2020 봄학기 '정보와 정보학' 프로젝트 보고서

#2017320233 김채령

강의명: 정보와 정보학

교수님: 김자미 교수님

제출일자: 2020년 6월 2일 정오

[목차]

1. 여행 관련 설문 데이터 분석
   A. 문항 선택 ... 2
   B. 전처리 ... 3
      i. 클리닝
      ii. 데이터 변환
   C. 연구 방법 및 결과 ... 5
      i. 빈도분석, 기술통계분석
      ii. t-Test, F-test, 교차 분석

2. 중간고사 성적 데이터 분석
   A. 데이터 선택과 분석 목표 ... 13
   B. 전처리 ... 13
      i. 클리닝
      ii. 가정 도입 및 데이터 변환
   C. 연구 방법 및 결과 ... 15
      i. 백분위수 값
      ii. 중심경향과 분산

1. 여행 관련 설문 데이터 분석

A. 문항 선택 (총 5 문항)

- 32번 문항 & 33번 문항

32번 문항은 여행지에서 쇼핑할 때의 품목 별 관심도를, 33번 문항은 여행지에서 쇼핑 품목 별 중요도(의미가 얼마나 있는지)를 묻고 있다. 좀 더 해석하자면, 32번 문항은 보통 여행을 가서 관심이 있어서 구경을 하게 되는 쇼핑 품목을 묻는 것이다. 그리고 33번 문항은 여행지에서 꼭 사야하는 쇼핑 품목을 묻는 것으로 이해하였다. 두 문항을 선택한 이유는 여행을 갔던 기억을 떠올려 봤을 때, 지나치지 못하고 다 구경을 해야 하는 관심 품목과 실제 구매까지의 결정을 내리게 하는 품목의 차이가 있었던 것으로 기억하기 때문에 이를 통계적으로 검증해보고 싶었기 때문이다. 따라서 해당 궁금점에 대한 검증을 t 값을 통해 두 종속표본 t 검증을 시도하고자 한다.

- 19번 문항

19번 문항은 선호하는 여행 인원을 묻는 문항으로 혼자, 2명, 3명, 4명, 5명 이상으로 구성된 5가지의 선택지가 있다. 여행과 관련된 설문에 참여하고 결과 데이터를 받는 과정에서, 해당 설문이 누구에게 가장 유용할지를 고민해보았다. 바로 여행과 관련한 마케팅을 하는 여행사나 여행지를 담당하는 지역 단체 혹은 여행지에 위치한 상점들이 해당 설문에 따른 통계를 유용하게 사용할 수 있을 거라는 생각이 들었다. 그 중 상점들이 판매 전략을 정비해 매출을 올리는 데에 해당 설문이 기여했으면 좋겠다는 바람에 여행 인원에 따른 품목별 관심도를 F 검증을 통해 살펴보고자 19번 문항을 선택하게 되었다. 소규모 인원이 주로 여행을 가는 여행지의 상점은 어떤 품목을 어필할 수 있고, 대규모 인원이 주로 여행을 가는 여행지의 상점은 어떤 품목을 전시하여 긍정적인 경제적 효과를 이끌어 낼 수 있는지 해석하고자 한다.

- 15번 문항 & 1번 문항

15번 문항은 여행의 기록을 타인과 공유하는 것을 즐기는 편인지를 묻는 문항으로 예와 아니요 2가지 중 하나로 답변할 수 있다. 해당 문항을 선택한 이유는 여행에 관한 설문을 하면서 가장 신선한 변인으로 작용할 수 있겠다는 예상을 했기 때문이다. 특히, 요즘과 같이 소셜 미디어가 활발하게 활용되는 시대에서 여행의 기록을 타인과 공유하기를 즐기는 지의 여부와 여행을 좋아하는 지의 여부가 통계적으로 유의미한 관계가 있는지를 교차 분석의 독립성 검정으로 해석해보고자 한다.

B. 전처리

i. 클리닝

- 32번 문항

결측치나 관심도를 표기하는 1, 2, 3, 4, 5를 벗어나는 이상치 및 잡음치가 존재하지 않았다.

- 33번 문항

결측치나 의미도(의미 정도)를 표기하는 1, 2, 3, 4, 5를 벗어나는 이상치 및 잡음치가 존재하지 않았다.

- 19번 문항

결측치나 선호하는 여행 인원을 표기하는 1, 2, 3, 4, 5를 벗어나는 이상치 및 잡음치가 존재하지 않았다.

- 15번 문항

결측치나 선호하는 여행 인원을 표기하는 1, 2, 3, 4, 5를 벗어나는 이상치 및 잡음치가 존재하지 않았다.

- 1번 문항

결측치나 여행을 선호하는 지를 표기하는 1, 2, 3을 벗어나는 이상치 및 잡음치가 존재하지 않았다.

ii. 데이터 변환

- 32번 문항 & 33번 문항

Q32-1

		빈도	퍼센트	유효 퍼센트	누적 퍼센트
유효	1	52	44.8	44.8	44.8
	3	32	27.6	27.6	72.4
	4	27	23.3	23.3	95.7
	5	5	4.3	4.3	100.0
	전체	116	100.0	100.0	

Q32-5

		빈도	퍼센트	유효 퍼센트	누적 퍼센트
유효	1	66	56.9	56.9	56.9
	3	23	19.8	19.8	76.7
	4	19	16.4	16.4	93.1
	5	8	6.9	6.9	100.0
	전체	116	100.0	100.0	

Q33-1

		빈도	퍼센트	유효 퍼센트	누적 퍼센트
유효	1	22	19.0	19.0	19.0
	2	26	22.4	22.4	41.4
	3	30	25.9	25.9	67.2
	4	32	27.6	27.6	94.8
	5	6	5.2	5.2	100.0
	전체	116	100.0	100.0	

Q33-5

		빈도	퍼센트	유효 퍼센트	누적 퍼센트
유효	1	37	31.9	31.9	31.9
	2	31	26.7	26.7	58.6
	3	31	26.7	26.7	85.3
	4	14	12.1	12.1	97.4
	5	3	2.6	2.6	100.0
	전체	116	100.0	100.0	

