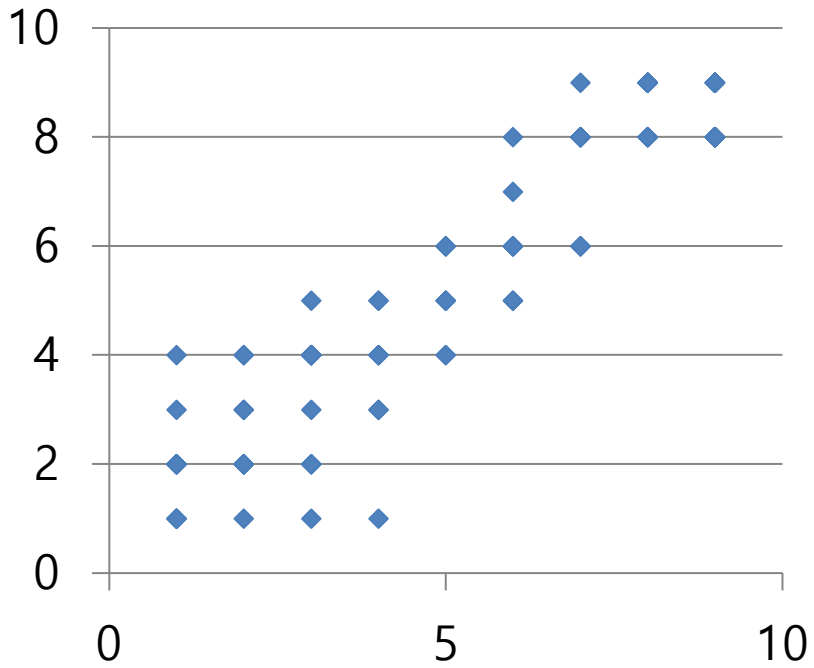
- p-값의 의미 (p = probability)
  - 상관관계분석의 영 (null) 가설: r = 0 (no relationship)
  - 상관관계분석의 연구가설: $r \neq 0$

- null 가설을 중심으로 p값을 생각 (즉, null가설이 맞을 확률)
  - p값이 0에 가까울 수록, r값이 0이 될 확률이 적다는 것.
  - 반대로 p값이 1에 가까울 수록, r값이 0이 될 확률이 크다는 것.
  - [예] $r = .90, p = .0001$ $r = .75, p = ,05$
    $r = .45, p = .45$ $r = .10, p = .89$

