

빅데이터 예측분석 6

서강대학교 정보통신 대학원 정 화민 교수 (MIS Ph.D.)

Independent t-Analysis

- t- 검증은 모집단을 대표하는 표본으로 부터 추정된 분산이나 표준편차를 가지고 모집단의 평균가 차이가 통계적으로 있는지 검증하나는 방법(기본샘플 30개 이상)
- t- 검증을 하기 위해서는 독립변수는 성별과 같은 질적 변수이고 종속변수는 양적변수 이어야 함 (종속변수는 평균을 낼 수 있어야 함)
- -> 만약 종속변수가 질적 변수이면 이미 배우 카이제곱검증을 해야 함
- 예) 연구가설(대립가설): 성별에 따른 강의 만족도의 차이는 있을 것이다. <-> 영가설: 성별에 따른 강의 만족도의 차이는 없을 것이다.
- t- 검증을 할 때 2 집단의 분산의 동질성 검증을 실시한 후 t -검증을 한다.

$$t_s = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$
 t= 집단 1의 평균 – 집단2의 평균/두 집단간 평균 차의 표준오차

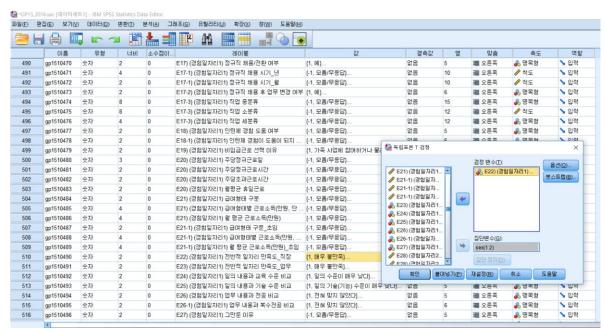
Independent t-Analysis 사례

https://survey.keis.or.kr/index.jsp 고용조사 분석시스템에서 회원가입 후대졸자 직업이동경로조사 자료 받기 (18,082명의 data)

- 남녀 성별에 따른 직장만족도 차이검증 실시 (Independent t-Analysis)



Independent t-Analysis(대졸자 직업이동경로조사 data 활용)



집단통계량

	성별	N	평균	표준화 편차	표준오차 평균
E22) (경험일자리1) 전반	사님	218	3.13	1.095	.074
적 일자리 만족도_직장	여자	287	3.01	1.033	.061

등분산 확인

3. 결론: 성별에 따른 직장

만족도 차이검증 결과 t값은

1.239, 유의확률 값이 .216 으로 그 값이 0.05 보다 큼

으로 통계적으로 성별에 따 른 만족도 차이는 없음.

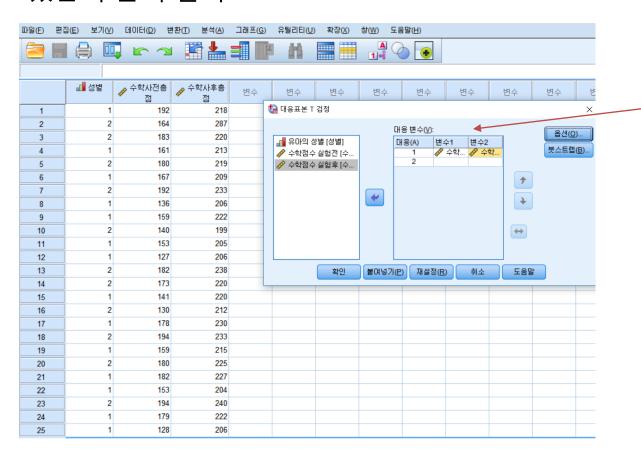
	독립표본 검정											
	Levene의 등	등분산 검정			평균의 등	동일성에 대한	T검정					
						유의확률 (양			차이의 95%	6 신뢰구간		
		F	유의확률	t	자유도	♣)	평균차이	표준오차 차이	하한	상한		
E22) (경험일자리1) 전반	동분산을 가정함	3.750	.053	1.239	503	.216	.118	.095	069	.305		
적 일자리 만족도_직장	동분산을 가정하지 않음			1.229	452.410	.220	.118	.096	071	.307		

분석결과 :

- 1. 남자의 전반적 일자리 만족도는 5점 만점에 평 균 3.13, 여자는 3.01로 나타남.
- 2. 등분산 검정에서는 유의 확률 값이 0.053으로 등분 산이 가정되었음.