문항 32번과 33번에 대해서 빈도 분석을 해보았더니, 두 문항 모두 공통적으로 전반적인 품목에 따른 답변에서 5에 해당하는 '매우 관심/의미 있다' 변수가 빈도수가 상이하게 적은 것으로 드러났다. 따라서 이러한 빈도에 따라 분석이 왜곡되는 것을 방지하기 위해서 5에 해당하는 '매우 관심/의미 있다'와 4에 해당하는 '관심/의미 있다'를 병합하여 '관심 있다'로 해석하기로 한다.

```
RECODE Q32_1(5=4).
RECODE Q32_2(5=4).
RECODE Q32_3(5=4).
RECODE Q32_4(5=4).
RECODE Q32_5(5=4).
RECODE Q32_6(5=4).
RECODE Q32_7(5=4).
RECODE Q33_1(5=4).
RECODE Q33_2(5=4).
RECODE Q33_3(5=4).
RECODE Q33_4(5=4).
RECODE Q33_5(5=4).
RECODE Q33_6(5=4).
RECODE Q33_7(5=4).
```

Q32-1

		빈도	퍼센트	유효 퍼센트	누적 퍼센트
유효	1	52	44.8	44.8	44.8
	3	32	27.6	27.6	72.4
	4	32	27.6	27.6	100.0
	전체	116	100.0	100.0	

Q32-5

		빈도	퍼센트	유효 퍼센트	누적 퍼센트
유효	1	66	56.9	56.9	56.9
	3	23	19.8	19.8	76.7
	4	27	23.3	23.3	100.0
	전체	116	100.0	100.0	

위의 데이터 변환을 통해 변수 간의 빈도 수 차이를 완화시켰다.

- 19번 문항

Q19

		빈도	퍼센트	유효 퍼센트	누적 퍼센트
유효	1	12	10.3	10.3	10.3
	2	44	37.9	37.9	48.3
	3	26	22.4	22.4	70.7
	4	31	26.7	26.7	97.4
	5	3	2.6	2.6	100.0
	전체	116	100.0	100.0	

19번 문항을 빈도 분석해본 결과, 5명 이상을 의미하는 5의 빈도수가 상대적으로 매우 낮다. 이 경우, 제대로 된 통계 검정을 위해서 데이터를 합쳐줄 필요성을 확인할 수 있다. 5명 이상을 뜻하는 5를 4로 바꾸어 4의 의미를 4명 이상으로 확장하겠다. 방법은 위의 32번과 33번 문항처럼 RECODE 명령어를 이용한다.

```
RECODE Q19 (5=4).
EXECUTE.
FREQUENCIES VARIABLES=Q19
  /ORDER=ANALYSIS.
```

Q19

		빈도	퍼센트	유효 퍼센트	누적 퍼센트
유효	1	12	10.3	10.3	10.3
	2	44	37.9	37.9	48.3
	3	26	22.4	22.4	70.7
	4	34	29.3	29.3	100.0
	전체	116	100.0	100.0	

변환 후, 5의 빈도 수가 4의 빈도 수에 더해져서 변인간의 빈도 차이가 조금 더 고르게 바뀌었음을 확인할 수 있다. 변인 1 즉, 혼자에 대한 빈도수가 상대적

으로 작다고 판단할 수 있으나 이에 대해서는 데이터 변환을 수행하지 않기로 결정했다. 왜냐하면, 선호하는 여행 인원에 있어서 혼자이냐 동행이 있느냐는 구분할 필요가 있다고 생각했기 때문이다.

C. 연구 방법 및 결과

i. 빈도분석, 기술통계분석

32번 문항의 6번째 변인, 여행지에서의 쇼핑에서 기념품으로 대표되는 제조 제품에 대한 관심도에 대해 간단한 빈도 분석과 기술 통계 분석을 해보고자 한다.

Q32-6

		빈도	퍼센트	유효 퍼센트	누적 퍼센트
유효	1	37	31.9	31.9	31.9
	3	28	24.1	24.1	56.0
	4	51	44.0	44.0	100.0
	전체	116	100.0	100.0	

기술통계량

	N	최소값	최대값	평균	표준편차
Q32-6	116	1	4	2.80	1.300
유효 N(목록별)	116				

앞서 전처리에 따라 답변 1은 전혀 관심 없음, 2는 관심 없음, 3은 보통임, 4는 관심 있음을 나타낸다. 기술 통계 분석에 따라 평균은 2.80으로 이번 학기 정보와 정보학을 수강하는 고려대학교 학생 116명은 평균적으로 여행지에서 제조 제품에 대해 관심이 보통임을 확인할 수 있다. 빈도 분석을 통해서는 최빈값이 4로, 약 44%의 학생들이 여행지에서의 제조 제품 쇼핑에 관심이 있음을 알 수 있다. 즉, 해당 설문자들이 여행을 가서 쇼핑을 할 경우 품목 중 제조 제품에 관심을 가지는 학생들이 가장 많지만 과반수는 아님을 의미한다. 따라서, 추가적으로 해당 설문자들을 대상으로 여행지에서 팔 품목을 선정할 때 제조 제품은 평균적으로 관심도가 보통임으로 관심을 끌기에 아주 적절한 품목은 아님을 해석해 낼 수 있다.

ii. t-Test, F-Test, 교차 분석

1. 문항 32 & 문항 33: **<u>두 종속 표본 t 검증</u>**

문항 32와 문항 33의 집단은 독립적이지 않은데, 그 이유는 쇼핑 품목별 관심도와 의미 정도이기 때문이다. 따라서 두 종속 표본 t 검증을 통해 쇼핑 품목 별 관심도와 의미도의 차이에 대한 분석을 시도하고자 한다.

이를 위한 기본 가정 3가지를 가정한다. 먼저, 두 변수는 등간 척도 이상의 양적 변수이며, 161개의 데이터로부터 중심 극한 정리에 따라 모집단의

분포가 정규분포임을 가정한다. 그리고 두 집단에 해당하는 모집단의 분산이 동일함 역시 가정한다.