상관계수의 방향성

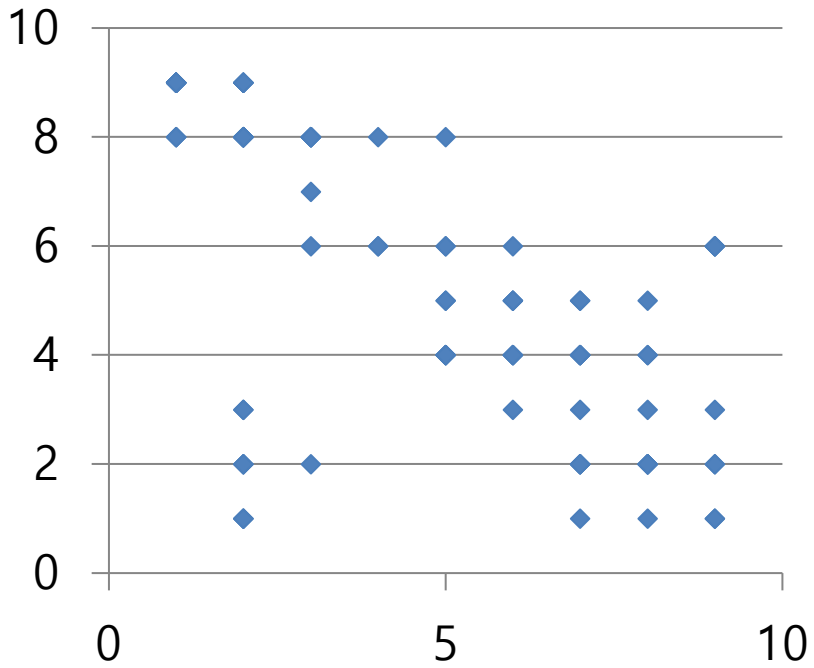
positive relationship

말하기 & 자신감



inverse relationship

말하기 & 쓰기



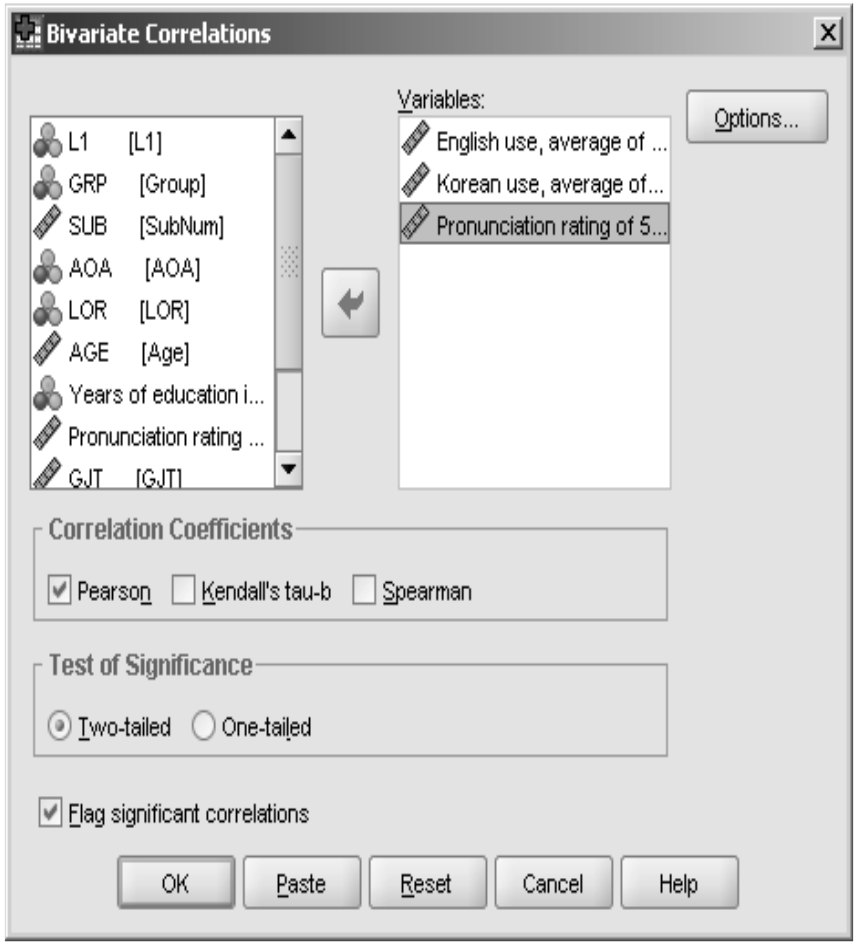
상관관계분석: Summary so far

- 말하기 점수와 쓰기 점수가 관계가 있는가?
  - 결과: $\boldsymbol{r}$ **= -.02,** $\boldsymbol{p}$ **= .85,** $n = 100$
- 말하기 점수와 자신감 점수가 관계가 있는가?
  - 결과: $\boldsymbol{r}$ **= .94,** $\boldsymbol{p}$ **= .0001,** $n = 100$
- 말하기 점수와 불안감 점수가 관계가 있는가?
  - 결과: $\boldsymbol{r}$ **= -.65,** $\boldsymbol{p}$ **= .0001,** $n = 100$

SPSS 실행

- 상단메뉴에서 …

분석(analysis)
→ 상관변수 (correlate)
→ 이변량 상관계수 (bivariate)



SPSS 결과물

		말하기	자신감
말하기	***Pearson* 상관계수**	***1***	***.936****
	유의확률 (양쪽)		***.000***
	N	***100***	***100***
자신감	***Pearson* 상관계수**	***.936****	***1***
	유의확률 (양쪽)	***.000***	
	N	***100***	***100***

**. 상관계수는 0.01 수준(양쪽)에서 유의합니다.

- 말하기 & 자신감
- 필요한 숫자
  - r 계수
  - p 값
  - Sample size (N)

상관분석 예시 결과작성

		말하기	자신감
말하기	Pearson 상관계수	1	.936**
	유의확률 (양쪽)		.000
	N	100	100
자신감	Pearson 상관계수	.936**	1
	유의확률 (양쪽)	.000	
	N	100	100

**. 상관계수는 0.01 수준(양쪽)에서 유의합니다.

- "두 번째 연구문제는 말하기 점수와 자신감 점수가 유의미한 관계가 있는지를 알아보는 것이었다.
- Pearson 상관관계 분석결과, 두 점수는 통계적으로 유의미한 양의 상관관계를 보였다 r= .94, p = .0001, n = 100."