Paired t-Analysis

Paired t-검증은 표본을 추출하는 과정이 상호 독립적인 경우가 아니라 비독립적인 경우 집단간 종속변수에 차이가 있는지를 알아보고자 할 때 상용한다. 카페의 수학능력점수 샘플 데이터를 이용하여 수학능력검사의 교육훈련 실험 전과 실험후의 점수차이가 통계적으로 있는지 분석 한다.



변수 1이 사전 변수 2가 사후 로 들어감

Paired t-Analysis

컴퓨터를 활용한 수학활동이 유아의 수학능력에 어떠한 효과가 있는지를 실험한 데이터임.

분석결과, 실험전/후의 인원은 60명 실험전 수학평균은 175.33, 시험후의점수는 203.23으로 평균점수가 올라갔으며, 유의확률값이 .000으로 통계적유의수준에서 컴퓨터를 활용한 수학활동의 평균의 차이가 있는 것으로나타났다. 그러므로 컴퓨터를 활용한 수학활동은 유아의 수학능력에 효과적인 것으로 나타났다.

대응표본 통계량

		평균	N	표준화 편차	표준오차 평균
대응 1	수학점수 실험전	175.33	60	23.091	2.981
	수학점수 실험후	203.23	60	26.376	3.405

대응표본 상관계수

	N	상관관계	유의확률
대응 1 수학점수 실험전 & 수학점 수 실험후	60	.094	.476

대응표본 검정

				대응차					
					차이의 959	6 신뢰구간			유의확률(양 /
		판	표준화 편차	표준오차 평균	하한	상한	t	자유도	\$
대응1	수학점수 실험전 - 수학점 수 실험후	-27.900	33.386	4.310	-36.525	-19.275	-6.473	59	.000

유의확률값 .000 으로 통계적으로 유의한 수준에서 효과가 있음. (p<0.05)

One way ANOVA

One way ANOVA 집단을 구분해 주는 독립변수에 따른 종속변수의 평균의 차이를 분석하는 방법이다.

Independent t-analysis에서의 등분산 가정을 확인한 것 처럼 One way ANOVA분석에서도 등분산 분석을 실시하며 그룹간 등분산을 가정하면 scheffe, 등분산이 가정되지 않았을 때는 Dunnett T3를 기준으로 사후분석을 실시한다.

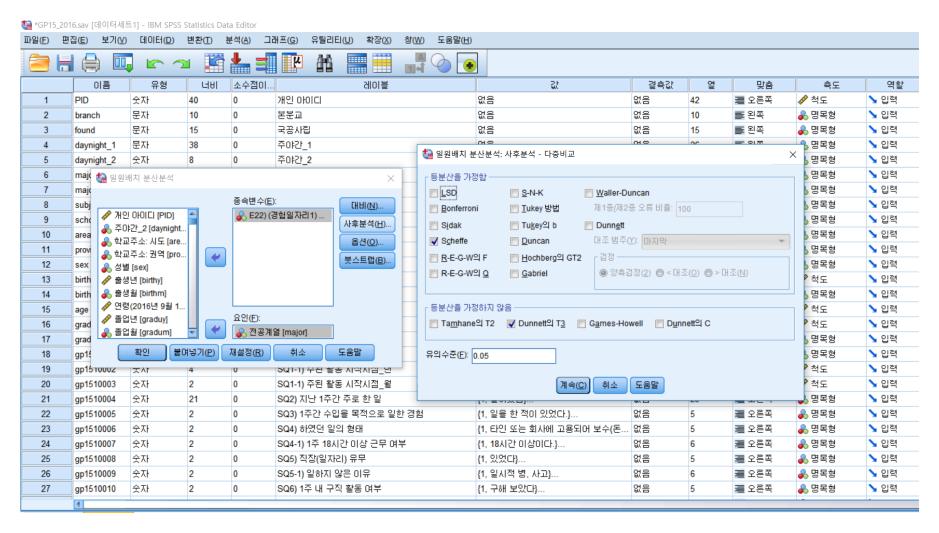
(사후분석은 독립변수가 3개 이상일 때 그룹 간의 차이를 분석하기 위하여 실시하는 방법이다).

