Chapter 01 연습문제

1. 기술통계와 추론통계를 각각의 예를 들어 설명하라.

기술통계(descriptive statistics)는 표본의 모집단에 대한 대표성을 설명하기 위해 연구자가 선택한 표본에 대한 통계량의 각종 수치들로 표본의 특성을 설명하는 것이다. 추론통계(inference statistics)는 기술통계를 통해 모집단의 대표성이 확인된 표본으로부터 모집단의 특성을 추정하여 나타내는 것이다. 기술통계는 표본의 구성, 비율 및 월별, 분기별, 연도별 성장률, 취업률 등에 적용할 수 있고, 추론통계는 모집단을 추정하는 평균, 분산, 표준편차, 왜도, 첨도 등의 다양한 통계지표를 통한 예측, 추론 등에 적용한다.

2. 인문/사회과학에서 통계학을 사용하는 이유와 통계학이 추론에 불과하다는 한계를 극복하기 위한 방법에는 무엇이 있는지 설명하라.

통계학은 응용수학의 한 분야이며 관찰이나 조사를 통해 얻는 불균형적인 데이터에 응용수학의 기법을 이용해 수치상의 성질과 규칙성 또는 불규칙성을 찾아내는 학문이다. 그렇기 때문에 통계적 기법으로 실험 계획과 데이터의 요약이나 해석을 할 때, 근거를 제공하는 학문이다. 근거로 제공하는 수치가 전체를 대상(모수)으로 조사된 값이면 좋겠지만 통계조사는 주로 표본을 통해 이루어지므로 표본은 전체를 대표하는 대표성이 있어야 한다. 결국 표본을 통한 전체에 대한 추론(inference)이기 때문에 추론의 결과가 어느 정도 모수와 일치할 것인가 하는 확률(probabilty)로 보충하여 나타낸다.

추론이기에 정확하지 않다는 한계를 논하지만, 과학적 수치화를 통한 사회현상에 대한 이해의 편익이 질적 연구에 의한 한계보다 크다. 제대로 된 정확한 방법에 의한 분석과 결과의 해석은 과학적 분석의 전제가 되며, 그로 인한 연구의 확장은 그 자체로서 의미가 있다.

3. 다양한 분야에서 통계학을 활용하는 목적이 무엇인지 설명하라.

통계학을 사용하는 목적은 여러 대안 중에서 하나를 선택할 수 있는 의사결정 과정에서 탁월한 답안을 선택할 수 있기 때문이다. 의사결정을 잘 하더라도 그 의사결정이 확실한지 아닌지에 대한 불확실성이 존재한다. 불확실성은 현재 경험적으로 수집된 데이터를 통해 해소할 수 있다. 또한 수집한 데이터를 요약하면서 불확실성을 줄일 수 있으며, 연관성을 파악하여 세부적인 판단을 할 수 있고, 인과관계 파악을 통해 알아낸 변화의 패턴을 하나의 추세로 가늠하는 예측을 할 수 있다. 이러한 이유에 따라 아래와 같은 세부 목적을 가질 수 있다.

① 의사결정

의사결정(意思決定, decision making)이란 여러 대안 중에서 하나를 선택하는 일을 해내는 정신적 지각 활동이다. 모든 의사결정에서는 하나의 최종적 선택을 하며, 이 선택의 결과로 어떤 행동과 그 행동으로 인한 결과가 뒤따른다. 정보처리 관점에서 의사결정은 정보와 반응 사이의 다대일 대응으로 나타난다고 볼 수 있다. 즉, 대개 많은 정보를 지각하고 평가하여 하나를 선택한다. 이러한 경우는 주로 여러 가지 변수나 데이터가 복잡하게 얽혀 있어 자료를

기준으로는 판단하기 어려운 분야에서 많이 활용되고 있다. 특히나 의사결정은 경영학에서 여러 가지 대안 가운데 하나의 안을 선택할 때 사용되는데 이때 통계학은 아주 듬직한 지원군이 될 수 있다.

② 불확실성의 해소
의사결정을 하더라도 '그 의사결정이 정확한 것인가'라는 불확실의 문제가 남는다. 불확실성에는 단순히 리더나 경영자 혹은 권력자가 가지는 어떤 특정 상황의 불확실성도 포함된다. 특히나 비즈니스에서는 기업의 효율성과 안정성을 제고하기 위해 다양한 시도와 솔루션을 개발해 제시하게 되는데, 최근에는 경험적으로 수집되는 데이터를 넘어 빅데이터의 개념을 들여와서 불확실성을 해소하려는 노력을 하고 있다. 더 나아가 기존의 유통과 물류체계에서의 불확실성을 제거하기 위해 통계를 적용하는 등 비즈니스 프로세스 분야에서 생산자 측면에서부터 소비자 측면까지 아우르며 미래 예측까지 접목하는 시도가 이루어지고 있다. 또한 제품의 생산과 판매가 접목되는 과정에서 CRM(customer relationship management), SCM(suply chain management) 등에서 정보 획득의 어려움과 이를 통한 고객 응대와 시장 대응의 어려움이 나타나게 되는데 이러한 약점을 극복할 수 있는 좋은 기반이 될 수 있다.

③ 요약
통계에서 요약이란 데이터의 요약을 의미한다. 데이터를 수집하여 이를 요약하는 이유는 불확실성을 줄이기 위해서다. Excel의 기능 중 피벗테이블 기능은 데이터 요약에 대한 좋은 예가 된다. 비즈니스 과정에서는 매출액이 발생하는데, 지속적으로 반복되어 생산되는 데이터라 하더라도 날짜별, 월별, 분기별 매출액을 정리하여 지역별 혹은 지점별 매출액에 대한 날짜별, 월별, 분기별 매출액이 정리된 보고서가 있다면, 의사결정권자는 비즈니스를 확장해야 할 부분과 축소해야 할 부분에 대한 판단을 할 수 있는 적절한 기준을 제시할 수 있게 된다.

④ 연관성 파악
단순히 매출액 자료만 보고 받았다면 요약 기능으로 끝난다. 하지만 날짜별, 월별, 주별, 분기별 광고비의 지출 내역 대비 매출액 규모를 동시에 정리한 보고서가 있다면 의사결정자는 광고비와 매출액의 연관성을 판단할 수 있다. 목표 매출액을 정하고 광고비, 매장 규모, 청결도 등의 다른 변수들을 조정할 수 있는 경우라면 더욱 다양한 세부적 판단을 할 수도 있다. 즉, 광고비의 단위별 증가나 감소를 원인으로 한 매출액의 증가나 감소를 확인할 수 있다. 통계는 이런 직접적인 연관성을 제시한다.

⑤ 예측
인과관계 파악을 통하여 매출액이 변화하는 패턴을 찾아낸 의사결정자는 이러한 패턴을 하나의 추세로 판단할 수도 있다. 이러한 패턴이 반드시 산술적으로 반복되는 정형화된 결과로 이어질 수는 없겠으나 경영자는 향후에 광고비나 기타 변수의 조작을 통하여 원하는 매출액을 예측해낼 수 있다. 이로부터 다양한 계량 기법을 적용해 여러 변수를 효율적으로 투입하여 최소 비용으로 최대 수익을 얻을 수 있는 조합점을 찾아낼 수 있다.

4. 조사나 연구에 통계분석을 적용하는 과정을 설명하라.
통계 프로세스에서 조사 목적에 맞는 자료를 수집하는 것에서 시작하여 분석에 적합한 자료를

선별하여 조사 목적과 자료가 적합하도록 만드는 정제 과정을 거친 후 정제된 데이터를 분석하는 것이 추정단계다. 그리고 조사의 목적에 따른 귀무가설과 대립가설의 채택 여부를 판단하는 검정단계다.

1) 자료의 수집
조사 목적에 맞게 자료를 수집하는 단계다. 자료의 수집은 자연스럽게 또는 설계를 통해 이루어진다. 자료의 성격에 따라 1차 자료와 2차 자료로 구분되며, 수집하고자 하는 자료의 성격에 따라 설계 단계에서 통계분석의 방법이 결정될 수 있다.

* 1차 자료 – 조사자가 직접 수집한 자료. 조사자가 직접 측정 도구를 설계 및 개발하며, 목적에 가장 부합하는 자료 수집 방법이다.
EX) 설문지, 우편, 전화 질의, 인터뷰 등

* 2차 자료 – 조사자가 목적을 가지고 직접 수집한 자료는 아니지만, 조사 목적에 맞아 활용할 수 있는 자료. 목적을 가지고 설계되어 측정된 자료가 아니므로 활용 가능성, 적합성, 신뢰성 등의 사전평가가 매우 중요.
EX) 학술 자료, 정부 간행물, 연구 보고서, 사내 자료 등

2) 자료의 정제
수집된 자료 중에서 분석에 적합한 자료를 선별하는 과정이다. 조사 목적과 자료가 서로 적합하도록 만드는 과정이므로 매우 중요하다.

3) 추정
통계학은 결국 모수를 추정하는 것이다. 표본을 분석하여 그 표본 대상의 특징을 설명하는 하는 목적도 있지만, 결국 궁극적으로 모수를 추정하는 데 목적이 있다.

4) 검정
믿어지는 사실이 실제로 옳은지 아닌지를 확인하기 위함이 조사나 연구의 목적이 된다. 이를 위해 연구방법과 통계를 통해 가설을 수립하고 검정을 한다. 검정은 수립된 가설에 유의미한 타당성이 있는지 통계적으로 확인하는 과정이다.

5. 추정과 예측을 비교하여 설명하라.
추정(statistical estimation)은 표본으로부터 얻은 통계량으로부터 모수를 추정하는 것이다. 대개의 경우에 있어 모집단에 대한 분석이 불가능하므로 모집단으로부터 표본을 구성하고 표본을 조사하여 얻어진 수치로 모수를 추정하는 것이다.

예측(statistical prediction)은 추정을 반복하여 얻어진 결과들을 기준으로 일정한 패턴을 찾아낸 후, 향후 미래에 활용할 수 있는 의미 있는 특정한 모수를 가늠하는 것이다. 예를 들어, 일기예보에서는 강수확률을 예측할 때 과거의 자료를 기준으로 확인하고 일정한 패턴을 매년 측정하여 내일(미래)의 사건을 예측한다.

6. 통계분석에 한계가 있는데도 통계분석을 하는 이유를 예를 들어 설명하라.

일부에서는 이러한 한계를 지적하며 '부정확한 것의 일반화'라 하는 경우도 있다. 하지만 모집단에 대한 조사가 불가능하다고 해서 신포도(sour grapes)로 취급하는 것이 더 문제다. 적극적으로 한계를 극복하려는 시도와 오류를 배제하려는 노력이 더 많은 학문적 결과와 인문사회과학의 발전을 가져올 것이다.