<쇼핑 품목 별 관심도와 의미 정도 차이 검증>

- 영가설_1: 생활 용품에서 관심도와 의미 정도는 같을 것이다.
- 대립가설_1: 생활 용품에서 관심도와 의미 정도는 다를 것이다.
- 영가설_2: 식품(특산물 등)에서 관심도와 의미 정도는 같을 것이다.
- 대립가설_2: 식품(특산물 등)에서 관심도와 의미 정도는 다를 것이다.
- 영가설_3: 의류에서 관심도와 의미 정도는 같을 것이다.
- 대립가설_3: 의류에서 관심도와 의미 정도는 다를 것이다.
- 영가설_4: 잡화(가방, 신발 등)에서 관심도와 의미 정도는 같을 것이다.
- 대립가설_4: 잡화(가방, 신발 등)에서 관심도와 의미 정도는 다를 것이다.
- 영가설_5: 전자 제품에서 관심도와 의미 정도는 같을 것이다.
- 대립가설_5: 전자 제품에서 관심도와 의미 정도는 다를 것이다.
- 영가설_6: 제조업 용품(관광민예품, 장난감 등)에서 관심도와 의미 정도는 같을 것이다.
- 대립가설_6: 제조업 용품(관광민예품, 장난감 등)에서 관심도와 의미 정도는 다를 것이다.
- 영가설_7: 기타 품목에서 관심도와 의미 정도는 같을 것이다.
- 대립가설_7: 기타 품목에서 관심도와 의미 정도는 다를 것이다.

대응표본 검정

		대응차					t	자유도	유의확률 (양측)
		평균	표준화 편차	표준오차 평균	차이의 95% 신뢰구간 하한	상한			
대응 1	Q32-1 - Q33-1	-.345	.943	.088	-.518	-.171	-3.939	115	.000
대응 2	Q32-2 - Q33-2	-.207	.786	.073	-.351	-.062	-2.835	115	.005
대응 3	Q32-3 - Q33-3	-.259	1.104	.103	-.462	-.056	-2.522	115	.013
대응 4	Q32-4 - Q33-4	-.112	1.011	.094	-.298	.074	-1.194	115	.235
대응 5	Q32-5 - Q33-5	-.147	1.144	.106	-.357	.064	-1.380	115	.170
대응 6	Q32-6 - Q33-6	-.397	1.062	.099	-.592	-.201	-4.021	115	.000
대응 7	Q32-7 - Q33-7	-.060	.761	.071	-.200	.080	-.854	115	.395

1종 오류 허용 한계인 p값이 0.05 미만인 대응을 통계적으로 유의미한 차이가 있다고 판단한다. 따라서, 유의 확률이 0.05 미만인 대응 1, 대응 2, 대응 3, 대응 6은 통계적인 차이가 있다고 판단하며 대립 가설을 채택한다. 유의 확률이 0.05 이상인 대응 4, 대응 5, 대응 7에 대해서는 통계적으로 차이가 없다고 분석한며 대립 가설을 채택할 수 없다.

T 검정

대응표본 통계량

		평균	N	표준화 편차	표준오차 평균
대응 1	Q32-1	2.38	116	1.303	.121
	Q33-1	2.72	116	1.116	.104
대응 2	Q32-2	3.41	116	1.038	.096
	Q33-2	3.61	116	.810	.075
대응 3	Q32-3	2.67	116	1.343	.125
	Q33-3	2.93	116	1.125	.104
대응 4	Q32-4	2.73	116	1.321	.123
	Q33-4	2.84	116	1.131	.105
대응 5	Q32-5	2.09	116	1.305	.121
	Q33-5	2.24	116	1.060	.098
대응 6	Q32-6	2.80	116	1.300	.121
	Q33-6	3.20	116	1.049	.097
대응 7	Q32-7	2.53	116	1.034	.096
	Q33-7	2.59	116	.895	.083

각 대응에 대한 통계량은 위와 같다. 해당 검증에 대한 해석을 위해 각 변수를 대응하는 문항 내용으로 대체하여 아래의 표로 재구성해보았다.

No.	품목	관심도	의미 정도	t	p
		M(SD)	M(SD)		
1	생활 용품	2.38(1.303)	2.72(1.116)	-3.939	.000
2	식품(특산물 등)	3.41(1.038)	3.61(.810)	-2.835	.005
3	의류	2.67(1.343)	2.93(1.125)	-2.522	.013
4	잡화(가방, 신발 등)	2.73(1.321)	2.84(1.131)	-1.194	.235
5	전자 제품	2.09(1.305)	2.24(1.060)	-1.380	.170
6	제조업 제품(관광민예품, 장난감 등)	2.80(1.300)	3.20(1.049)	-4.021	.000
7	기타	2.53(1.034)	2.59(.895)	-.854	.395

생활 용품에 대한 관심도는 2.38이며, 의미 정도는 2.72로 유의수준 .05에서 통계적으로 의미 있는 차이를 보인다. 따라서 여행지에서의 생활 용품 관련 쇼핑 품목은 관심도에 비해 의미 정도가 큰 것으로 해석할 수 있다. 이는 여행자들이 여행지에서 쇼핑할 때, 그 품목이 생활 용품에 해당하면 해당 제품의 구매가 더 중요하고 의미 있다고 판단한다는 것이다. 이 통계치에 따라 관광 명소의 가게에서는 생활 용품에 해당하는 물품을 많이 전시하고 판매함으로써 매출을 올릴 수 있다는 해석이 가능하다.