상관분석 vs 회귀분석 1

- Good news
  - 수학적으로는 회귀분석과 correlation분석은 같음 (Y= aX + e [최단거리 수직선 찾기])

- **<u>목적의 차이</u>**
- 상관분석:
  - 두 변수가 **관계**가 있는가?
  - **[예시 연구문제] 말하기 점수와 불안감 점수가 관계가 있는가?**

- 회귀분석:
  - 하나의 변수가 다른 변수(들)을 **설명/예상**할 수 있는가?
  - **Explain/predict** → 반드시 이론/선행연구기반
  - Y= aX + e
  - Y can be explained/predicted by X
  - **[예시 연구문제] 말하기 점수가 불안감 점수에 의해 설명될 수 있는가?**

상관분석 vs 회귀분석 2

- **변수설정의 차이**
- Y= aX + e [최단거리 수직선 찾기]
- 상관분석:
  - X와 Y가 바뀌어도 상관없음 [단순히 두 변수가 관계가 있는가 탐색]
  - e.g., The relationship between X and Y
- 회귀분석
  - X와 Y가 결정됨 [이 결정은 반드시 이론, 선행연구에 기반해야 됨]
  - e.g. The effect of X on Y
  - BUT**인과관계 아님**!!! [NO causal relationship]

회귀분석 must-know

- Y= aX + error
  - **Y = 종속변수, dependent variable (DV), outcome variable**
  - **X = 독립변수, independent variable (IV), predictor variable**

회귀분석의 종류

- **단순회귀분석 [simple regression]**
- $Y = aX + e$

  X, 즉 독립변수 (IV)가 하나인 경우
  예) 말하기 = 불안감 + error

- **다중회귀분석 [multiple regression]**
- $Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n + e$

  X, 즉 독립변수 (IV)가 여러 개인 경우
  예) 말하기 = 불안감 + 쓰기 + 자신감 + error

  Note1. Y는 단순/다중 회귀분석 **모두 1개의 연속형 변수**
  Note2. IV는 연속형 (continuous), 범주형 (categorical) 상관 없음.

단순회귀분석 (simple regression)

- $Y = aX + e$
- 독립변수 (IV)가 하나인 경우
- 계산되는 계수 a가 결국 상관관계계수 (r)과 같음

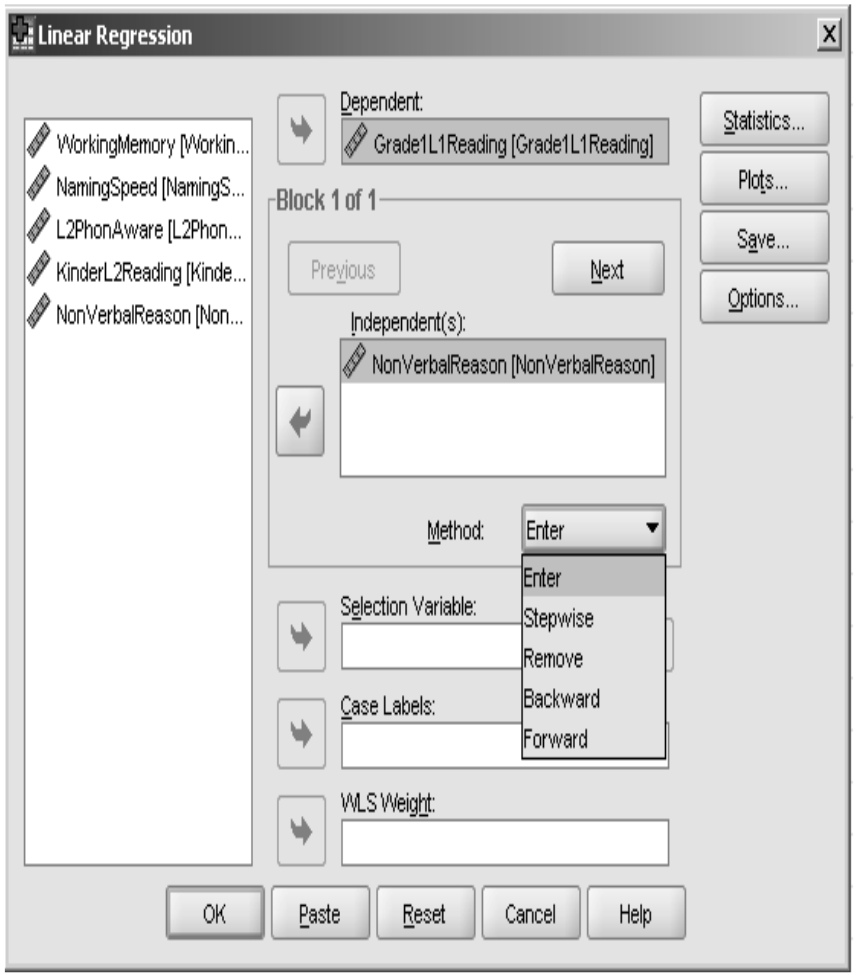
단순회귀분석의 결과제시

- $Y = aX + e$
- 위의 식을 Model이라고 부름
  - [결과제시1] Model이 통계적으로 중요한가?
    - 즉, X가 Y를 설명해 낼 수 있는가?
  - [결과제시2] a의 값 (a가 통계적으로 유의미한가?)
  - [결과제시3] a의 방향성 [positive/negative]
  - [결과제시4] model은 몇 퍼센트의 변이(variable)를 설명해 낼 수 있는가?