7. [컴퓨터 실습] Excel의 [데이터] 탭에 [데이터 분석] 메뉴가 없는 때 메뉴를 어떻게 노출해야 하는지 설명하라.

파일-옵션-추가기능-분석도구, 관리: 이동-분석도구:확인

맥에서는 도구-추가기능-분석도구 체크

Chapter 02 연습문제

1. 모집단, 표본, 모수, 통계량을 각각 예를 들어 설명하고, 표본과 통계량을 조사하는 이유를 설명하라.

모집단(population)이란 통계분석 방법을 적용하는 관심 대상의 전체 집합을 말한다. 통계학을 적용하여 조사할 때, 모집단에 대한 조사를 진행하기 불가능한 경우가 많다. 예를 들어, 모든 여성 대한민국국민, 202X년에 수입된 모든 쇠고기, A쇼핑몰 회원 전체, B통신회사 전체 가입자는 모집단에 대한 예이지만 모든 여성 대한민국국민, 202X년에 수입된 모든 쇠고기는 유한한 규모이므로 인력과 시간을 최대한 투입하면 원하는 모든 통계자료를 계산할 수 있다. 하지만 모든 여성 대한민국국민, 202X년에 수입된 모든 쇠고기에 대해 전수를 조사하는 것이 좋기는 하겠지만 비용이라는 벽에 부딪히게 될 것이므로 전수조사를 진행하기란 쉽지 않다. 그러므로 모집단을 대표할 수 있을 정도의 과학적인 방법으로 추출한 모집단의 일부분에 대한 조사를 실시하게 되는데, 모집단의 일부분을 표본(sample)이라 한다. 모수(parameter)란 모집단을 분석하여 얻어진 특성을 나타내는 수치이다. 알고자 하는 평균, 분산, 표준편차, 비율 등의 모집단의 특성을 모평균(μ), 모분산(σ^2), 모표준편차(σ), 모비율(p)이라 한다. 통계량(statistic)이란 표본을 분석하여 얻어진 특성을 나타내는 수치이다. 보통 모집단을 대상으로 한 분석이 불가능하므로 표본을 분석하여 얻어지는 수치를 의미하며, 알고자 하는 평균, 분산, 표준편차, 비율 등의 표본의 특성을 표본평균($\bar{x}$), 표본분산(s^2), 표본표준편차(s), 표본비율($\hat{p}$)이라 한다.

2. 평균, 분산, 표준편차, 비율에 대해 모집단과 표본에서 표시하는 방법이 다르다. 표시 방법을 다르게 하는 이유를 설명하고, 각각의 표시 방법을 나열하라.

평균, 분산, 표준편차, 비율은 모집단과 표본에서 각각 구할 수 있는데, 일반적으로는 모수를 알 수 없으므로 통계량과 비교를 하기 위해 구분을 한다. 모수는 모집단으로부터 계산되며 모평균, 모분산, 모표준편차, 모비율로 나타내며, 각각의 표현 방법은 모평균 : μ, 모분산 : σ^2, 모표준편차 : σ, 모비율 : p로 나타낸다.

통계량은 표본으로부터 계산되며 표본평균, 표본분산, 표본표준편차, 표본비율로 나타내며, 각각의 표현 방법은 표본평균 : $\bar{x}$, 표본분산 : s^2, 표본표준편차 : s, 표본비율 : $\hat{p}$로 나타낸다.

3. 모집단을 대상으로 하는 조사가 가장 정확함에도 불구하고 실제 연구 조사에서 표본을 대상으로 연구하는 이유를 설명하라.

가장 정확한 조사는 모집단을 조사하는 것이다. 하지만 대부분의 인문사회과학에서는 표본으로 조사를 진행하는데, 그 이유는 모집단의 구성이나 범위가 너무 넓어 조사 자체가 힘든 경우가 대부분이며, 농수산물, 배터리, 전구 등의 제품처럼 사용을 하게 되면 없어지거나 가치가 달라지는 대상은 전수조사의 대상이 될 수 없는 경우도 있다. 때문에 표본조사는 조사

자의 입장에서 시간과 비용 대비 효율성이 가장 큰 조사방법이다.

4. 확률적 표본추출과 비확률적 표본추출을 구분하는 방법과 그 각각의 기준을 설명하라.

확률적 표본추출이란 모집단으로부터 표본을 추출할 때 표본으로 선택될 확률이 모두 동일한 방식이다. 즉, 모집단을 구성하는 대상들 중 일부분이 표본을 구성하게 되는데, 표본으로 선택될 확률이 모두 동일해야 한다. 이와 반대로, 비확률적 표본추출은 모집단으로부터 표본을 추출할 때 표본으로 선택될 확률이 서로 다른 표본추출방법이다. 조사자가 판단하여 확률적 표본추출을 하기 힘들거나 불가능한 경우에 사용한다. 주로 표본을 추출할 수 있는 특수한 상황 하에서 가장 적합하게 조사할 수 있는 표본을 구성하는 방법이다.

확률적 표본 추출 : 표본으로 추출될 확률이 동일해야 함

1) 단순 무작위 표본 추출
   모집단에서 일정한 규칙에 따라 기계적으로 추출하는 방법
   Ex> 난수표, 컴퓨터 추출
2) 체계적 표본추출
   모집단을 대상으로 각각의 경우에 번호를 부여하고 일정한 순서대로 n개의 간격을 정해서 표본을 추출하는 방법
   Ex> 선거 출구조사
3) 비례 층화 표본추출
   모집단을 여러 개의 이질적 집단으로 구분한 후, 각 집단의 구성 개수에 따라 비례하도록 추출하는 방법
   Ex> 모 고등학교 1 2 3 학년 비율 1 : 2 : 3 일 때, 60명 추출 시 각각 학년별 10명, 20명, 30 명으로 표본을 구성하는 방법
4) 다단계 층화 표본추출
   비례 층화 표본추출에서 상·하위 표본단위를 미리 설정하고 그에 맞추어 다시 추출하는 방법
   Ex> 모 회사 총원 50 명 50명을 표본 추출 시 첫 번째로 부서별로 먼저 구분한 후 팀별로 구성 숫자를 맞추어 추출하는 방법
5) 군집 표본추출
   모집단의 구성을 살펴보았을 때, 내부 이질적이면서 외부 동질적으로 구성되어 있다면 모집단 전체를 조사하지 않고 몇 개의 군집을 표본으로 선택해서 조사하는 방법
   Ex> 서울 시민 대상으로 서울시장 만족도 조사시 표본으로 몇 개의 구를 선택. 선택된 구를 표본으로 선정하여 조사하는 방법

비확률적 표본 추출 : 표본으로 추출될 확률이 서로 다른 표본추출 방법

1) 편의 표본 추출
   조사자가 자신의 편의에 따라 시간, 장소에 구애받지 않고 임의적으로 표본을 추출하는 방법
2) 판단 표본추출
   조사자의 판단에 따라 적합하다 생각되는 구성원들을 표본으로 선택하는 방법
3) 할당 표본추출

모집단의 속성을 대표할만한 연령, 학력, 직업 등 구분을 결정하고, 각각에 대한 표본 개수를 미리 결정한다. 이후 조사자가 결정한 표본의 개수에 따라 임의적으로 표본을 추출하는 방법

4) 자발적 표본추출

조사자의 의지와는 별도로 응답자가 원하여 조사에 응하는 경우를 표본으로 선택하는 방법

5. 다음의 조사에서 모집단 조사와 표본조사 중 어떤 방법을 사용하는 것이 더 적합한지 결정하고, 표본조사인 경우 어떤 방법으로 표본을 구성할지 설명하라.

(a) 선거 후보자들에 대한 지지도 조사=확률적 표본추출(비례층화 표본추출)

(b) 음료수 용기의 250ml의 용량이 맞는지에 대한 조사
=확률적 표본추출(체계적 표본추출)

(c) 대학생의 스마트폰 만족도에 대한 조사=확률적 표본추출(단순무작위 표본추출)

(d) A대학교 신입생의 한 달 용돈 금액 조사=확률적 표본추출(체계적 표본추출)

(e) 대한민국 성인의 1일 SNS 이용시간에 대한 조사=확률적 표본추출

(f) SNS이용자의 이용의도와 만족도에 관한 조사=확률적 표본추출

(g) 프랜차이즈 패밀리 레스토랑과 일반 패밀리 레스토랑에 대한 음식맛, 친절도, 만족도, 선호도에 대한 조사=비확률적 표본추출

(h) 청년창업가의 진취성, 혁신성, 위험 감수성에 대한 조사=비확률적 표본추출

6. z분포와 t분포의 개념과 두 분포의 관계를 설명하라.

z분포란 표본의 개수가 충분하면서 표준화과정을 거친 정규분포이며, 표준정규분포라고도 한다. 또한 '평균=0, 분산=1'인 정규분포를 따른다.

t분포란 표본의 개수가 30개를 넘지 못하는 경우 t 분포를 사용하며, t분포도 '평균=0, 분산>1'인 분포를 가진다.

이 두 분포의 공식을 보면 다음과 같다. $z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$과 $t_{n-1}=\frac{\bar{x}-\mu}{s/\sqrt{n}}$이므로 n과 n−1을 제외하면 동일하다. t분포에서 n−1이라는 자유도를 사용하는 이유는 분산을 구할 때 평균이 미리 정해져 있는 경우라면 변수들 중 평균값을 맞추기 위해 하나는 특정한 고정값으로 정해져 자유를 상실하기 때문에 자유도는 n−1이 된다.

7. 표본분산의 확률분포에는 어떤 것이 있으며, 이들이 어떤 경우에 활용되는지 설명하라.

표본분산과 관련된 확률분포는 χ^2분포와 F분포가 있으며, χ^2분포는 1개의 분산(표본)으로 추론하는 것에 비하여 F분포는 2개의 분산(표본)으로 추론하는 분포이다.

χ^2분포는 한 개의 분산에 대한 추론이며 정규분포로부터 도출된다. Z분포의 제곱에 대한 분포이므로 항상 0보다 큰 값을 가지게 된다. χ^2분포는 주로 모분산의 추정이나 계수값에 대한 해석을 위해 주로 사용한다.

F분포는 두 개의 분산에 관한 추론이기에에 $F_{(v_1, v_2)}$로 나타내며 v_1, v_2는 각각의 χ^2에 대한 자유도를 의미한다. F분포는 주로 분산의 동일성 여부를 판단하는 수단으로 사용한다.

8. 표본으로부터 모수를 추정할 때, 오차를 줄이기 위해서 조사자가 할 수 있는 방법에 대해 설명하라.