특산물 등의 식품에 대한 관심도는 3.41이며, 의미 정도는 3.61로 유의수준 .05에서 통계적으로 의미 있는 차이를 나타낸다. 이 품목 역시 관심도에 비해서 의미 정도가 크다. 따라서 여행자들이 여행지에서 쇼핑을 할 때, 식품 상품에 대해 보이는 관심에 비해 그 의미나 중요도를 더 크게 둔다는 것이다. 따라서 관광 명소의 가게에서 그 지역의 특산물 등의 식품에 해당하는 품목을 더 많이 홍보하면 매출에 긍정적인 변화를 줄 수 있다는 제안이 가능하다. 또한, 품목별 관심도와 중요도 전반에서 식품에 대한 관심도와 중요도가 가장 높게 나타났다. 이로부터 관광지의 상점이 길목에 식품에 해당하는 품목을 더 많이 내보이는 것은 여행자의 관심을 집중시킬 뿐만 아니라, 통계적으로 여행자들이 해당 품목의 쇼핑을 중요하고 의미 있다고 여기기 때문에 구매까지 이어질 수 있음을 기대할 수 있다고 해석할 수 있다

의류 품목에 대한 관심도는 2.67이고 의미 정도는 2.93이다. 그리고 일반적으로 기념품으로 통칭되는 제조업 제품에 대한 관심도는 2.80이며 그 의미 정도는 3.20이다. 의류와 제조업 제품 모두 유의 수준 .05에서 통계적을 차이가 있으며, 따라서 두 품목 모두 관심도보다 의미 정도 즉, 중요도가 높다고 할 수 있다. 따라서 의류와 제조업 제품 품목에 대한 전시나 홍보가 관광지에 위치한 가게에서 여행자를 대상으로 활발히 이루어져 관심을 끌면, 여행자들의 소비로 이어질 수 있음을 제안할 수 있다.

통계적으로 차이가 있다고 해석되는 4가지 품목 즉, 생활 용품, 식품, 의류, 제조업 제품에 대해서는 모두 관심도보다 의미 정도가 높다고 나타나 있다. 따라서, 여행지의 상점의 판매 전략을 포함한 마케팅 계획을 세울 때 위의 4가지 품목의 홍보 및 전시를 눈에 띄게 하는 전략을 채택하여 관광지 활성화뿐만 아니라 해당 관광지의 경제 성장에도 기여할 수 있다고 해석할 수 있다.

유의 수준 .05에서 관심도와 의미 정도에 통계적으로 차이가 있다고 해석할 수 없는 품목으로는 가방과 신발 등으로 대표되는 잡화, 전자 제품, 그리고 기타 품목이 있다. 따라서, 해당하는 3가지 품목에 대해서는 여행지에서 쇼핑 품목으로써 여행자에게 받는 관심도와 중요도가 같다고 해석할 수 있다. 즉, 그 품목들에 대한 관심이 있다고 해서 그 정도보다 더 강하게, 혹은 더 약하게 해당 품목의 쇼핑에 의미나 중요성을 두지 않는다는 것이다. 다르게 해석하면, 그 품목들에 대한 관심이 있는 만큼 쇼핑을 하는 데에 중요도나 의미 정도를 둔다는 것이다. 추가적인 설문을 통해서 고객 타겟에 대한 특성을 파악하여 해당 품목에 관심이 높은 고객 특성에 따라 마케팅 전략을 세우는 것으로 위 3개의 품목에 대한 t 값을 연장하여 활용할 수 있겠다.

2. 문항 19 & 문항 32: **F 검증**

F 검증을 수행하기 위해서 기본적으로 등간 척도 이상의 양적 변수인 종속 변수와 각 집단의 모집단에 대해서 정규성과 등분산성을 가정한다. 중심 극한 정리를 통해 정규성을 가정하고 등분산성도 가정한다. 해당 검증에서 종속 변수에 해당하는 문항 32의 품목 별 관심도는 계량화할 수 있는 양적 변수임으로 역시 가정을 만족한다.

<선호하는 여행 인원에 따라 관심 있는 쇼핑 품목 알아보기>

- 영가설_1: 선호하는 여행 인원에 따라 생활 용품에 대한 관심도는 같을 것이다.
- 대립가설_1: 선호하는 여행 인원에 따라 생활 용품에 대한 관심도는 다를 것이다.
- 영가설_2: 선호하는 여행 인원에 따라 식품에 대한 관심도는 같을 것이다.
- 대립가설_2: 선호하는 여행 인원에 따라 식품에 대한 관심도는 다를 것이다.
- 영가설_3: 선호하는 여행 인원에 따라 의류에 대한 관심도는 같을 것이다.
- 대립가설_3: 선호하는 여행 인원에 따라 의류에 대한 관심도는 다를 것이다.
- 영가설_4: 선호하는 여행 인원에 따라 잡화에 대한 관심도는 같을 것이다.
- 대립가설_4: 선호하는 여행 인원에 따라 잡화에 대한 관심도는 다를 것이다.
- 영가설_5: 선호하는 여행 인원에 따라 전자 제품에 대한 관심도는 같을 것이다.
- 대립가설_5: 선호하는 여행 인원에 따라 전자 제품에 대한 관심도는 다를 것이다.

- 영가설_6: 선호하는 여행 인원에 따라 제조업 제품에 대한 관심도는 같을 것이다.
- 대립가설_6: 선호하는 여행 인원에 따라 제조업 제품에 대한 관심도는 다를 것이다.
- 영가설_7: 선호하는 여행 인원에 따라 기타 품목에 대한 관심도는 같을 것이다.
- 대립가설_7: 선호하는 여행 인원에 따라 기타 품목에 대한 관심도는 다를 것이다.

일원배치 분산분석

ANOVA

		제곱합	자유도	평균제곱	F	유의확률
Q32-1	집단-간	17.620	3	5.873	3.702	.014
	집단-내	177.691	112	1.587		
	전체	195.310	115			
Q32-2	집단-간	1.775	3	.592	.542	.654
	집단-내	122.182	112	1.091		
	전체	123.957	115			
Q32-3	집단-간	13.379	3	4.460	2.572	.058
	집단-내	194.173	112	1.734		
	전체	207.552	115			
Q32-4	집단-간	10.748	3	3.583	2.112	.103
	집단-내	189.968	112	1.696		
	전체	200.716	115			
Q32-5	집단-간	3.641	3	1.214	.707	.550
	집단-내	192.316	112	1.717		
	전체	195.957	115			
Q32-6	집단-간	1.878	3	.626	.364	.779
	집단-내	192.562	112	1.719		
	전체	194.440	115			
Q32-7	집단-간	1.108	3	.369	.339	.797
	집단-내	121.815	112	1.088		

F 검정 결과 유의 수준 .05에서 문항 32의 1번 변인 즉, 생활 용품에 대응하는 대립 가설을 채택할 수 있음을 확인하였다.