-> 시중에 시판하는 일부 교재에서는 독립변수가 2개 이상일때 One way ANOVA분석을 한다고 해 놓은 것들이 많은데 독립변수가 질적 변수이고 2개이면 Independent t-analysis를 하면 된다.

예제 파일 : 다운 받아 놓은 GOMS(대졸자 직업이동경로조사)파일을 다시 열어보자 .

One way ANOVA

전공계열에 따른 직장만족도 차이검증 사후분석으로 scheffe(등분산), Dunnett T3(비 등분산)으로 검증



One way ANOVA 결과

기술통계

E22) (경험일자리1) 전반적 일자리 만족도_직장

					평균에 대한 9	5% 신뢰구간		
	Ν	평균	표준화 편차	표준화 오류	하한	상한	최소값	최대값
인문	65	3.15	1.202	.149	2.86	3.45	1	5
사회	112	2.96	1.048	.099	2.77	3.16	1	5
교육	34	3.74	.864	.148	3.43	4.04	2	5
공학	100	2.94	1.153	.115	2.71	3.17	1	5
자연	60	3.05	1.064	.137	2.78	3.32	1	5
의약	28	3.14	.891	.168	2.80	3.49	1	5
예체능	106	2.99	.921	.089	2.81	3.17	1	5
전체	505	3.06	1.061	.047	2.97	3.15	1	5

분산의 등질성 검정

		Levene 통계 랑	자유도1	자유도2	유의확률
E22) (경험일자리1) 전반	평균을 기준으로 합니다.	2.292	6	498	4 .034
┃ 적 일자리 만족도_직장 ┃	중위수를 기준으로 합니다.	1.930	6	498	.074
	자유도를 수정한 상태에서 중위수를 기준으로 합니다.	1.930	6	483.185	.074
	절삭평균을 기준으로 합니 다.	2.384	6	498	.028

ANOVA

E22) (경험일자리1) 전반적 일자리 만족도_직장

	제곱합	자유도	평균제곱	F	유의확률
집단-간	19.252	6	3.209	2.917	.008
집단-내	547.845	498	1.100		
전체	567.097	504			

등분산이 가정되 지 않았음으로 Dunnett T3로 사 후분석 확인

분석결과:

- 1. 전공계열별 직장만족도에 대한 기술통계표가 작성 되었다. (교육계열의 평균 이 가장 높은 3.74로 나타 났다.)
- 2. 등분산 검정에서는 유의확 률 값이 0.034로 등분산이 가 정되지 않았다.
- 3. ANOVA 의 분석결과: F값은 2.917, 유의확률 값이 .008이 므로 그 값이 0.05 보다 작음으로 통계적으로 전공계열에 따른 만족도 차이는 있음.