표본으로부터 모수를 추정할 때, 모수와 통계량 간의 차이를 표본 평균의 오차라 한다. 오차를 아예 줄일 수는 없겠지만 최소한으로 하는 것이 중요하다. 이러한 오차를 줄이기 위해서는 표본의 수를 늘려서 표본이 모수의 특성을 잘 반영하게 하는 것으로 오차를 줄일 수 있다.

9. 중심극한정리의 개념에 대해 설명하고, 표본의 개수는 어느 정도로 해야 적절한지 설명하라.

중심극한정리란 모수를 모르는 상황에서 표본통계량으로 표본의 개수가 충분하다면 정규분포를 구성하여 모수 추정이 가능하다는 것이다. 모수를 모르지만 표본의 개수(n)이 충분하다면 정규분포를 구성하여 모수를 추정할 수 있다는 가정으로 모집단이 정규분포를 이루지 않아도 표본의 수가 충분하다면 정규분포를 이루게 된다는 것이다. 정규분포를 구성하는 표본의 최소개수는 30개이다.

Chapter 03 연습문제

1. 범주형 척도의 의미를 설명하고, 그 종류를 예를 들어 설명하라.

범주형 척도(categorical scale)는 명목척도와 서열척도로 나누어진다. 명목척도(nominal scale)는 수(數) 또는 순서의 개념과는 상관없이 이름만 붙여지는 척도이다. 설문 문항에 대한 보기는 대부분 '① 남자 ② 여자'로 주어지며, 응답자는 이에 대해 '①' 혹은 '②'로 답을 한다. 이때 '남자'나 '여자'는 '1'과 '2'라는 숫자 또는 연산이라는 개념과는 아무런 연관이 없다. 즉 '남자+남자'를 수식으로 나타내자면 '1+1=2'이지만 남자를 두 명 더하더라도 하여 '여자'가 될 수는 없다는 의미이다. 서열척도(ordinal scale)는 '순서척도'라고도 한다. 서열척도 역시 숫자 혹은 연산과는 관련이 없고, 단순하게 순서(서열)를 구분하기 위해 만들어진 척도를 의미한다. 마라톤 경기 결과처럼 '1등', '2등', '3등' 혹은 '금메달', '은메달', '동메달'을 구분할 때, '1등+2등'이 '3등' 혹은 '은메달+동메달'이 '금메달'이 될 수 없다. 이때 주의해야 할 것은 '2-1=1'이고 '3-2=1'로서 모두 1이라는 차이가 나지만, 1등과 2등 사이의 시간적 차이가 2등과 3등 사이의 시간 차(혹은 거리)와 같다고 할 수 없다는 점이다.

2. 연속형 척도의 의미를 설명하고, 그 종류를 예를 들어 설명하라.

연속형 척도(continuous scale)는 등간척도와 비율척도로 나누어진다. 등간척도(interval scale)는 명목척도나 서열척도와 달리, 측정된 자료들 간에 더하기와 빼기가 가능한 척도를 의미한다. '섭씨 영상 15도'인 경우를 생각해 보자. 온도는 '0도 보다 15도만큼 높은 온도'이다. 그러나 0도라는 것은 영상과 영하의 구분점이 되는 온도이지 온도 자체가 없는 무(無)를 의미하는 '절대 0'의 개념이 아니다. 비율척도(ratio scale)는 등간척도의 성질과 함께 무(無)의 개념인 0 값도 가지는 척도를 의미한다. '길이', '무게', '부피', '경력' 등과 같이 다양한 기준을 비율척도로 활용할 수 있다.

3. 중심경향도가 무엇인지 설명하고, 중심경향도를 나타내는 지표들의 개념을 설명하라.

중심경향도란 데이터들을 종합하여 그 중심을 이루는 값은 어느 정도가 될 것인지를 확인하는 값이다. 집단특성을 대표적으로 표현하는 값에는 평균, 중간값, 최빈값 등이 있으며 이를 중심경향도(measure of central tendency)라 한다. 평균 : 통계에서 가장 많이 활용되는 중심경향도이다. 모든 통계분석에서 사용되며 표본의 특성을 제시할 때 가장 먼저 사용되는 수치다. 중간값 : 관측된 자료의 편중과는 상관없이 가장 작은 값에서 가장 큰 값까지 정렬한 후 이들의 좌우의 중간값이다. 관측된 자료가 홀수 개라면 표본 중의 값이 선택되겠지만, 짝수 개라면 중간 2개의 표본값에 대한 중간값을 찾으면 된다. 최빈값 : 표본에서 가장 많이 나타나는 관측치를 의미한다. 여러 번 확인된다는 특성으로 중심경향도에 있으나, 최소부분과 최대부분으로 쏠림현상이 나타날 수도 있기 때문에, 실제로 특별한 필요성이 없는 한 잘 사용하지 않는다.

4. 중심경향도 이외에 산포도까지 파악하는 이유를 예를 들어 설명하고, 산포의 정도를 나타내는 지표들의 개념을 설명하라.

중심경향도의 결과를 확인하는 것만으로도 어느 정도 집단에 대한 성격과 분포를 파악할 수 있지만, 이렇게 분석된 결과만으로 결론을 내린다면 실수를 범할 수 있다. 측정된 데이터가 어떻게 분포하고 있는지에 대해 파악해야 최종적인 실수를 줄일 수 있기 때문에 산포도를 파악해야 한다.

예를 들어 4개씩 구성된 표본 A와 B의 2개가 있는데, A=0, 100, 0, 100, B=50, 50, 50, 50으로 구성되어 있다. A와 B의 중심경향도인 평균을 보면 모두 50이지만, B그룹에서는 평균=50에서 직선으로 나타날 것이며, A는 0과 100에 각각 멀리 떨어진 산포를 나타낸다. 통계학에서 산포의 정도를 나타내는 지표는 분산, 표준편차, 범위, 사분위수, 백분위수가 있다.

분산 : 모집단의 모분산, 표본의 표본분산이 있다. 모분산(population variance)은 모평균과 모집단의 개별 측정치들 간의 차를 구해서 제곱을 하여 모두 더한 후 다시 모집단을 구성하는 갯수로 나누어 계산한다. 또한 본을 선정해서 표본의 개수 갸 로 계산한 분산을 표본 분산(sample variance)이라 한다.

표준편차 : 모표준편차와 표본표준편차가 있으며, 분산과 편차의 개념은 평균으로부터 측정치들이 어느 정도 흩어져 있는지의 정도를 나타내는 것이다. 다만, 편차는 평균을 기준으로 음(−)과 양(+)으로 흩어져서 총합이 0이 되니, 이를 피하기 위하여 편차에 제곱을 하는 것이라 했다. 수학적 계산에 의해 제곱합에서 분산을 구했으니, 이를 다시 제곱 이전으로 되돌리는 방법으로 표준편차를 구한다. 5. n−1이라는 자유도를 사용하는 이유는 분산을 구할 때 평균이 미리 정해져 있는 경우라면 변수들 중 평균값을 맞추기 위해 하나는 고정값으로 하나의 변수는 고정되어 자유를 상실하므로 자유도는 총변수−1이 된다.

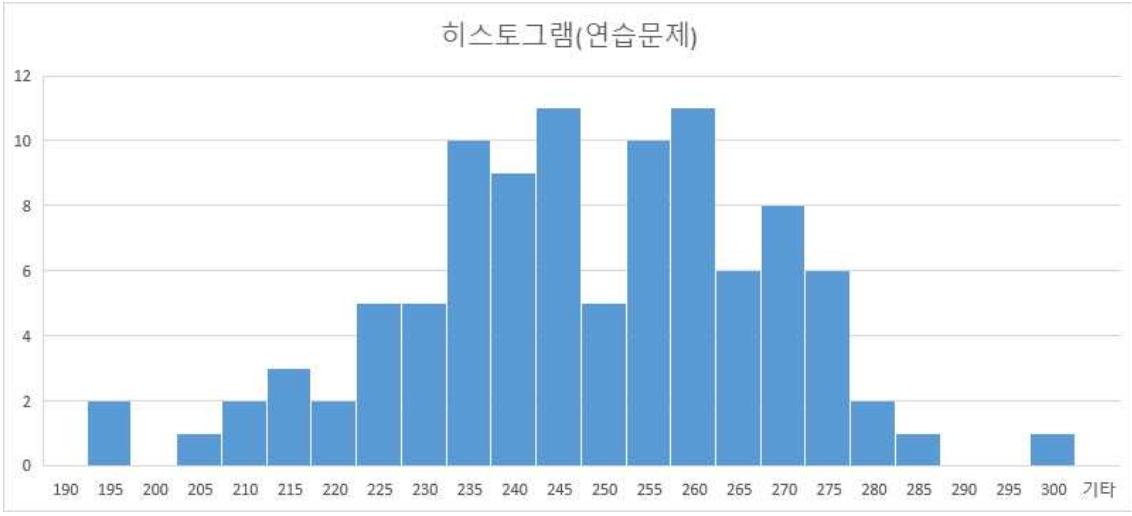
5. 통계량을 계산할 때 자유도를 이용하여 계산하는 이유를 설명하라.

n−1이라는 자유도를 사용하는 이유는 분산을 구할 때 평균이 미리 정해져 있는 경우라면 변수들 중 평균값을 맞추기 위해 하나의 변수는 고정되어 자유를 상실하기 때문이다. 따라서 자유도는 '표본 개수−1'이 된다.

6. [컴퓨터 실습] 연습문제.xlsx의 '3장_전선길이' 시트는 공사에 활용하기 위해 눈대중으로 자른 전선의 실제 길이를 측정한 수치이다. 190cm부터 300cm까지 5cm 단위로 도수분포표를 출력하라.

계급	빈도수
190	0
195	2
200	0
205	1
210	2
215	3
220	2
225	5
230	5
235	10
240	9
245	11
250	5
255	10
260	11
265	6
270	8
275	6
280	2
285	1
290	0
295	0
300	1
기타	0

7. [컴퓨터 실습] [연습문제 6]에서 구한 도수분포표를 히스토그램으로 나타내라.



8. [컴퓨터 실습] A기업에서 근무 교육의 성과를 파악하기 위해 평가 점수의 상승폭을 조사했다. 총 10명의 직원에 대해 조사한 결과, 직전 고과 점수보다 각각 5, 5, 10, 2, 15, 20, 18, 13, 10, 11점만큼 향상되었다. Excel을 이용하여 A기업 근무 교육 성적의 상승폭에 대한 평균, 중간값, 최빈값을 계산하라.

번호	상승폭
1	5
2	5
3	10
4	2
5	15
6	20
7	18
8	13
9	10
10	11
평균값	10.9
중간값	10.5
최빈값	5

9. [컴퓨터 실습] 연습문제.xlsx의 '3장_전선길이' 시트를 기준으로 모분산과 표본분산을 구하고, 이 두 값에 차이가 발생하는 이유를 설명하라. 또한 어떤 경우에 모분산과 표본분산을 사용하는지 설명하라.