SPSS 실행: 단순회귀분석

- [연구문제]
  - 말하기 = 불안감 + error

- 상단메뉴에서
- 분석 → 회귀분석 → 선형



SPSS 결과물 및 해석: 단순회귀분석

진입/제거된 변수[b]

모형	진입된 변수	제거된 변수	방법
1	불안감[a]	.	입력

a. 요청된 모든 변수가 입력되었습니다.

b. 종속변수: 말하기

분산분석[b]

모형		제곱합	자유도	평균 제곱	F	유의확률
1	회귀	307.949	1	307.949	71.242	.000[a]
	모형					
	잔차	423.611	98	4.323		
	합계	731.560	99			

a. 예측값: (상수), 불안감

b. 종속변수: 말하기

- 말하기점수를 종속변수로, 불안감점수를 독립변수로 한 모델을 이용하여, 회귀분석을 실행하였다.
- (회귀분석에 이용된 모델은 통계적으로 유의미하였다 $F(1, 98) = 71.24$, $p = .0001$.)

SPSS 결과물 및 해석: 단순회귀분석

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	8.306	.404		20.548	.000
	불안감	-.643	.076	-.649	-8.441	.000

a. 종속변수: 말하기

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차
1	.649[a]	.421	.415	2.07908

a. 예측값: (상수), 불안감

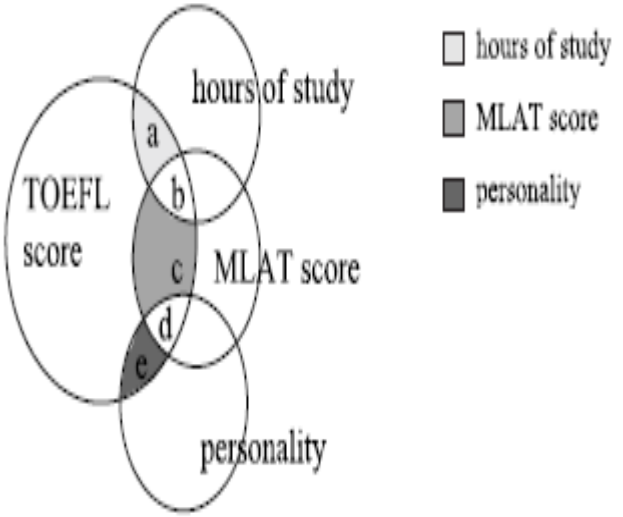
- 독립변수인 불안감 점수는 종속변수인 말하기 점수를 통계적으로 유의미하게 설명해 낼 수 있었으며 [t = -8.44, p = .0001,], 역비례관계를 보였다 B = -.64, beta = -.65.
- 전체변이 중 약 42%를 설명해 낼 수 있었다 R^2 = .42.

다중회귀분석 (multiple regression)

- $Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n + e$

  X, 즉 독립변수 (IV)가 여러 개인 경우

  예) 말하기 = 불안감 + 쓰기 + 자신감 + error



- 주요 연구문제
  - 여러 가지 IV중 DV를 설명/예측해 낼 수 있는 IV는 무엇인가?
  - 만일 여러가지 IV가 통계적으로 유의미하다면, 그 중 상대적으로 더 중요한 것은 무엇인가?