사후검정

다중비교

Scheffe

종속변수	(I) Q19	(J) Q19	평균차이(I-J)	표준화 오류	유의확률	95% 신뢰구간 하한	95% 신뢰구간 상한
Q32-1	1	2	.402	.410	.811	-.76	1.57
		3	-.590	.440	.616	-1.84	.66
		4	-.225	.423	.963	-1.43	.97
	2	1	-.402	.410	.811	-1.57	.76
		3	-.991*	.312	.021	-1.88	-.11
		4	-.627	.288	.197	-1.44	.19
	3	1	.590	.440	.616	-.66	1.84
		2	.991*	.312	.021	.11	1.88
		4	.364	.328	.746	-.57	1.30
	4	1	.225	.423	.963	-.97	1.43
		2	.627	.288	.197	-.19	1.44
		3	-.364	.328	.746	-1.30	.57

위의 사후 검정 결과에 따라, 선호하는 여행 인원에 따른 생활 용품 품목에 대한 쇼핑 관심도를 좀 더 보기 쉽게 아래의 표로 재구성했다.

품목	여행 인원	M	F	p	사후 분석
생활 용품	혼자	1.93	3.702	.014	(2, 3)
	2인	2.33			
	3인	2.56			
	4인 이상	2.92			

선호하는 여행 인원에 따른 여행지에서의 생활 용품 품목 쇼핑에 대한 관심도를 분석한 결과, 4인 이상의 경우가 2.92로 가장 높은 관심도를 나타냈고, 3인, 2인, 혼자의 순서로 뒤따른다. 검정 결과, 유의수준 .05에서 통계적으로 유의한 차이를 나타냈다. 이에 몇 명의 여행 인원과 다른 몇 명의 여행 인원이 차이를 나타내는지 유의수준 .05의 수준에서 사후 분석한 결과, 2인인 집단과 3인인 집단 간에 차이가 있었다. 즉, 여행지에서의 쇼핑 중 식품 품목에 대한 관심도가 여행집단이 2명인지 혹은 3명인지에 따라 통계적으로 유의미한 차이가 있다는 것이다. 그리고 3인인 경우 그 관심도가 더 높다. 따라서 여행지에서의 쇼핑에서 생활 용품 품목에 대한 관심도는 여행의 인원이 많을 수록 높으며, 혼자인 경우에는 관심이 없다고 결론 내릴 수 있다. 해당 통계치에 대해서 여행지의 상점들이 취할 수 있는 마케팅 전략을 제시하고자 한다. 여행자들이 주로 혼자 방문하는 예를 들어 순례길과 같은 여행지의 상점의 경우 생활 용품에 대한 홍보나 전시가 여행자들의 관심을 끌지 않을 것이므로 해당 품목을 덜 고려하고, 동행과 함께 여행하는 관광지의 가게의 경우 생활 용품에 대한 고려를 상대적으로 조금 더 하는 것이 여행자들의 관심을 끌 수 있는 방안이라고 제시할 수 있다. 정리하자면, 본 분석결과를 토대로 여행하는 인원에 따라서 어필 혹은 홍보해야 하는 쇼핑 품목이 다를 수 있는 것으로 해석할 수 있다.

3. 문항 15 & 문항 1: **교차분석 – 독립성 검정**

교차 분석은 각 표본의 모집단으로부터의 추출, 질적 혹은 범주 변수의 종속 변수, 독립적인 각 범주의 응답, 그리고 최소 기대 빈도가 5이하인 셀이 전체의 20%이하여야 함을 기본 가정으로 삼는다. 최소 기대 빈도에 대한 가정은 아래 검정 결과에서 확인할 수 있고, 나머지는 모두 성립함을 가정한다.

<여행 기록을 공유하기를 즐기는 지 여부와 여행을 좋아하는 지의 관계>

- 영가설: 여행 기록을 공유하기를 즐기는 지와 여행을 좋아하는 지는 관계가 없다.
- 대립가설: 여행 기록을 공유하기를 즐기는 지와 여행을 좋아하는 지는 관계가 있다.

Q1 * Q15 교차표

			Q15		
			1	2	전체
Q1	1	빈도	50	27	77
		Q15 중 %	80.6%	50.0%	66.4%
	2	빈도	7	11	18
		Q15 중 %	11.3%	20.4%	15.5%
	3	빈도	5	16	21
		Q15 중 %	8.1%	29.6%	18.1%
전체		빈도	62	54	116
		Q15 중 %	100.0%	100.0%	100.0%

카이제곱 검정

	값	자유도	근사 유의확률 (양측검정)
Pearson 카이제곱	13.031[a]	2	.001
우도비	13.380	2	.001
선형 대 선형결합	12.749	1	.000
유효 케이스 수	116		

a. 0 셀 (0.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다. 최소 기대빈도는 8.38입니다.

카이 제곱 검정에서 빈도수가 5보다 작은 셀이 전체의 20% 미만을 차지하고, 유의 수준 .05에서 대립가설을 채택할 수 있음을 볼 수 있다. 즉, 여행 기록을 타인과 공유하기를 즐기는 지와 여행을 좋아하는 지의 여부가 관련성이 있다는 것이다. 공유하기를 즐기는 사람이 여행을 더 좋아하고, 공유하기를 즐기지 않는 사람이 여행을 덜 좋아하는 것으로 나타난다. 이를 통해서, 여행 기록을 공유하기를 즐기는 사람들을 대상으로 하는 여행 상품 출시를 늘리기 위한 방안을 여행사에서 모색할 필요가 있음을 이끌어낼 수 있다.