One way ANOVA 사후분석 결과

	(A) TI TI TI					95% 신뢰구간		
	(I) 전공계열		평균차이(I-J)	표준화 오류	유의확률	하한	상한	
		사회	0.190	0.179	0.999	-0.36	0.74	
		교육	-0.581	0.210	0.132	-1.24	0.07	
	0.0	공학	0.214	0.188	0.997	-0.37	0.80	
	인문	자연	0.104	0.203	1.000	-0.52	0.73	
		의약	0.011	0.225	1.000	-0.70	0.72	
		예체능	0.163	0.174	1.000	-0.38	0.70	
		인문	-0.190	0.179	0.999	-0.74	0.36	
		교육	−.771 *	0.178	0.001	-1.33	-0.21	
	사회	공학	0.024	0.152	1.000	-0.44	0.49	
	사회	자연	-0.086	0.169	1.000	-0.61	0.44	
		의약	-0.179	0.195	1.000	-0.80	0.44	
		예체능	-0.026	0.133	1.000	0.44	0.38	
		인문	0.581	0.210	0.132	-0.07	1.24	
		사회	.771*	0.178	0.001	0.21	1.33	
	교육	공학	.795*	0.188	0.001	0.21	1.38	
	112 4	자연	.685*	0.202	0.022	0.05	1.32	
		의약	0.592	0.224	0.191	-0.12	1.30	
		예체능	.745*	0.173	0.001	0.20	1,29	
	공학	인문	-0.214	0.188	0.997	-0.80	0.37	
		사회	-0.024	0.152	1.000	-0.49	0.44	
Dunnett T3		교육	−.795 [*]	0.188	0.001	-1.38	-0.21	
Darmett 13		자연	-0.110	0.179	1.000	-0.66	0.44	
1		의약	-0.203	0.204	0.999	-0.85	0.44	
		예체능	-0.051	0.146	1.000	-0.50	0.40	
		인문	-0.104	0.203	1.000	-0.73	0.52	
		사회	0.086	0.169	1.000	-0.44	0.61	
	자연	교육	−.685 [*]	0.202	0.022	-1.32	-0.05	
	A.C.	공학	0.110	0.179	1.000	-0.44	0.66	
		의약	-0.093	0.217	1.000	-0.78	0.59	
		예체능	0.059	0.164	1.000	-0.45	0.57	
		인문	-0.011	0.225	1.000	-0.72	0.70	
		사회	0.179	0.195	1.000	-0.44	0.80	
	의약	교육	-0.592	0.224	0.191	-1.30	0.12	
	_, ,	공학	0.203	0.204	0.999	-0.44	0.85	
		자연	0.093	0.217	1.000	-0.59	0.78	
		예체능	0.152	0.191	1.000	-0.46	0.76	
		인문	-0.163	0.174	1.000	-0.70	0.38	
		사회	0.026	0.133	1.000	-0.38	0.44	
	예체능	교육	−.745 [*]	0.173	0.001	-1.29	-0.20	
	VII.7II 0	공학	0.051	0.146	1.000	-0.40	0.50	
		자연	-0.059	0.164	1.000	-0.57	0.45	
	의약		-0.152	0.191	1.000	-0.76	0.46	
		*. 평균	^균 차이는 0.05 수	준에서 유의합	니다.			

사후분석결과:

- 1. 가장 높은 만족도를 나타 낸 교육계열을 기준으로 사회, 공학, 자연, 예체능 계열은 직장만족도가 통 계적으로 교육계열에 비 하여 낮은 것으로 나타났 다.
- 2. 등분산 가정이면 scheffe 사후분석을 기준으로 그룹별 차이를 확인해야 한다.

등분산이 가정되 지 않았음으로 Dunnett T3와 유 의확률값으로 사 후분석 확인함

Correlation Analysis

Correlation (상관)은 어떤 것과 다른 것의 관계를 말한 것으로 두 변수의 관계를 말합니다.

Correlation Analysis는 두 변수 또는 그 이상의 변수 간에 상관관계가 존재하는지를 파악하고 상관관계 정도를 알아보는 방법으로 한 변수가 커지거나 작아질 때 다른 변수가 어떻게 변화하는지를 알아보면서 그 변동의 정도와 방향을 예측하여 알아보는 방법입니다. 단 상관분석에서의 기본가정은 두 변수나 그 이상의 변수는 모두 등간 및 비율척도로 측정된 것으로 한다. (Pearson 상관분석).