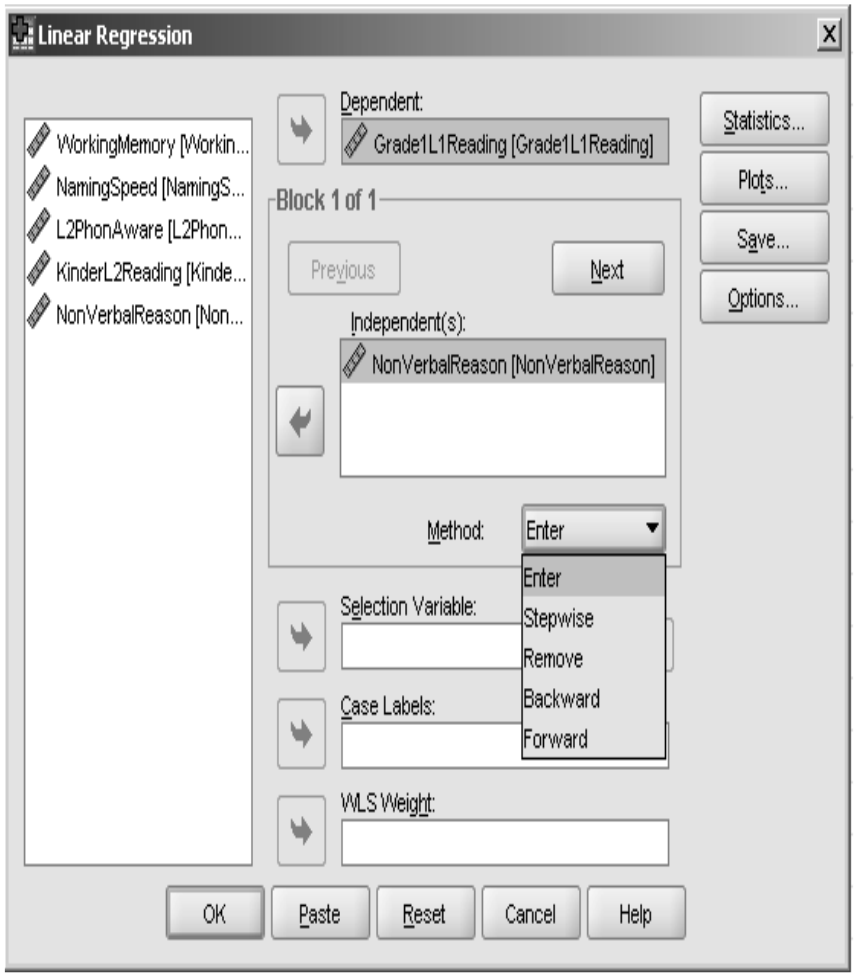
(표준)다중회귀분석의 결과제시

- $Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n + e$
  - [결과제시1] Model이 통계적으로 중요한가?
  - [결과제시2] 여러 IV중 통계적으로 중요한 변수는 무엇인가? 즉, 여러 개의 a 값 중 통계적으로 유의미한 것은 무엇인가?)
  - [결과제시 3] 유의미한 변수의 상대적 기여도
  - [결과제시3] a값들의 방향성 [positive/negative]
  - [결과제시4] model은 몇 퍼센트의 변이(variable)를 설명해 낼 수 있는가?

SPSS 실행: 다중회귀분석

- [연구문제]
  - 말하기 = 쓰기 + 자신감 + 불안감 + error

- 상단메뉴에서
- 분석 → 회귀분석 → 선형



SPSS 결과물 및 해석: 다중회귀분석

진입/제거된 변수[b]

모형	진입된 변수	제거된 변수	방법
1	불안감, 쓰기, 자신감[a]	.	입력

a. 요청된 모든 변수가 입력되었습니다.

b. 종속변수: 말하기

- 입력된 변수들 확인

분산분석[b]

모형		제곱합	자유도	평균 제곱	F	유의확률
1	회귀 모형	651.734	3	217.245	261.260	.000[a]
	잔차	79.826	96	.832		
	합계	731.560	99			

a. 예측값: (상수), 불안감, 쓰기, 자신감

b. 종속변수: 말하기

- 모델이 통계적으로 유의미한가 확인

SPSS 결과물 및 해석: 단순회귀분석

계수[a]

모형		비표준화 계수		표준화 계수		유의확률
		B	표준오차	베타	t	
1	(상수)	1.893	.389		4.863	.000
	쓰기	-.027	.035	-.026	-.777	.439
	자신감	.827	.041	.852	20.311	.000
	불안감	-.142	.042	-.143	-3.402	.001

a. 종속변수: 말하기

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차
1	.944[a]	.891	.887	.91188