모분산	405.3219
표본분산	409.4161

모분산의 경우 표본갯수 n 그대로 이지만, 표본분산의 경우 특정 값을 정하는 고정값으로 인해 자유를 상실하게 되어 n-1의 값을 가지게 되는데 이를 자유도라 한다. 따라서 자유도로 인해 모분산보다 분모의 값이 작아지므로 분산은 값은 커지게 되며, 분산의 증가는 표본분산 안에 모분산이 속할 확률이 높아진다는 것을 모분산에 대한 추정 확률 또한 높아진다. 모집단 전체를 조사하여 모분산을 사용하는 것이 정확한 답을 도출 할 수 있지만, 인문사회과학에서 모집단을 조사하는 것이 쉽지않기 때문에 표본분산을 구하여 모분산을 추정한다.

Chapter 04 연습문제

1. 확률함수의 개념을 설명하고, 실제 어떻게 활용할 수 있는지 예를 들어 설명하라.

미래에 발생할 사건에 대하여 확률을 나열한 것이 확률분포이다. 확률분포는 표나 그래프로 나타낼 수도 있다.

Ex> 월드컵 기간의 치킨 판매량에 대한 데이터와 당시의 판매상황에 대한 정보를 수집하여 향후 매출에 대한 예측을 통해 최대 매출 및 최적의 수익 획득

2. 7개의 사건에 대한 확률분포가 다음과 같이 주어졌을 때, 이 사건의 평균과 분산을 구하라.

사건	1	2	3	4	5	6	7
확률	0.04	0.18	0.19	0.24	0.18	0.13	0.07

평균(기대값) :

$$0.04\times\frac{1}{7}+0.18\times\frac{1}{7}+0.19\times\frac{1}{7}+0.24\times\frac{1}{7}+0.18\times\frac{1}{7}+0.13\times\frac{1}{7}+0.07\times\frac{1}{7}=0.147$$

분산 :

$$(\frac{1}{7}-0.147)^2\times0.04+(\frac{1}{7}-0.147)^2\times0.18+(\frac{1}{7}-0.147)^2\times0.19+(\frac{1}{7}-0.147)^2\times0.24$$
$$+(\frac{1}{7}-0.147)^2\times0.18+(\frac{1}{7}-0.147)^2\times0.13+(\frac{1}{7}-0.147)^2\times0.07=1.768$$

3. 정부에서 청년 창업 스타트업 패키지를 통해 새로운 아이디어를 가진 성인 남녀 35세 미만의 청년 기업 100개를 선발하여 지원하려고 한다. 이를 위해 1년 차부터 5년 차까지 창업 후 생존율을 확인하고자 한다. 창업한 기업의 존속 여부의 평균과 분산을 구하고, 평균과 분산이 나타내는 수치가 의미하는 것을 설명하라.

연차	1	2	3	4	5
생존율 (%)	0.89	0.68	0.59	0.42	0.31

이 문제는 연차를 별도로 가중하지 않아야 하는 단순한 평균과 분산에 대한 문제이다.

평균 : $\frac{0.89+0.68+0.59+0.42+0.31}{5}=0.578$

분산 : $\sigma^2=\frac{1}{n}\sum(x_i-\overline{x})^2\fallingdotseq0.041$

표본분산으로 계산하지 않은 이유는 정부에서 지원하여 관리되는 100의 업체는 모두 전수조사될 것이기 때문에 모분산으로 계산했으며, 5년까지 약 57%의 생존율이 평균이고, 평균에서 예외로 될 경우가 0.041%라는 의미이다. 표준편차로 확인하면 0.20으로 2.5년 내에는 약 77%가 생존할 것이고 5년까지는 약 28% 정도가 생존할 수 있음을 추정할 수 있다.

4. COVID19와 같은 전염병으로 인해 새롭게 뉴 노멀(New normal)에 적응해야 하는 상황이지만, 전 세계적인 협력 끝에 이 상황을 극복하고자 한다. 7개의 크고 작은 제약회사에서 치료제를 개발하고자 하는데, 개발기간은 역량에 따라 1년의 시간이 필요하다고 한다. 각자 독자적인 방법을 통해 1개사부터 7개사가 개발할 확률이 아래와 같다고 할 때, 개발이 되는 평균과 표준편차를 구하라.

개발 완료	1	2	3	4	5	6	7
확률	0.11	0.32	0.21	0.15	0.11	0.09	0.01

평균 : $0.11\times1+0.32\times2+0.21\times3+0.15\times4+0.11\times5+0.09\times6+0.01\times7=3.14$

분산 : $(1-3.14)^2\times0.11+(2-3.14)^2\times0.32+(3-3.14)^2\times0.21+(4-3.14)^2\times0.15$
$+(5-3.14)^2\times0.11+(6-3.14)^2\times0.09+(7-3.14)^2\times0.01\fallingdotseq2.3$

표준편차 : $\sqrt{2.3}\fallingdotseq1.517$

5. SNS를 이용하여 친구들과 소통하는데, 글을 남기면 많은 댓글이 달린다. 다음은 최근 1달 동안 1시간에 달리는 댓글 수를 평균적으로 나타낸다.

댓글 수	1	2	3	4	5	6	7	8	9
확률	0.02	0.05	0.07	0.15	0.22	0.20	0.14	0.08	0.02

① SNS에 달리는 댓글의 평균은?

$1\cdot0.02+2\cdot0.05+3\cdot0.07+4\cdot0.22+5\cdot0.22+6\cdot0.20+7\cdot0.14+8\cdot0.08+9\cdot0.02$

$=5.03$

② SNS에 달리는 댓글의 분산은?

$(1-5.03)^2\cdot0.02+(2-5.03)^2\cdot0.05+(3-5.03)^2\cdot0.07+(4-5.03)^2\cdot0.15+(5-5.03)^2$

$\cdot0.22+(6-5.03)^2\cdot0.2+(7-5.03)^2\cdot0.14+(8-5.03)^2\cdot0.08+(9-5.03)^2\cdot0.02=2.984$

③ SNS에 시간당 10개 이상의 댓글이 달리는 확률은?

$1-(0.02+0.05+0.07+0.15+0.22+0.20+0.14+0.08+0.02)=0.05$

6. 전 세계적으로 K-pop의 인기기 높고, 국내 가전 및 자동차용 배터리 기술에 대한 만족도도 매우 높다. 이 기회에 대한민국이라는 국가 브랜드의 가치를 더욱 높이려고 하는데, 현재의 경기 상황이 썩 좋기만 한 것은 아니다. 그래서 글로벌 경제 상황을 침체, 유지, 상승의 세 가지 국면으로 구분하고, 다양한 경기선행지수를 바탕으로 세 가지 국면에 대한 확률을 도출했다. 국내 투자자들이 국내에 투자하는 것과 해외에 투자하는 것으로 수익을 낼 수 있는 확률을 가지고 있다면 국내에 투자하는 것이 좋을지 해외에 투자하는 것이 좋을지

다음 표를 보고 판단하라.

경기별 확률 \ 투자 지역		국내 투자		해외 투자	
		수익	확률	수익	확률
경기 침체	0.7	100	0.3	200	0.2
경기 유지	0.5	200	0.4	400	0.4
경기 상승	0.2	400	0.5	800	0.6

국내 투자의 기대값 : $400(0.7\times0.5)+200(0.5\times0.4)+100(0.2\times0.3)=186$
해외 투자의 기대값 : $800(0.7\times0.6)+400(0.5\times0.4)+200(0.2\times0.2)=424$

국내 투자의 기대값 :
$(100-186)^2\times0.2\times0.3+(200-186)^2\times0.5\times0.4+(400-186)^2\times0.7\times0.5=16,511.56$
해외 투자의 기대값 :
$(200-186)^2\times0.2\times0.2+(400-186)^2\times0.5\times0.4+(800-186)^2\times0.7\times0.6=61,500.16$

기대값을 계산하면 국내 투자의 경우 186이고 해외 투자는 424이다. 그렇다면 당연히 기대값이 높은 해외 투자를 하는 것이 맞을 것이다. 그런데 분산을 확인해 보면 국내 투자는 16,511.56이고 해외 투자는 61.500.16으로 계산되었다. 기대값은 해외 투자가 약 2.3배가 크지만, 분산의 경우는 3.7배가 넘는다. 분산이 크다는 의미는 편차가 그만큼 클 수 있다는 의미다. 그런데 표준편차로 환산해보면 국내 투자는 약 128, 해외 투자는 약 248이 계산됨을 알 수 있다. 해외 투자의 경우가 국내 투자의 편차를 감안하더라도 높은 것을 알 수 있다. 때문에 해외 투자를 할 것이다.

Chapter 05 연습문제

1. 불확실한 미래에 대한 위험을 줄이는 의사결정과 확률분포의 관계를 설명하라.

모든 미래는 불확실하기 때문에 미래를 알수 있다는 것은 불확실성을 줄일 수 있고, 불확실성을 줄인다는 것은 위험을 줄인다는 것이기 때문이다. 불확실성을 줄인다는 것은 예측가능성을 높일 수 있다는 것이고, 우리의 생활 모든 부분이 이러한 위험을 줄인 상황에서의 의사결정으로 이루어져 있다. 이러한 의사결정은 많은 경험의 축적이 이루어져 데이터로 정보를 얻을 수 있다면, 충분한 경험을 바탕으로 올바른 판단을 할 확률이 높아진다.

2. 균등분포의 개념과 특성에 대해 설명하라.

균등분포(uniform distribution)는 시간에 상관없이 일정한 분포이므로 과거의 경험이 미래를 예측하는데 어떤 도움도 되지 않는 분포를 말한다. 균등분포는 변수의 특성에 따라 이산균등분포(discrete uniform distribution)와 연속균등분포(continuous uniform distribution)의 두 가지로 구분되는데, 이산균등분포는 확률분포함수가 정의된 구간에서 모든 확률이 동일한 분포이며, 연속균등분포는 특정 범위 내에서 동일한 확률분포를 가지는 경우를 말한다.

3. 정규분포와 표준정규분포의 개념과 특성에 대해 설명하라.