2. 중간고사 성적 데이터 분석

A. 데이터 선택과 분석 목표

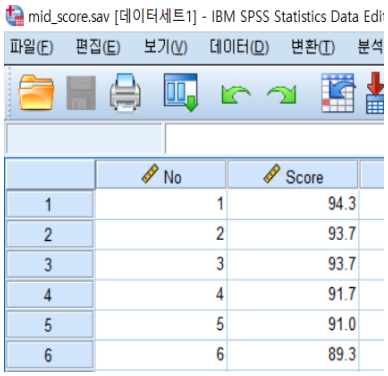
모든 표본 분포는 표본의 크기가 커짐에 따라 정규 분포에 유사한 형태로 변해간다는 즉, 표본 평균이 정규 분포를 따르게 된다는 '중심 극한 정리'를 실습하고자 한다. '중심 극한 정리'를 통해 가정한 정규성이 다양한 통계 분석의 기반 가정이 된다는 사실은 중심 극한 정리의 유용성을 강조한다. 중심 극한 정리를 평균, 분산, 표준편차, 최빈값 등을 이용해서 데이터를 분석하는 것을 목표한다.

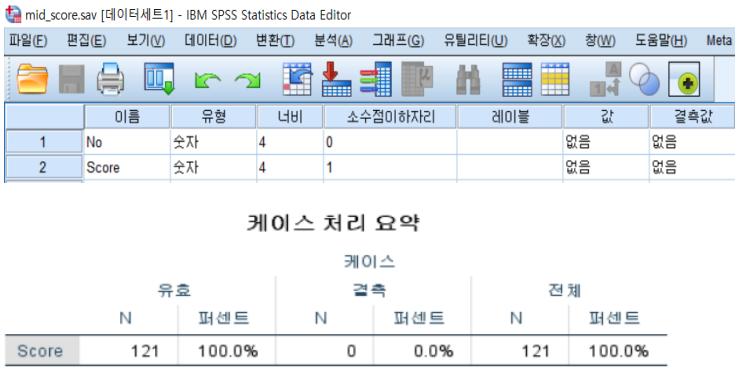
해당 분석이 어떤 데이터에서 가장 의미 있는 해석을 제공할지 고민해보았다. 그 결과, 강의로부터 제공된 2020학년도 봄학기 정보와 정보학 중간고사 성적 데이터가 가장 적절하다고 생각했다. 보편적으로 어떤 시험의 성적 분포가 정규 분포를 따르면 그 시험의 난이도가 적절했다고 판단하는데, 이 데이터에 중심 극한 정리를 적용해 봄으로써 이번 중간고사 문항의 난이도와 성적에 대한 해석을 시도하고자 한다.

B. 전처리

i. 클리닝

주어진 중간고사 성적 데이터의 변수에 대한 기본 정보를 확인해보았다.





해당 성적 분포는 총 121명의 학생들의 중간고사 성적으로, 식별자 역할을 하는 No라는 변수와 실제 중간고사 점수인 Score 변수로 구성되어 있다. 양적 변수인 Score는 원칙상 0부터 100까지의 값을 가질 수 있는데, 내림차순으로 정렬된 해당 데이터를 통해서 두 극 값은 존재하지 않음을 확인했다. 변수 정보를 통해서는 Score가 소수점 이하 1자리까지 주어져 있고 두 변수 모두에 대해서 결측치가 없음을 알 수 있었다.

121개의 데이터를 모두 확인해본 결과, 극단적인 값 혹은 0에서 100까지의 범위를 벗어나는 이상치나 기타 잡음치를 발견할 수 없었다. 따라서 본 중간고사 성적 데이터에 대해서는 특정 데이터 행 삭제나 보정값 사용과 같은 데이터 클리닝에 해당하는 전처리가 불필요하다고 판단했다.

대신, No라는 변수는 식별자의 역할일 뿐 분석과 해석에 아무런 의미가 없다고 판단하여 해당 열 전체를 삭제했다.

	Score
1	94.3
2	93.7
3	93.7
4	91.7
5	91.0

또 주어진 표본의 수는 121개로 중심 극한 정리(보통 표본의 크기가 30이상이면 적용함)를 적용하기에 충분한 크기로 보인다.

ii. 가정 도입 및 데이터 변환

주어진 원본 데이터의 비틀림(상하좌우로의 치우침)을 확인하여 분석에 앞서 데이터 변환을 시도할지 판단해보았다. 단순히 표본의 크기가 크다고 해서 어떤 표본의 데이터 자체가 정규 분포에 근사하는 것이 아니라, 표본 평균이 정규 분포에 근사하게 되는 것이기 때문이다.

정규성 검정

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	통계량	자유도	유의확률	통계량	자유도	유의확률
Score	.084	121	.034	.973	121	.015

a. Lilliefors 유의확률 수정

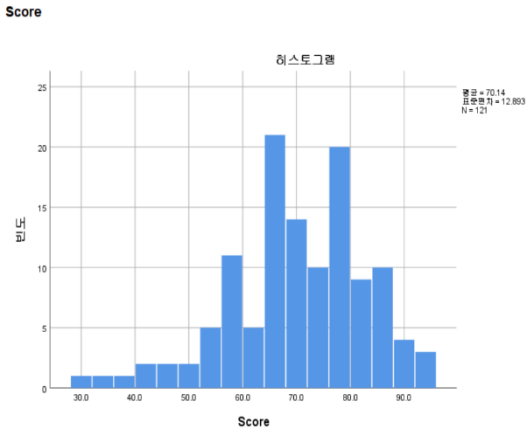
좀 더 중심 극한 정리를 명확히 하자면, 주어진 데이터는 SPSS에서 제공하는 두 정규성 검정 모두에서 유의확률이 0.05 미만인 수준이기 때문에 정규성 검정의 영가설 즉, '데이터 자체가 정규분포이다'를 채택할 수 없지만, 중심 극한 정리에 따라 정규성을 만족하지 않더라도 표본의 크기가 30 이상일 경우 정규 분포를 가정할 수 있다는 것이다.