a. 예측값: (상수), 불안감, 쓰기, 자신감

- 각 계수에 대한 정보 확인
  - 유의미한 계수
  - 계수의 방향성
  - 계수의 기여도
- 모델의 설명력

계층적 회귀분석
Hierarchical regression

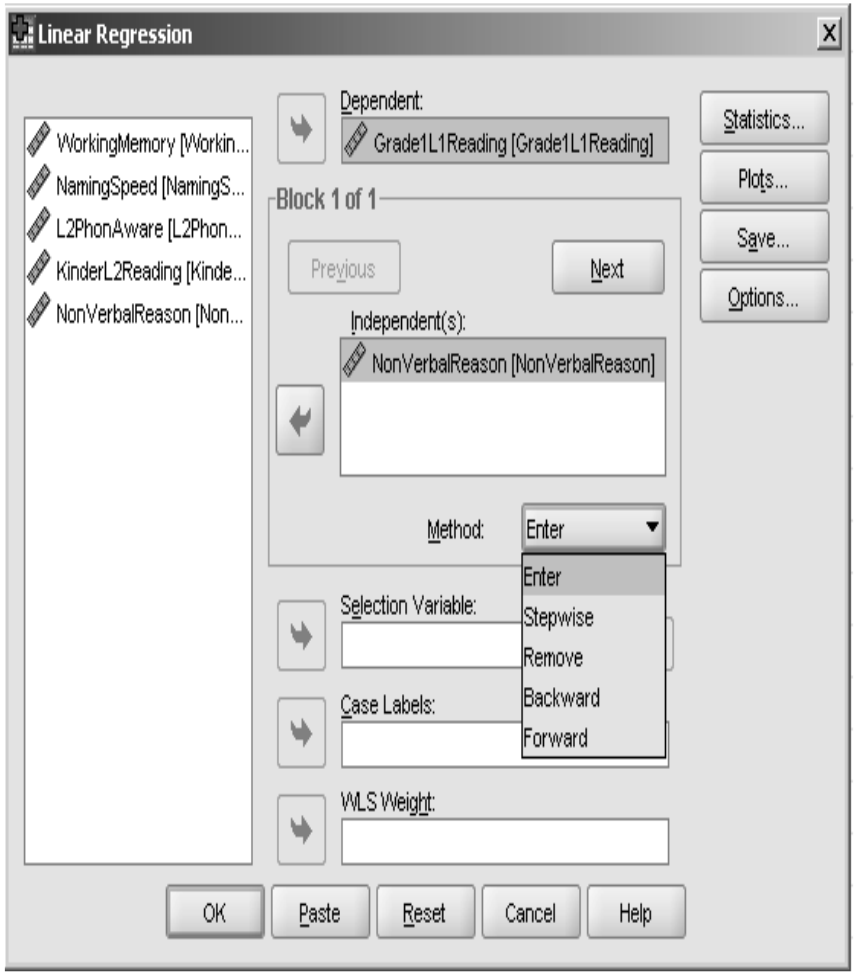
- 여러 가지 IV중 DV를 설명/예측해 낼 수 있는 최적의 모델은 무엇인가?

- Standard/Simulatneous multiple regression
  - 말하기 = 쓰기 + 불안감 + 자신감 + error
- Hierarchical/Stepwise multiple regression
  - [model 1] 말하기 = 쓰기 + 불안감 + error
  - [model 2] 말하기 = 쓰기 + 불안감 + 자신감 + error

SPSS 실행: 다중회귀분석 (hierarchical)

- [연구문제]
  - 말하기 = 자신감 + error
  - 말하기 = 불안감 + 자신감 + error

- 상단메뉴에서
- 분석 → 회귀분석 → 선형
  - 우측메뉴 "통계량"선택 → R 제곱 변화량 선택



SPSS 결과물 및 해석: 계층적 다중 회귀분석

진입/제거된 변수[b]

모형	진입된 변수	제거된 변수	방법
1	자신감[a]	.	입력
2	불안감[a]	.	입력

a. 요청된 모든 변수가 입력되었습니다.

b. 종속변수: 말하기

분산분석[c]

모형		제곱합	자유도	평균 제곱	F	유의확률
1	회귀 모형	640.979	1	640.979	693.482	.000[a]
	잔차	90.581	98	.924		
	합계	731.560	99			
2	회귀 모형	651.232	2	325.616	393.195	.000[b]
	잔차	80.328	97	.828		
	합계	731.560	99			

a. 예측값: (상수), 자신감

b. 예측값: (상수), 자신감, 불안감

c. 종속변수: 말하기

SPSS 결과물 및 해석: 계층적 다중 회귀분석

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차	통계량 변화량				
					R 제곱 변화량	F 변화량	df1	df2	유의확률 F 변화량
1	.936[a]	.876	.875	.96140	.876	693.482	1	98	.000
2	.944[b]	.890	.888	.91002	.014	12.380	1	97	.001

a. 예측값: (상수), 자신감

b. 예측값: (상수), 자신감, 불안감

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	.695	.202		3.435	.001
	자신감	.908	.034	.936	26.334	.000
2	(상수)	1.791	.366		4.898	.000
	자신감	.824	.040	.849	20.360	.000
	불안감	-.145	.041	-.147	-3.519	.001

a. 종속변수: 말하기

Differences

- Univariate vs Multivariate
- How many Ys?

- ANOVA vs MANOVA
- ANCOVA vs MANCOVA

Differences

Between-subject differences

- Independent t-test:

> Y = aX + e

- One-way ANOVA:

> Y = aX + e

- Multi-way ANOVA:

> Y = a1X1 + … + anXn +e

Within-subject differences

- Paired t-test

> Y1 = Y2

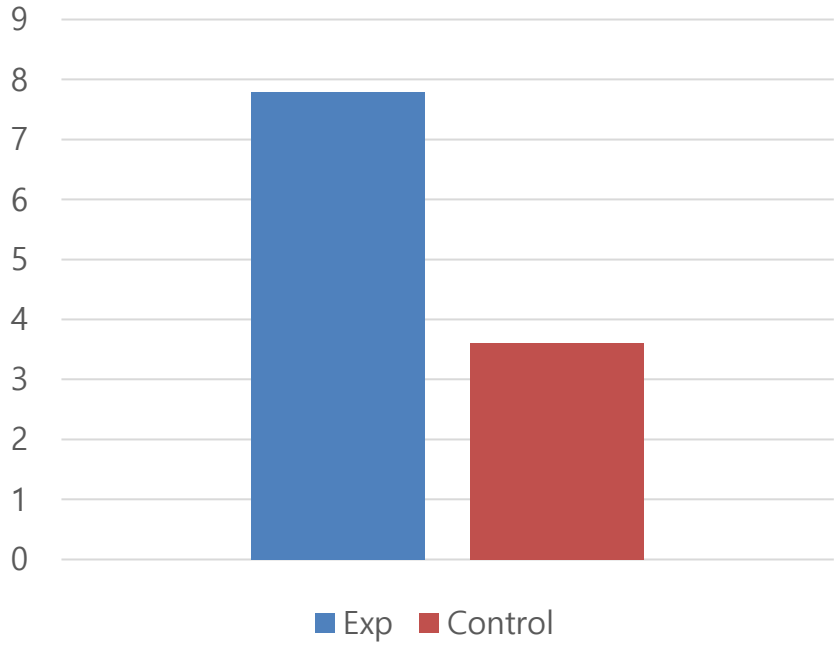
- Within-subject repeated measure ANOVA

> Y1 = Y2 … = Yn

Between-subject differences (2 groups)

Independent t-test

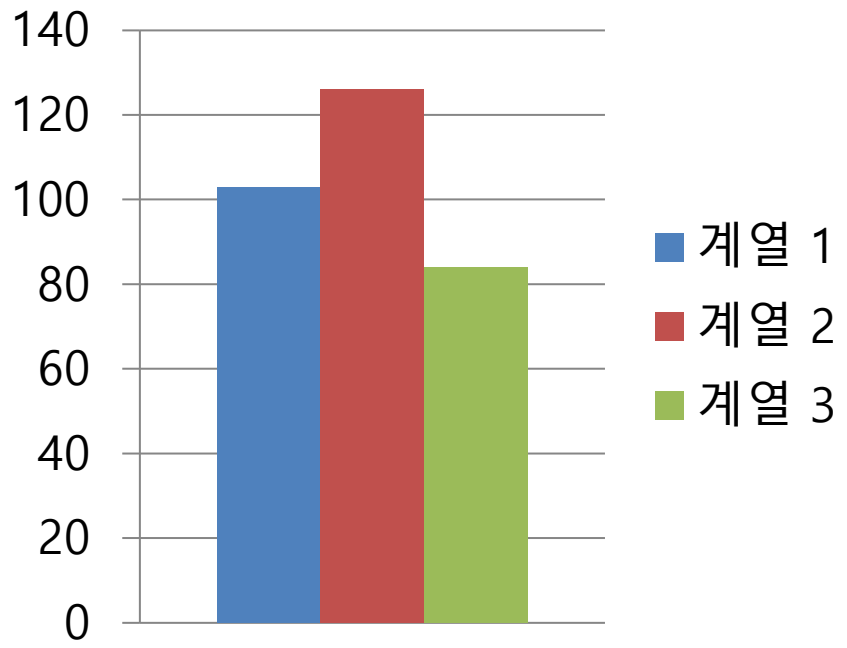
ID	Group	Scores
1	Experimental	9
2	Experimental	8
3	Control	4
4	Control	3
5	Experimental	7
…		
100	Control	5



Between-subject differences (>3 groups)

One-way ANOVA

ID	Group	Scores
1	1	98
2	1	101
3	2	124
4	2	134
5	3	125
…	…	…
100	3	67



Within-subject differences (2 measures)