축적된 데이터를 기준으로 미래의 예측이 가능한 분포가 정규분포(normal distribution)이다. 정규분포는 통계학에서 가장 많이 사용하는 분포이며, 평균과 분산만으로 그 특성을 모두 설명할 수 있으므로 아주 편리하게 사용할 수 있다. 평균을 중심으로 좌우 대칭인 종(鍾)의 모양을 하고 있으며 분포가 모여 있으면 분산이 작고, 넓게 퍼져 있으면 분산이 큰 것을 나타낸다. 정규분포로 된 분포들을 서로 비교하는 경우가 있는데, 아무리 정규분포라 하더라도 서로 분포가 다르므로 비교하는 것이 상당히 곤란하다. 여러 개의 분포를 어떤 하나의 기준을 세우고 같이 모아 놓는다면, 세워진 기준 아래서 각 분포들을 비교가 가능하다. 즉, 표준정규분포를 구성하는 기준이 되는 평균=0, 표준편차=1로 다시 재구성 한 것이 표준정규분포이다.

* 정규분포

균등분포와 달리 축적된 데이터를 기준으로 미래의 예측이 가능한 분포를 정규분포라 하며 평균과 분산만으로 그 특성을 모두 설명할 수 있으므로 편리하다. 평균을 중심으로 한 좌우 대칭이며, 분산의 크기에 따라 크면 넓게, 작으면 좁게 분포한다.

*표준정규분포

분포들의 비교를 용이하게 위해 정규분포의 기준이 되는 평균=0, 표준편차=1로 재구성하는 표준화 과정을 거친 것을 표준정규분포라 한다.

4. 정규분포와 표준정규분포는 분포상으로 모두 정규분포를 이루기 때문에 분포의 모양이 동일하다. 그러나 실제 조사에서는 표준정규분포로 계산한다. 표준정규분포를 이용하는 이유와 그 특징을 설명하라.

표준정규분포는 평균=0, 분산=1로 표준화시킨 분포를 의미한다. 그런데, 좌우 대칭인 정규분포 그래프가 일반적으로 많이 사용되지만, 왜 굳이 표준화 과정을 필요로 하는지 알아야 한다. 단순히 하나만을 대상으로 통계량을 얻는다면 굳이 표준화를 하지 않아도 상관없다. 하지만 서로 비교해야 하는 대상이 있는 경우는 비교할 수 있는 기준이 필요하게 된다.

예를 들어 수학 60점과 영어 80점의 성적을 서로 비교했을 때 80점이 더 좋은 성적이라 할 수 있을까? 서로 다른 대상의 비교에서 기준을 0으로 설정해 주면 +와 − 어느 쪽에 있는지를 바로 확인 가능해진다. 표준정규분포는 평균=0, 표준편차=1의 분포를 가지는 것으로 정규분포 확률변수 $N(\mu, \sigma^2)$를 $z = \frac{X-\mu}{\sigma}$를 통해 표준정규분포 $N(0,1)$로 표준화 시킨 것을 말한다. 정규분포는 평균과 분산으로 모양이 결정 되므로 서로 다른 정규분포끼리 기준을 맞추어 비교를 용이하게 만드는 것이다.

5. 확률변수 X의 평균은 100이고, 분산이 100이다. 이때 확률변수 X가 107보다 클 확률과 97보다 작을 확률을 구하라.

$P(x \geq 107)$

$z = \frac{x-\mu}{\sigma}$ $\quad z = \frac{107-100}{10}$, $z = 0.7$ $\quad \therefore P(z \geq 0.7)$

$z = 0.7$에서의 확률은 0.758036으로 확인되었다. 커야 하는 확률을 구해야 하는 것이므로 $P(z < 0.7)$의 확률을 뺀 나머지가 답이 된다.

$\therefore 1 - 0.758036 = 0.241964$ 확률은 24%이다.

$P(x \leq 97)$

$z = \frac{x-\mu}{\sigma} \Rightarrow z = \frac{97-100}{10}$, $z = -0.3$ $\quad \therefore P(z < -0.3)$

$z = 0.3$에서의 확률은 0.617911로 확인되었다. 작아야 하는 확률을 구해야 하는 것이므로 $P(z < -0.3)$의 확률을 뺀 나머지가 답이 된다.

$\therefore 1 - 0.617911 = 0.3821$ 확률은 38%가 된다.

6. 우리나라는 전 세계에서 5G 통신망이 가장 발달한 나라다. 5G가 보급되면서 통신사마다 고객 이탈을 막기 위해 다양한 프로모션과 고객 관리 업무를 진행하고 있다. 1년 동안 고객 유치 업무를 진행해보니 평균적으로 5.4명을 만나 5G에 대해 설명하면 현재 이용하고 있는 상품을 변경했다. 이때 분산은 1.5명이었다.

(a) 4명 이내로 고객을 만나 새롭게 유치하게 될 고객은 전체의 몇 %가 되는지 구하라.

$P(x < 4)$

$z = \frac{x-\mu}{\sigma} \Rightarrow z = \frac{4-5.4}{\sqrt{1.5}} = -0.409$, $z = -0.409$ $\quad \therefore P(z < -0.409)$

$z = -0.409$에서의 확률은 0.3412698이므로 4명 이내로 고객을 만나 5G 상품으로 변경할

확률은 34%이다.

(b) 고객의 10%는 기존 상품을 오래 사용한 분들이다. 이 고객들에게 5G로 이전하도록 안내한다고 할 때 몇 명을 만나야 새 상품으로 변경하게 될지 구하라.
10%에 해당하는 z값을 구하면, 확률이 0.90인 $z=1.3$을 구할 수 있다.

$P(x>1.3)$

$z=\frac{x-\mu}{\sigma} \Rightarrow x=1.3\times\sqrt{1.5}+5.4 \quad \therefore x=6.99$

그러므로 약 7명을 만나야 새로운 상품으로 변경하는 고객을 만날 수 있다.

7. B손해사정법인에서 수임하는 보험 관련 사건은 1년간 약 150건에 이른다. 그중 약 6건 정도에서 제대로 된 보상을 받지 못하는 것으로 나타났다. 이 손해사정법인에서 이런 경우가 나타나는 확률을 이항분포와 포아송분포를 이용하여 계산한 후, 그 결과를 비교하여 나타나는 특징을 설명하라. 그리고 어떤 확률분포를 이용하는 것이 바람직한지 설명하라.

1) 이항분포

$n=150, x=6, p=0.04$를 $P(X=r)=\frac{r!}{r!(n-r)!}\times p^r\times(1-p)^{(n-r)}$에 대입하면

$\frac{150!}{6!(150-6)!}\times 0.04^6\times(1-0.04)^{(150-6)} \fallingdotseq 0.164$

약 16%의 확률로 제대로 된 보상을 받지 못했다.

2) 포아송분포

$P(X=x)=\frac{\lambda^x e^{-\lambda}}{x!}$로 함수식에 $\lambda=150, x=6$을 직접 대입해서

$\frac{150^6 e^{-150}}{6!}$을 계산하면 확률이 1.13517E-55로 거의 0에 가깝게 나타났다.

그러므로 0%에 가까운 것도 문제가 있으나 150건 중 6건이 16%로 계산된 것은 더 문제가 있다.

8. L지역에 거주하는 주민들이 병원을 찾아 진료를 받는 평균을 조사해보니 연간 3.5일이고, 분산은 9일이었다.

(a) 이 지역 주민 중에서 무작위로 표본을 선정했을 때, 연간 진료일이 5일을 넘을 확률을 구하라.

(b) 이 지역의 90%, 95%, 99% 확률로 주민이 진료를 받는 날짜의 범위를 구하라.

(a)

$P(x>5)$ 을 구하는 문제이므로

$z=\frac{5-3.5}{\sqrt{9}}=0.5$ 이므로, $z=0.5, \therefore P(z>0.5)$, z = 0.5 에서의 확률은 0.691462 으로 확인되었다.

클 확률이므로 $P(z\le 0.5)$ 의 확률을 뺀 나머지가 답이 된다.

$\therefore 1-0.691462=0.308538$ 확률은 31% 가 된다.

(b)

1) 90%의 진료 z=1.3 일 때, 0.9032 약 90%의 확률 이므로

$1.3=\frac{x-3.5}{\sqrt{9}}, \rightarrow x=1.3\cdot\sqrt{9}+3.5 \quad \therefore x=6.8$

2) 95%의 진료 z=1.6 일 때, 0.945201 약 95%의 확률 이므로

$1.6=\frac{x-3.5}{\sqrt{9}}, \rightarrow x=1.6\cdot\sqrt{9}+3.5 \quad \therefore x=8.3$

3) 99%의 진료 z=2.4 일 때, 0.991802 약 99%의 확률 이므로

$2.4=\frac{x-3.5}{\sqrt{9}}, \rightarrow x=2.4\cdot\sqrt{9}+3.5 \quad \therefore x=10.7$

9. 축구 경기를 관람하려고 하는데, 지방에서 열리는 경기라 돌아오는 차편의 터미널 막차 시간을 확인해보니 21시 30분이다. 터미널에서 경기장까지 15분이 걸리고 경기는 오후 7시에 시작한다. 최근의 100경기의 경기 시간은 전후반 45분씩 90분과 휴식 시간 15분, 예상 지연 시간이 전후반 각각 7분씩 평균 119분으로 계산되었으며, 분산은 100분이었다. 이 경기를 보러 갔을 때 막차를 놓치게 될 확률을 구하라.

경기시간 평균=119분

경기시작 시간은 7시, 경기 종료시간은 8시 59분

경기장~터미널 소요시간=15분

따라서 9시 14분에 터미널에 도착가능

막차시간=9시 30분 이므로 16분 지연되면 막차를 놓치게 된다.

따라서, 경기의 평균시간 119분에서 16분이 초과되면 막차를 놓치게 된다.

$P(x>235)$

$z=\frac{235-119}{\sqrt{100}}=1.6$이므로, z=1.6 $\therefore P(z>1.6)$, z = 1.6 에서의 확률은 0.9452로 16분 보다 지체될 확률은 1- 0.9452= 0.0548의 확률이므로 5.48%의 확률로 막차를 놓치게 된다.

Chapter 06 연습문제

1. 추정의 개념과 모수를 추정하는 방법에 대해 설명하고, 추정치와 추정량을 구분하여 설명하라.

추론통계에서는 표본을 통해 모집단의 성격을 파악하는 것에 중점을 둔기 때문에 모수에 대한 정보가 전혀 없는 상황에서 표본을 추출해서 모집단의 특성을 유추하여 파악하게 되는데, 이때 모수를 특정 수치나 수치의 범위로 표현하게 된다. 모수를 추정(estimation)하는 방법은 점추정(point estimation)과 구간추정(interval estimation)이 있다. 점추정은 모수가 특정한 수치로 표현되는 것이고, 구간추정은 모수를 최소값과 최대값의 범위로 추정하는 것이다. '통학시간'에 대해 점추정은 30분, 40분의 특정한 수치로 표현되고, 구간추정은 30분~40분과 같이 범위로 표현된다. 추정치(estimate)는 모수를 추정하기 위해 선택된 표본을 대상으로 통계량을 계산하며, 구체적으로 도출된 통계량을 추정치라 한다. 추정량(estimator)이란 표본을 추정치의 도출함수에 직접 대입하여 구하게 되는데, 이 과정에서 표본에서 관찰된 값으로 추정치를 계산하기 위한 함수를 추정량이라 한다.