따라서 여기서 한가지 가정을 해야 한다. 중심 극한 정리는 모집단의 분포와 상관없이 표본의 크기가 커지면 표본 평균이 정규 분포를 따르게 되기 때문에 여기서 표본이 더 커질 경우 정규성을 가정할 수 있음을 다시 상기시켰을 때, 본 분석에서 사용하는 성적 데이터는 표본이 아니라 모집단이기 때문이다. 따라서, 편의상 그리고 중심 극한 정리에 더 적합한 적용을 위해서 주어진 데이터가 121명보다 더 많은 학생들로 이루어진 모집단에서 표본의 크기를 121(30 이상)로 수 없이 랜덤 추출한 표본 평균들이라고 가정한다. 주어진 데이터에서 표본

의 크기를 정하고, 그 크기의 표본을 여러번 랜덤 추출해서 그 표본 평균들의 분포를 만들어 내는 방향도 생각해보았으나, 이는 해당 프로젝트와 강의 범주를 넘어선다고 판단했기에 위의 가정을 도입했다.

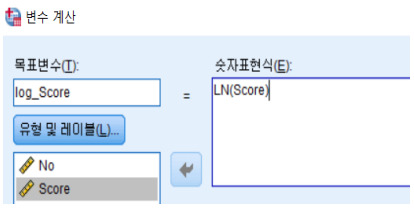
기술통계

			통계량	표준화 오류
Score	평균		70.140	1.1721
	평균의 95% 신뢰구간	하한	67.819	
		상한	72.460	
	5% 절사평균		70.681	
	중위수		70.700	
	분산		166.228	
	표준화 편차		12.8930	
	최소값		29.7	
	최대값		94.3	
	범위		64.6	
	사분위수 범위		15.7	
	왜도		-.608	.220
	첨도		.443	.437



다시 데이터 변환과 관련해서 논의를 해보자면, 위의 가정에 따라 해석했을 때 주어진 데이터 즉, 표본 평균의 분포는 왜도가 -0.608이고 첨도가 0.443인 우측으로 기울어진 뾰족한 분포를 띔을 알 수 있다. 이는 히스토그램을 통해서도 대략적으로 확인할 수 있다.

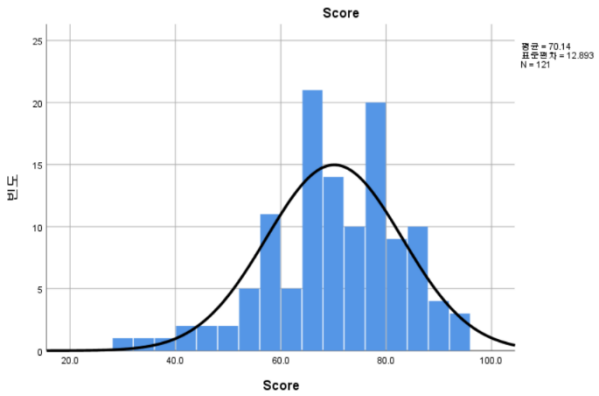
따라서, 중심 극한 정리를 본 분석에서는 주어진 데이터(표본의 크기가 121인 표본 평균의 분포로 가정한 데이터)와 주어진 데이터를 로그 변환한 데이터에 각각 적용하여 분석을 진행하고자 한다. 그 이유는 0보다 큰 변수가 기울어진 경우 로그 변환을 보편적으로 수행하기 때문이다.

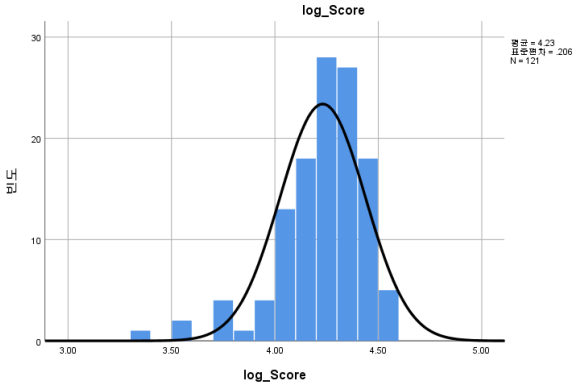


	Score	log_Score
1	94.3	4.55
2	93.7	4.54
3	93.7	4.54
4	91.7	4.52
5	91.0	4.51

C. 연구 방법 및 결과

i. 백분위수 값





두 데이터는 꼬리가 치우쳐져 있고 0미만의 점수가 존재하지 않으므로 히스토그램의 요약에 백분위 수가 좀 더 유용하다. 다만, 위의 히스토그램과 정규 분포 곡선을 비교했을 때, 로그 변환을 거친 성적 변수가 육안 상으로는 좀 더 위로 뾰족한 분포를 보인다. 하지만, 사실 이 현상은 로그 변환을 통해 성적 변수 범주를 축소했기 때문이다.

1. 백분위 수

이 데이터에서 백분위는 자신의 성적이 혹은 자신의 성적 점수대가 특정 상위 혹은 하위 몇 %에 해당하는지를 보여준다. 거꾸로 해석하면, 해당 과목의 성적 기준에 따라 자신이 어떤 성적 등급에 속하는지 혹은 비슷한 수준으로 다음 학기 중간고사가 출제되면 특정 등급에 속하기위해서는 몇 점을 중간고사에서 목표로 해야 하는지를 백분위를 통해서 확인할 수 있다는 것이다. 이러한 백분위 수의 활용과 관련하여 로그 변환을 거친 log_Score는 직관적이지 못하다는 한계가 뒤따를 수밖에 없지만, 100점 만점의 시험 성적이 로그 변환을 거쳐 전체 성적에 합산되는 경우 유용성을 찾을 수 있다.