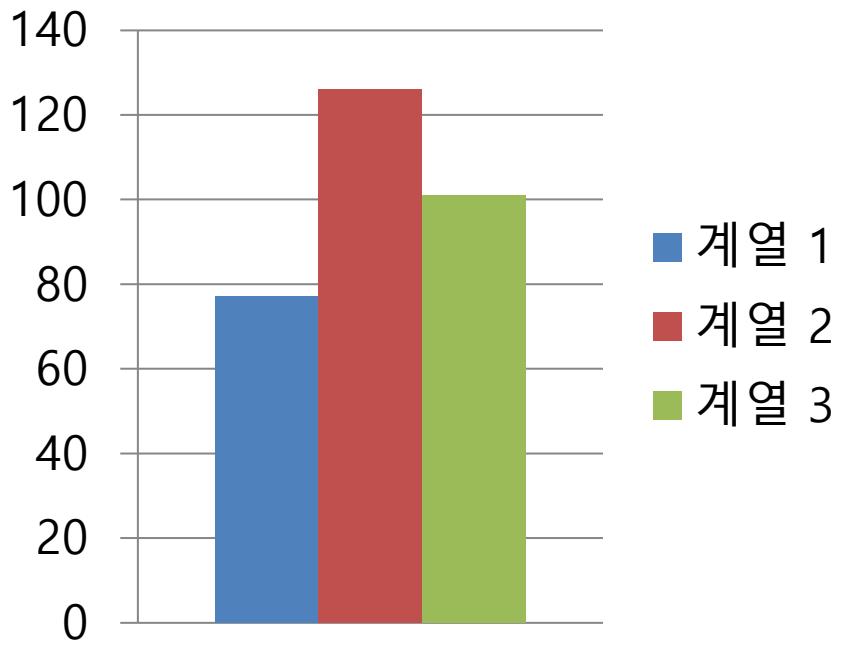
paired t-test

ID	Pre-test	Post-test
1	2	9
2	3	8
3	1	4
4	4	6
5	2	7
…		
100	2	5

Within-subject differences (>3 measures)

Within-subject RMA

ID	Pre-test	Post-test1	Post-test2
1	71	101	98
2	73	112	101
3	79	152	124
4	81	162	134
5	66	145	125
…		…	…
100	62	99	67



Multi-way ANOVA

Y = Group + Gender
+ Group x Gender + (Age + error)

- **Are there differences between groups** with different language backgrounds (three groups) and different gender (two genders) in how accurately they recognize the affect in someone's voice (when the participants' age is statistically controlled for)?

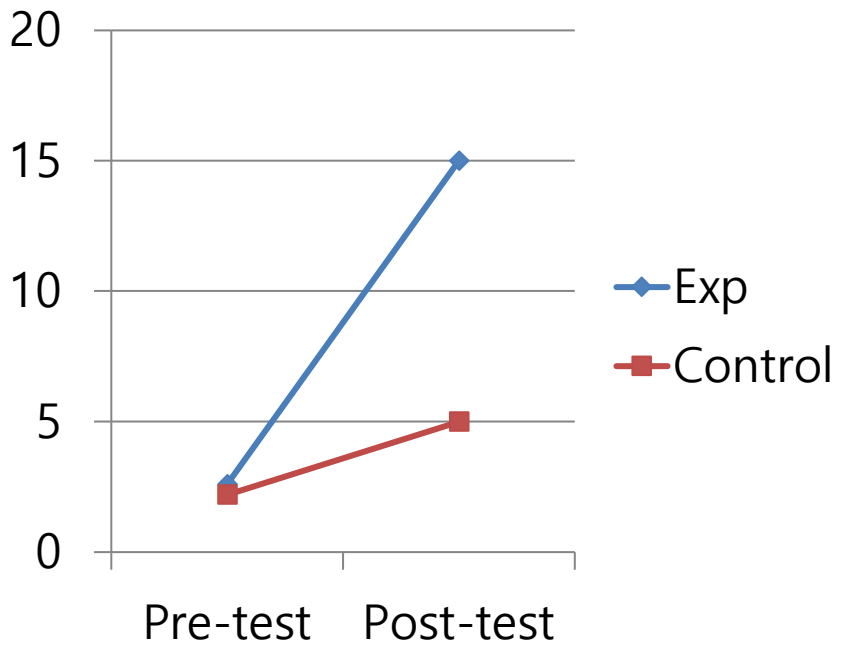
- Dromey, Silveira, and Sandor (2005)

3 x 2 ANCOVA

ID	Group	Gender	Scores	Age
1	1	1	3	**24**
2	2	1	5	**29**
3	3	1	7	**22**
4	1	2	8	**32**
5	2	2	9	**45**
6	3	2	3	**26**
7	1	1	7	**38**
…	…	…	…	**…**
100	3	2	5	**33**

Y = Group + Time + Group x Time

ID	Group	Pre-test	Post-test
1	Exp	1	12
2	Exp	2	15
3	Exp	5	18
4	Control	6	19
5	Control	4	9
…	…	…	…
100	Control	3	6



Mixed ANOVA (Btw-subject factor + Within-subject Factor)

Y = Instruction + Time
+ Instruction x Time + error

Two-way RMA; 2 x 2 RMA; mixed ANOVA

- Will instruction that helps learners to "notice" formulaic sequences **improve** their perceived fluency in L2 English **more than** a group taught by the same teacher with the same materials but without a focus on formulaic sequences?

- Y = perceived fluency
- Instruction = 2 levels = Between-subject
- Time = 2 levels = Within-subject

- Boers, Eyckmans, Kappel, Stengers, and Demecheleer (2006):

- Group
- Time
- Group X Time

Descriptive vs Inferential stats

- **Component 1: Description of your data**
  - Continuous data > Mean, Standard deviation, range
  - Discrete data > Frequency
  - **Meaningful description?**

- **Component 2: Statistically important?**
  - Related?
  - Different?

- **Component 3: The meaning/interpretation of your findings**
  - Is your hypothesis confirmed?
  - So what? (How are your findings related to previous literature?)

- **All the 3 components should be included.**