2. 바람직한 점추정량이 되기 위한 조건을 나열하고, 이에 대해 설명하라.

점추정량은 반드시 오차를 동반할 수밖에 없으므로 오차를 최소로 만들기 위한 몇 가지 조건이 필요하다. 일치성, 불편성, 유효성, 평균제곱오차, 충분성이 바람직한 점추정량이 되기 위한 조건이다.

❶ 일치성(consistency) : 표본의 크기가 모집단 규모에 근접해야 한다.

일치성은 표본이 모집단의 규모에 근접할수록 오차가 작아진다는 의미다. 표본의 개수가 $n \to \infty$로 되어 모집단과 일치하면 오차는 0이 된다. 즉, 표본이 커질수록 위험이 감소한다.

❷ 불편성(unbiased estimator) : 추정량이 모수와 같아야 한다.

추정량 $\hat{\theta}$로 모수 θ를 추정하여 $E(\hat{\theta}) = \theta$가 되면 가장 바람직한 추정이다. 이때의 추정량을 불편추정량이라 한다. $E(\hat{\theta}) \neq \theta$가 되면 추정량과 모수에 차이가 있다는 의미이며, 이때 편의 (biased)가 있다고 한다. 추정량에 대한 기대값이 모수와 동일하게 나타나면, 이는 추출된 표본에 오류가 나타날 만한 영향이 없다는 뜻이다. 즉, 표본추출이 불편성(unbiasedness)을 만족한다는 의미다.

❸ 유효성(efficiency) : 추정량의 분산이 최소값이어야 한다.

유효성은 모수에 대한 추정량의 분산(분포)이 작을수록 추정량이 바람직하다는 의미다. 이러한 조건은 추정량이 여러 개일 경우, 이들을 서로 비교하여 가장 유효한 추정량을 확인할 때 필요하다. 모수 θ에 대한 불편추정량이 $\hat{\theta_1}$과 $\hat{\theta_2}$로 두 개가 있다고 할 때, $Var(\hat{\theta_1}) > Var(\hat{\theta_2})$라면 $\hat{\theta_1}$보다 $\hat{\theta_2}$가 더 바람직한 추정량이라고 볼 수 있다.

❹ 평균제곱오차(Mean Squared Error, MSE) : 평균제곱오차가 최소값이어야 한다.

각각의 측정치에서 평균을 뺀 나머지를 오차라 하는데, 이 차이가 최소가 되어야 한다. 모든

편차의 합은 0이 되므로 오차에 제곱을 하여 더한 평균제곱오차 $E[(\hat{\theta}-\theta)^2]$의 값이 최소가 되는 추정량이어야 한다. 즉, 추정량과 모수의 차이가 최소가 되어야 한다.

❺ 충분성(sufficiency) : 표본이 모집단의 대표성을 가져야 한다.

표본 $x_1, x_2, x_3, \cdots, x_n$으로부터 추정량 $\hat{\theta}$을 추정할 때, 확률함수를 $T(x_1, x_2, x_3, \cdots, x_n)=x_1$이라 하면 $\hat{\theta}=x_1$이 된다. 즉, 확률함수에 어떤 값을 대입해도 x_1만 도출되기 때문에, 추정량 $\hat{\theta}$가 표본 $(x_1, x_2, x_3, \cdots, x_n)$의 정보를 모두 포함한다고 보기 어렵다. 표본은 모집단에 대해 대표성을 가져야 통계적인 의미가 있으므로, $\hat{\theta}=T(x_1, x_2, x_3, \cdots, x_n)$의 정보가 모수 θ에 대한 모든 정보를 포함할 때, 추정량 $\hat{\theta}$을 모수 θ에 대한 충분한 추정이라 하고, 이를 충분성이 확보되었다고 한다.

❶ 일치성, ❷ 불편성, ❸ 유효성, ❹ 평균오차제곱은 바람직한 추정량에 한정되는 조건이지만 ❺ 충분성은 통계학 전체에 적용될 수 있는 일반적인 통계량에 관한 조건이다.

3. 구간추정의 개념과 통계에서 구간추정이 필요한 이유에 대해 설명하라.

점추정을 할 때 표본의 수 n→∞로 하여 모집단을 조사하지 않는 한 오차가 발생할 수밖에 없다. 또한 점추정치를 계산해 내더라도 오차에 대한 어떠한 정보도 확인할 수 없다. 점추정은 명확한 수치를 제시하면서 신뢰도의 문제를 동반하기 때문에 이를 보완하기 위하여 신뢰도를 제시하면서 상한값과 하한값으로 추정하며, 이를 구간추정(interval estimation)이라 한다. 구간추정이 필요한 이유는 100%의 신뢰도가 가장 중요하다고 간주한 상태에서 구간추정을 하게 된다면 무조건 맞아야 하므로 범위가 많이 넓어지게 된다. 하지만, 현실적으로 그렇게 할 수 없으므로 신뢰할 수 있을 만한 정도인 90%, 95%, 99%, 99.9% 수준에서 조사의 목적에 맞는 신뢰도를 선택하여 계산하게 된다.

4. 통계학 중간고사를 실시한 결과 평균이 81점, 표준오차가 9점으로 확인되었다. 신뢰수준 90%, 95%, 99%으로 구간추정을 하여 모평균을 추정하라.

90% 에서의 평균의 구간추정 : $\mu=$ 평균, $\bar{x}=81, SE=9, z=1.64$이므로
$81-1.64\times9 \le \mu \le 81+1.64\times9$
$66.24 \le \mu \le 95.76$

95% 에서의 평균의 구간추정 : $\mu=$모평균, $\bar{x}=81, SE=9, z=1.96$이므로
$81-1.96\times9 \le \mu \le 81+1.96\times9$
$63.36 \le \mu \le 98.64$

99% 에서의 평균의 구간추정 : $\mu=$모평균, $\bar{x}=81, SE=9, z=2.58$이므로
$81-2.58\times9 \le \mu \le 81+2.58\times9$
$57.78 \le \mu \le 1.4.22$

5. 신뢰수준과 신뢰구간의 개념을 설명하라.

신뢰수준(confidence level)은 모수가 추정값이 존재할 수 있는 범위를 어느 정도 확신할 수 있을 것인가에 대한 확률이다. 95%, 99%, 99.9%의 신뢰수준은 $100(1-\alpha)\%$로 계산되

며, 여기에서의 는 조사에서 인정되는 오차수준이다. 신뢰구간(confidence interval)은 하한값과 상한값의 구간으로 표시되며, 신뢰수준을 기준으로 추정된 점으로부터 −방향과 +방향으로 하한과 상한을 표시한다. 신뢰구간이 좁을수록 조사자는 더 의미 있는 결과를 제시할 수 있는 것처럼 보이지만, 신뢰수준이 낮아지는 것을 감수해야 한다. 따라서 조사대상이나 조사상황에 맞는 신뢰수준을 선택할 수 있어야 한다.

6. 통계조사에서 100%의 신뢰구간을 사용하지 않는 이유는 무엇인지 설명하라.

신뢰도를 100%로 하면, 조사결과가 100% 맞는 것이므로 최상의 결과가 될 것 같다. 신뢰도 90%, 95%, 99%에서 모평균의 구간추정을 한 결과에서 보는바와 같이 신뢰도가 올라갈수록 구간이 점점 넓어지는 것을 알 수 있다. 정값이 올라가기 때문에 그런 결과가 발생하는데, 100%에 해당하는 $z=$ 이다. 이는 $-\infty \le \mu \le \infty$의 값을 갖는다는 의미이며, 틀릴 확률은 당연히 0%가 된다. 하지만, 전혀 의미가 없는 결과가 되기 때문이다.

7. [컴퓨터 실습] 연습문제.xlsx의 '6장_손흥민 골기록' 시트는 손흥민이 프로 구단에 데뷔한 첫 해부터 2019/2020 시즌까지의 경기수와 골 기록을 보여준다. 풀이 과정을 적고, Excel을 이용하여 2020/2021 시즌에 몇 골을 기록할 수 있을지 90%, 95%, 99%의 신뢰구간에서 예측하라. 단, 90% $z=1.64$, 95%의 $z=1.96$, 99%의 $z=2.58$의 z분포를 이용한다. (예측을 해본 후 2020/2021의 실제 골 기록과 비교해보기 바란다.)

연수	시즌	경기수	골	팀
1	10/11	13	3	함부르크
2	11/12	30	5	
3	12/13	34	12	
4	13/14	43	12	레버쿠젠
5	14/15	42	17	
6	15/16	42	8	토트넘
7	16/17	47	21	
8	17/18	53	18	
9	18/19	47	20	
10	19/20	30	11	

● 직접 풀이

$\bar{x}=12.7$, 90%의 $z=1.64$, 95%의 $z=1.96$, 99%의 $z=2.58$, $\sigma=5.9$, $n=10$이므로

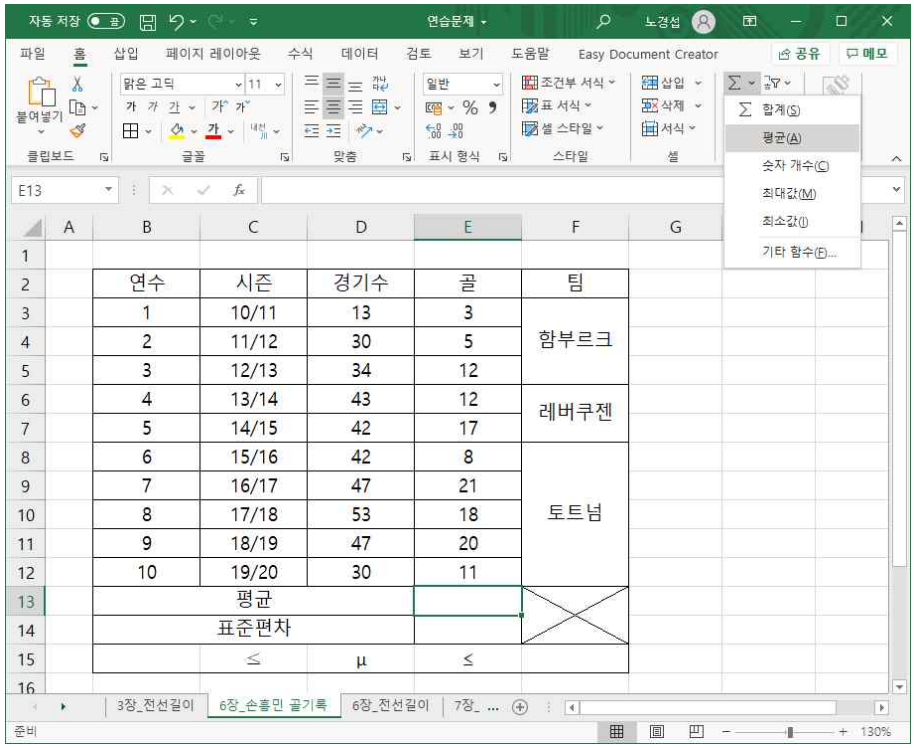
$\bar{x}-z_{\frac{\alpha}{2}}\times\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+z_{\frac{\alpha}{2}}\times\frac{\sigma}{\sqrt{n}}$에 대입하면

$$90\% : 12.7-1.64\times\frac{5.9}{\sqrt{10}} \le \mu \le 12.7+1.64\times\frac{5.9}{\sqrt{10}} = 9.64 \le \mu \le 15.76$$