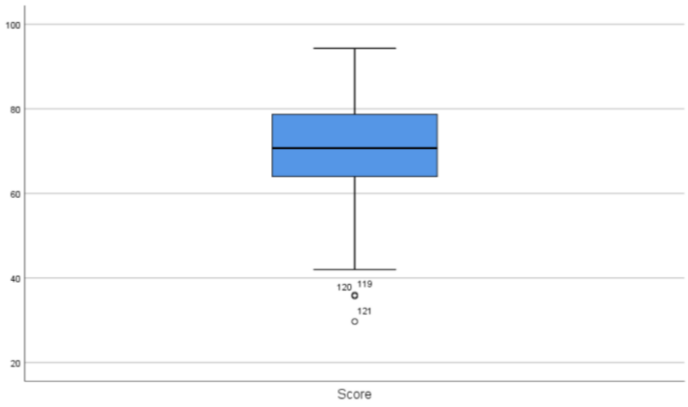
		Score	log_Score
N	유효	121	121
	결측	0	0
평균		70.140	4.2311
중위수		70.700	4.2584
최빈값		78.0	4.36
표준화 편차		12.8930	.20635
분산		166.228	.043
최소값		29.7	3.39
최대값		94.3	4.55
백분위수	25	63.150	4.1454
	50	70.700	4.2584
	75	78.850	4.3675

언급한 내용을 토대로 예시를 들어 해석을 해보자면, 첫번째로 자신의 성적이 백분위로 어느 구간에 해당하는지를 확인할 수 있다. 만약 78점 언저리의 점수를 얻었다면, 대략 상위 25% 혹은 하위 75%의 경계점이라고 생각

할 수 있고, 로그 변환을 거친 환산 점수가 약 4.36 근처일 경우에도 비슷하게 판단할 수 있다. 따라서 만약 정보와 정보학 중간고사에서 상위 25%까지를 A 커트라인으로 한다면, 자신이 A 등급에 속하는지 알기 위해서는 78.850점 이상인지만 알면 된다. 또 내년의 중간고사 난이도가 올해의 난이도와 비슷할 것으로 예상한다면, 그리고 자신이 A 등급을 목표로 한다면 대략 80점 이상을 맞도록 시험 공부를 계획해야 함을 데이터의 백분위 수로부터 해석해낼 수 있다.

2. 사분위 수와 상자 그림

백분위수를 좀 더 집약적으로 확인하기 위해서 사분위 수를 상자 그림으로 나타내 보았다.



파란 상자의 가장 밑변은 63.150으로 하위 25%에 해당하는 점수이며, 가장 윗변은 78.850으로 상위 25%를 나타낸다. 상자 중앙의 짙은 실선은 70.700으로 상위, 그리고 하위 50%에 해당하는 성적을 가시적으로 직관적으로 보여준다. 상자 그림은 이와 같은 사분위 수를 요약하여 보여준다는 장점과 함께, 상자 안의 변수 값들이 전체 분포의 50%를 이루고 있음을 나타낸다. 즉, 자신의 점수대를 세로 축에서 찾아 상자의 위치와 비교했을 때, 상자 아래로 맞닿으면 자신이 하위 25%의 성적임을, 상자 위로 맞닿으면 자신이 상위 25%임을, 상자 중앙의 선 아래에 맞닿으면 하위 25%~50%임을, 그리고 상자 중앙의 선 위에 맞닿으면 상위 25%~50%에 해당하는 성적임을 쉽게 확인할 수 있다는 것이다. 따라서 사분위 수에 따라 성적이 A, B, C, D로 나눠지는 경우는 해당 상자 그림이 더 유용하게 된다.

ii. 중심경향과 표준편차

		Score	log_Score
N	유효	121	121
	결측	0	0
평균		70.140	4.2311
중위수		70.700	4.2584
최빈값		78.0	4.36
표준화 편차		12.8930	.20635
분산		166.228	.043
최소값		29.7	3.39
최대값		94.3	4.55
백분위수	1	31.020	3.4316
	10	53.900	3.9871
	25	63.150	4.1454
	50	70.700	4.2584
	75	78.850	4.3675
	90	86.000	4.4543
	99	94.168	4.5451

먼저, Score 변수부터 살펴보자면 100점 만점에 평균은 70.140이며 표준 편차는 12.8930, 중위수(중앙값)은 70.700으로 평균보다 살짝 큰 값을 보이지만 평균과 매우 가깝다. 최빈값은 78.0으로 누적 퍼센트는 71.9%다. 평균과 중위수가 매우 가깝다는 것에서 시험 성적이 평균을 기준으로 고르게 분포함을 판단할 수 있다. 이는 위의 히스토그램에서 평균을 기준으로 좌우 막대 그래프들이 어느정도 균일한 빈도를 보인다는 것을 통해서도 알 수 있다. 여기서 표본이 커졌을 경우 정규 분포를 가정할 수 있다는 중심 극한 정리를 다시 한번 확인할 수 있다. 보통 100점 만점의 시험의 경우 평균과 중앙값이 60점대 이하이면 시험이 어려웠다고 판단하고, 80점대 이상이면 시험이 비교적 쉬웠다고 판단하는데 위의 데이터에서 평균이 거의 70점이므로 시험의 난이도가 평균과 중앙값의 측면에서 매우 이상적이라고 할 수 있다. 이는 다음의 정보와 정보학 중간고사도 이번 중간고사의 난이도를 목표로 하면 적당한 성적 분포를 얻을 수 있음을 의미한다.

중심 극한 정리에 따라 정규성을 가정하게 되면, 해당 데이터는 평균과 중앙값이 매우 가까우므로 둘 중 어느 기준에서 표준 편차 곱하기 2를 더하거나 뺀 점수대 즉, 대략 95점대나 45점대가 각각 약 상위 혹은 하위 2.5%에 해당한다고 상정할 수 있게 된다.

log_Score는 평균이 4.2311, 중위수가 4.2584, 최빈값이 4.36, 표준 편차가 0.20635로 Score 변수와 마찬가지로 평균과 중위수가 매우 가까우며 최빈값은 중위수를 웃돌고 있다. 히스토그램과 정규 분포 곡선으로 보아, 조금 더 중심이 뾰족한 분포를 보인다. 하지만 사실상, 왜도와 첨도는 더 극대화되었다는 점에서 이 성적 데이터에 적합한 변환은 아니었다는 결론을 얻을 수 있다. 또한, 성적 데이터는 100점 만점을 기준으로 해석을 하기 때문에 로그 변환을 거친

log_Score는 직관성이 떨어진다는 점에서 해석이 덜 용이하다. 하지만, 점수 값이 큰, 예를 들어 990점 만점의 토익 점수 등의 성적을 다른 성적과 합산하기 위해 더 작은 기준으로 변환하여 적용하는 경우 등에서는 로그 변환을 거친 성적을 이용하면 변수 값을 축소할 수 있다는 장점이 있다.

고맙습니다😊