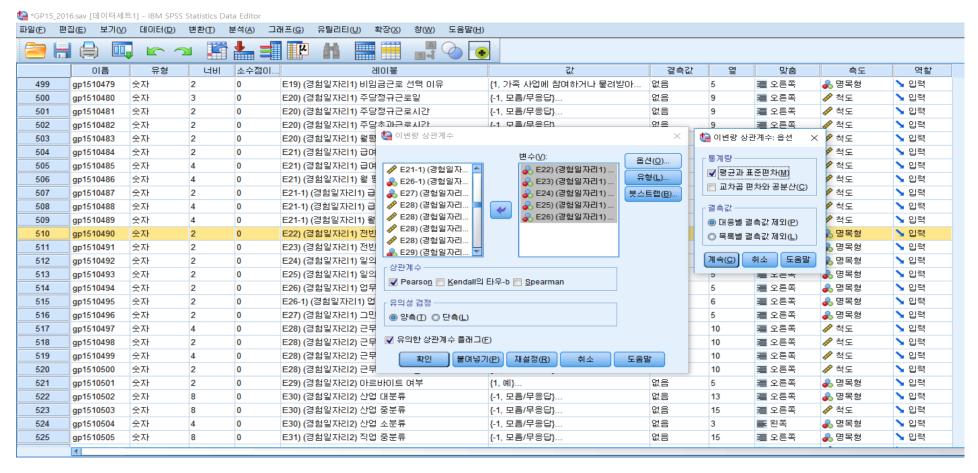
두 변수간 상관분석 공식
$$= \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \times \sum (y-\bar{y})^2}}$$

(각 변수 별 데이터와 평균을 기준으로 계산)

예제 파일 : 다운 받아 놓은 GOMS(대졸자 직업이동경로조사)파일을 다시 열어보자 .

Correlation Analysis

샘플 예제 E22~E26번 까지 상관분석을 해보자. E22(직장만족도), E23(업무만족도),E24(일의 수준), E25(일의 기술수준), E26(업무내용과 전공일치)



Correlation Analysis 결과

직장 만족도와 가장 상관관계가 높은 변수는 업무 만족도로 상관계수가 .738로 두 변수 간에는 73.8%의 통계적 유의수준(p<0.05)에서 긍정적 상관관계가 있는 것으로 나타났다. (-면 부정의 상관관계임)

삼관관계 <table-cell-columns>

			E22) (경험일 자리1) 전반적 일자리 만족도 _직장	E23) (경험일 자리1) 전반적 일자리 만족도 _업무	E24) (경험일 자리1) 일의 내 용과 교육 수 준 비교	E25) (경험일 자라1) 일의 내 용과 기술 주 준 비교	E26) (경험일 자리1) 업무 내 용과 전공 비 교
	E22) (경험일자리1) 전반	Pearson 상관	1	.738**	.257**	.236**	.216**
	적 일자리 만족도_직장	유의확률 (양측)		.000	.000	.000	.000
		N	505	505	505	505	505
•	E23) (경험일자리1) 전반	Pearson 상관	.738**	1	.316**	.266**	.290**
	적 일자리 만족도_업무	유의확률 (양측)	.000		.000	.000	.000
		N	505	505	505	505	505
	E24) (경험일자리1) 일의	Pearson 상관	.257**	.316**	1	.855**	.359**
	내용과 교육 수준 비교	유의확률 (양측)	.000	.000		.000	.000
		N	505	505	505	505	505
	E25) (경험일자리1) 일의	Pearson 상관	.236**	.266**	.855**	1	.313**
	내용과 기술 수준 비교	유의확률 (양측)	.000	.000	.000		.000
		N	505	505	505	505	505
	E26) (경험일자리1) 업무	Pearson 상관	.216**	.290**	.359**	.313**	1
	내용과 전공 비교	유의확률 (양측)	.000	.000	.000	.000	
		N	505	505	505	505	505

^{**.} 상관관계가 0.01 수준에서 유의합니다(양측).

± 0.9 이상 : 매우 높은 상관관계 ± 0.7~0.9 미만: 높 은 상관관계 ± 0.4~0.7 미만: 다 소 높은 상관관계 ± 0.2~0.4 미만: 비교적 낮은 상관 관계 ± 0.2 미만:상관관 계가 거의 없다.