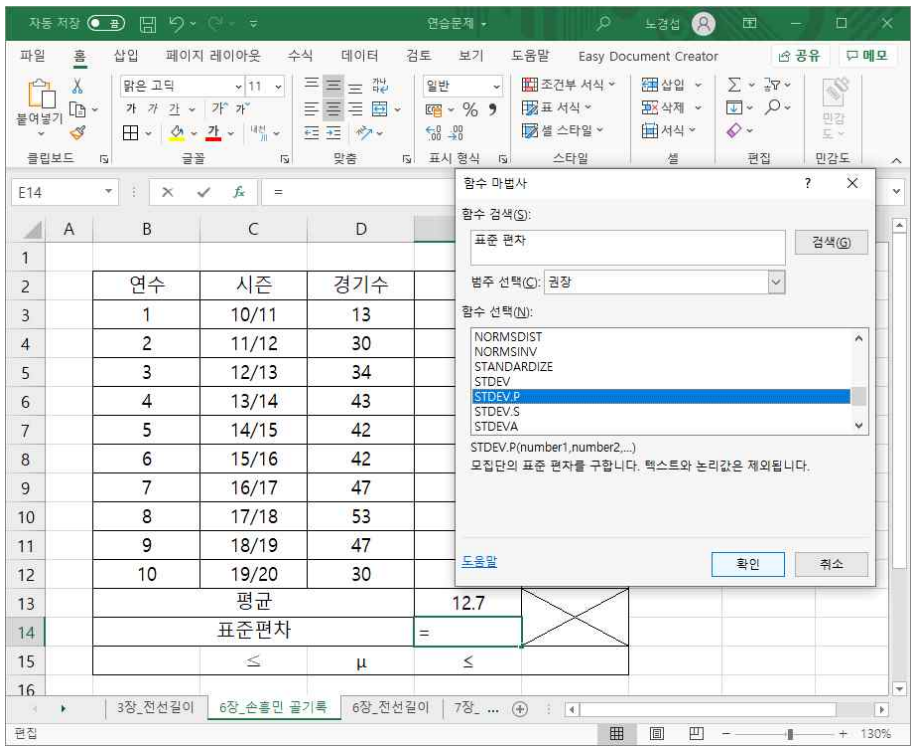
$$95\% : 12.7-1.96\times\frac{5.9}{\sqrt{10}} \le \mu \le 12.7+1.96\times\frac{5.9}{\sqrt{10}} = 9.00 \le \mu \le 16.36$$

$$99\% : 12.7-2.58\times\frac{5.9}{\sqrt{10}} \le \mu \le 12.7+2.58\times\frac{5.9}{\sqrt{10}} = 7.89 \le \mu \le 17.51$$

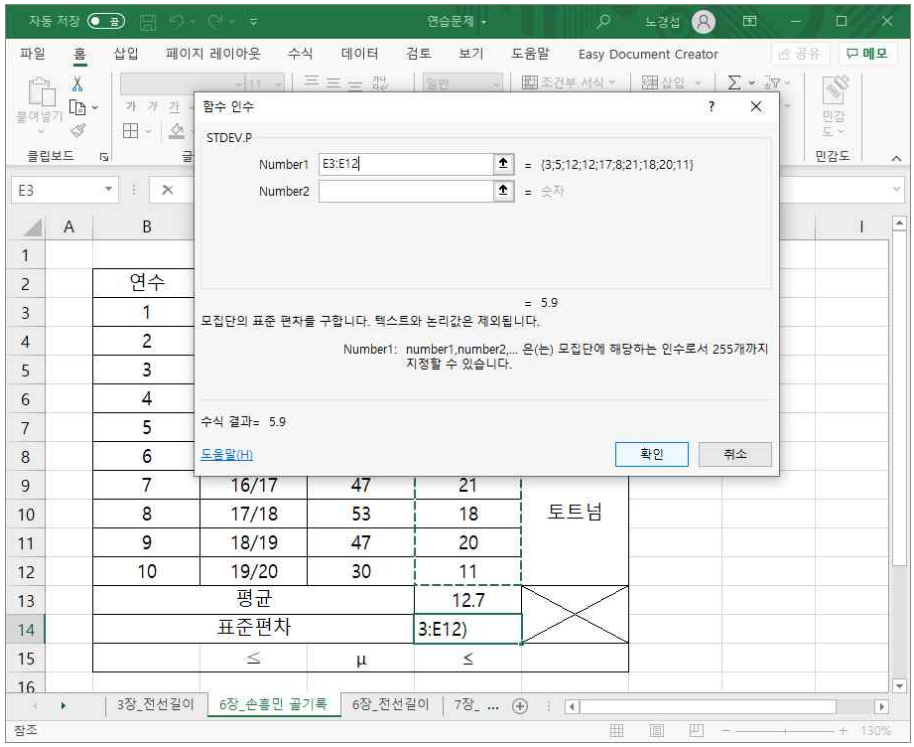
● Excel 풀이



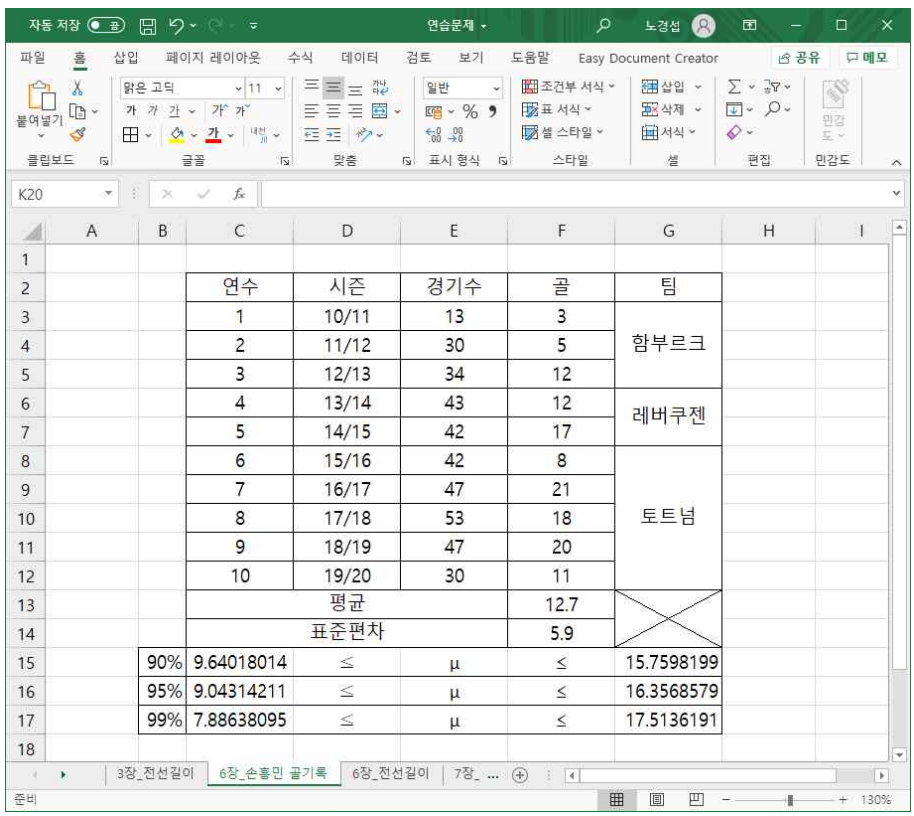
시그마 드롭다운 버튼을 클릭해서 '평균(A)'를 클릭한다.



함수 마법사를 통해 표준편차를 클릭한다.



표준편차를 구하기 위한 인수 E3:E12를 입력한다.



연수	시즌	경기수	골	팀
1	10/11	13	3	함부르크
2	11/12	30	5	
3	12/13	34	12	
4	13/14	43	12	레버쿠젠
5	14/15	42	17	
6	15/16	42	8	토트넘
7	16/17	47	21	
8	17/18	53	18	
9	18/19	47	20	
10	19/20	30	11	
평균			12.7	
표준편차			5.9	

90%	9.64018014	≤	μ	≤	15.7598199
95%	9.04314211	≤	μ	≤	16.3568579
99%	7.88638095	≤	μ	≤	17.5136191

90%에서의 z=1.64, 95%에서의 z=1.96, 99%에서의 z=2.58을 이용하여 공식에 대입하면 다음과 같다.

90%에서의 하한은 '=F13-(1.64*(F14/SQRT(10)))'
상한은 '=F13+(1.64*(F14/SQRT(10)))'
95%에서의 하한은 '=F13-(1.96*(F14/SQRT(10)))'
상한은 '=F13+(1.96*(F14/SQRT(10)))'
99%에서의 하한은 '=F13-(2.58*(F14/SQRT(10)))'
상한은 '=F13+(2.58*(F14/SQRT(10)))'

위의 그림과 같이 확인할 수 있다.

8. 표준편차와 표준오차의 개념을 설명하고, 두 개념의 차이를 설명하라.

표준편차는 표본평균으로부터 표본들의 흩어져있는 산포를 나타내기 위하여 분산을 먼저 구한 후 다시 제곱근을 취한 값이다. 모평균을 알 수 없으므로 표본평균으로 모평균을 추정했으며, 표본의 분포를 확인하고자 표준편차를 구했다. 즉 모평균을 추정하기 위해 표본을 추출해서 표본의 평균과 표본의 특성을 나타내는 것이 표준편차이다.

표준편차의 공식은 $s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$ 이다.

표준오차는 모평균을 추정하는 표본평균의 산포도를 나타내는데, 표본평균의 산포라는 말에서 표본이 여럿이라는 것을 알 수 있다. 2개 이상의 표본평균으로 모수를 추정한다면 더 정확한 추정이 될 수 있다. 표본의 추출 횟수를 최대한 늘려서 n→∞로 할 수 있다면 이들의 평균은 모수와 일치하게 될 것이다. 표준오차는 추출된 표본들의 숫자를 늘려서 평균을 구한 후 이들 간의 표준편차를 나타낸 것이며, (표본평균의) 표준오차라 한다. 표준편차를 표준오차와 비교할 때의 가장 큰 장점은 표준오차가 줄어들수록 평균을 나타내는 점들이 집중적으로 모여 있는 것이므로 모수의 추정이 정확하게 이루어졌음을 판단할 수 있다.

평균의 표준오차를 구하는 공식은 모분산을 알고 있다면 $S.E=\frac{\sigma}{\sqrt{n}}$이 되고,

모분산을 모르고 있다면 표본에서 표준편차를 구한 $S.E=\frac{s}{\sqrt{n}}$이다.

9. [컴퓨터 실습] 명절을 맞이하여 전국에서 사과 깎기의 달인이 모였다. 대결 과제는 '사과 껍질이 끊어지지 않게 최대한 길게 깎기'다. 무작위로 25명의 출전자들의 깎은 껍질의 길이를 측정하여 연습문제.xlsx의 '6장_사과껍질' 시트에 정리하였다. 이 데이터를 이용하여 사과 껍질 길이의 90%, 95%, 99%에 해당하는 신뢰구간을 구하라.

구분	t	하한	상한
90%	1.711	86.60644	93.31356
95%	2.064	85.91456	94.00544
99%	2.79694	84.4779976	95.4420024

● 직접 풀이

$\bar{x}=89.96,\ s=9.8,\ n=25,$ 표준오차는 $\frac{9.8}{\sqrt{25}}=1.96$

$\bar{x}-t_{\frac{\alpha}{2}}\times\frac{s}{\sqrt{n}}\le\mu\le\bar{x}+t_{\frac{\alpha}{2}}\times\frac{s}{\sqrt{n}}$를 이용하면

90% : 자유도 24인 $t_{\frac{\alpha}{2}}=1.711$

$$89.96-1.711\times1.96\le\mu\le89.96+1.711\times1.96$$
$$86.66.644\le\mu\le93.31356$$

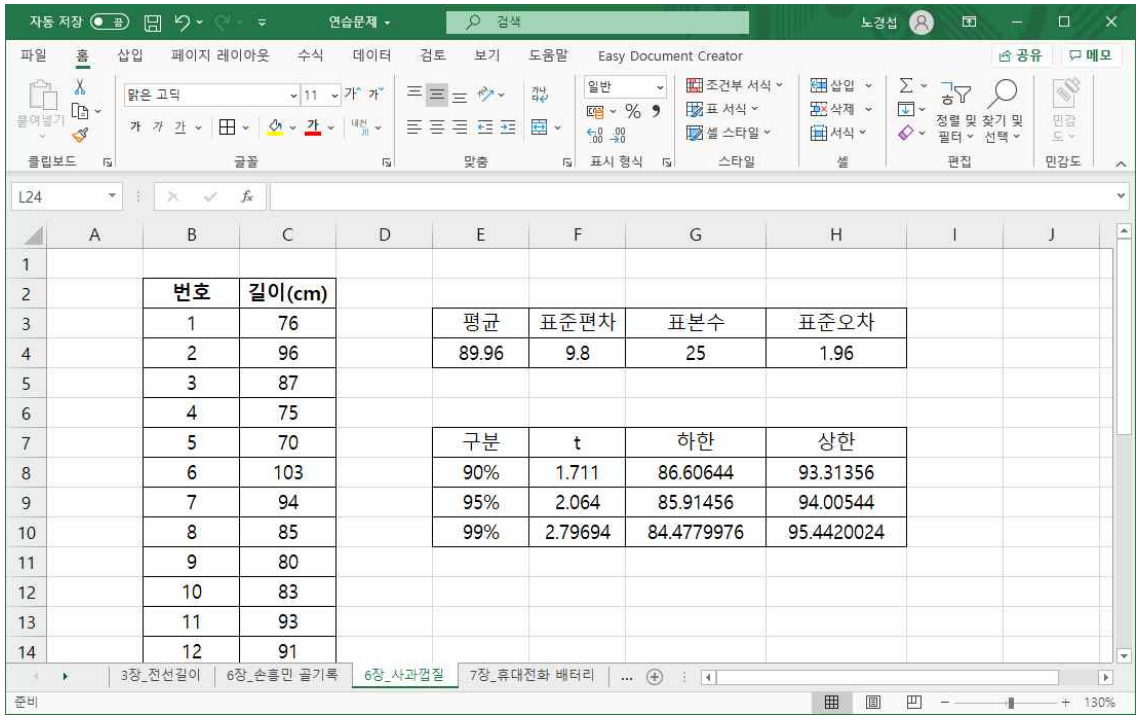
95% : 자유도 24인 $t_{\frac{\alpha}{2}}=2.064$

$$89.96-2.064\times1.96\le\mu\le89.96+2.064\times1.96$$
$$85.91456\le\mu\le94.00544$$

99% : 자유도 24인 $t_{\frac{\alpha}{2}}=2.79694$

$$89.96-2.79694\times1.96\le\mu\le89.96+2.79694\times1.96$$
$$84.4779976\le\mu\le95.4420024$$

● Excel 풀이



번호	길이(cm)
1	76
2	96
3	87
4	75
5	70
6	103
7	94
8	85
9	80
10	83
11	93
12	91

평균	표준편차	표본수	표준오차
89.96	9.8	25	1.96

구분	t	하한	상한
90%	1.711	86.60644	93.31356
95%	2.064	85.91456	94.00544
99%	2.79694	84.4779976	95.4420024

10. 문구점에서 판매되는 스카치테이프의 길이를 조사하고자 한다.

(a) 최소 몇 개의 표본으로 계산해야 하는지 구하라.

(신뢰구준 95%, 허용오차 ±30cm, 표준편차 22cm)

$1-\alpha=0.95$이므로 $z_{\frac{\alpha}{2}}=1.96$, 오차한계=30, $s=20$

$n=(\frac{z_{\frac{\alpha}{2}}\times\sigma}{d})$에 대입해 보면, $n=(\frac{z_{\frac{\alpha}{2}}\times\sigma}{d})^2=(\frac{1.96\times20}{30})^2=1.71$

그러므로, 최소한 2개의 표본을 구하여 용량을 확인해야 한다.

(b) 알려진 표준편차가 없어서 10번의 표본을 추출해서 (표준편차) 16cm를 계산했다. 신뢰수준 99%, 허용오차 ±30cm일 때, 몇 개의 표본으로 조사해야 하는지 구하라.

$1-\alpha=0.95$이므로 $z_{\frac{\alpha}{2}}=1.96$, 오차한계=30, $s=16$

$n=(\frac{z_{\frac{\alpha}{2}}\times s}{d})$에 대입해 보면, $n=(\frac{z_{\frac{\alpha}{2}}\times s}{d})^2=(\frac{1.96\times16}{30})^2=19.2370$

그러므로, 최소한 20개의 표본을 구하여 용량을 확인해야 한다.

11. 달걀을 고를 때 달걀 껍데기에 각인된 기호의 의미를 알면 좋다. 첫 번째 숫자 그룹은 계란 산란일, 가운데 기호는 축산업에 따른 사업자의 고유번호, 마지막 숫자는 사육 환경을 의미한다. 이 마지막 숫자 중 1은 방사, 2는 축사 내 평사, 3은 개선된 케이지, 4는 기존 케이지를 의미한다. 당연히 1이 적힌 달걀이 품질이 가장 좋고 가격도 비싸다. 그리고 3 이상의 숫자가 찍힌 달걀은 시중의 일반적인 달걀보다 약 25% 이상 비싸다.(일반적인 달걀은 대부분 4에 해당한다.) 일반 소비자에게 이런 정보를 제공하고 100명의 표본을 조사하여 3 이상의 달걀을 구입할 것인지에 대한 조사를 진행했다. 그 결과 20명의 선택이 달라졌을 때, 달걀을 구입하는 전체 소비자의 90% 신뢰구간으로 3 이상의 건강한 달걀을 선택할 비율의 신뢰구간을 구하라.

100명의 소비자 중 20명의 선택이 달라졌으므로, $\hat{p}=\frac{20}{100}=0.2$

$\hat{p}\pm z_{\frac{\alpha}{2}}\times\sqrt{\frac{\hat{p}\times(1-\hat{p})}{n}}$ 을 이용해야 하므로, 나머지 필요한 수치는 $n=100$, $z_{\frac{\alpha}{2}}=1.64$

대입해보면

$0.2-1.64\times\sqrt{\frac{0.2\times0.8}{100}}\le p\le0.2+1.64\times\sqrt{\frac{0.2\times0.8}{100}}$

$0.1344\le p\le0.2656$

12. 학교를 다니며 강의를 들었을 때와 Zoom으로 강의를 들었을 때를 비교하여 어떤 경우에 성적이 더 향상되었는지를 조사하고자 한다. 수강생의 비율에 대한 95%의 신뢰구간을 알려고 할 때, 오차한계가 10% 이하가 되는 표본의 크기를 구하라. (단, $\hat{p}=0.95$)

$z_{\frac{\alpha}{2}}=1.96,\ \hat{p}=0.95,\ d=0.1$

$n=0.95\times0.05\times(\frac{1.96}{0.1})^2 \fallingdotseq 18.25$

최소 19명의 표본설정이 필요하다.

13. 정규분포를 따르는 모집단의 분산에 대해 구간추정을 할 때, 상한과 하한을 다르게 표현하는 이유를 설명하라.

정규분포의 경우 신뢰구간의 상한과 하한이 중심으로부터 ±를 이용하여 동일한 위치로 계산 된다. 그러나, χ^2분포는 분산 자체의 산포도를 나타내는 분포이며, 분산이 제곱값을 가지므로 항상 양수이다. 그러므로 좌우 대칭인 정규분포가 아니므로 서로 다르게 표현된다.

14. 대학생 120명을 대상으로 한 달 평균 용돈을 조사하니 28만 3천 원이었다. 표준편차를 17만 원이라고 가정하고 신뢰수준 95%에서의 분산을 추정하라.

$n=120,\ s^2=17,\ \chi^2_{1-\frac{\alpha}{2}}=128.42,\ \chi^2_{\frac{\alpha}{2}}=73.36$

$$\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}$$

$15.753 \le \sigma^2 \le 27.576\cdots$

Chapter 07 연습문제

1. 가설의 개념과 종류에 대해 설명하라.

가설(hypothesis)이란 주어진 사실 혹은 조사하고자 하는 사실이 어떠하다는 주장이나 예측을 말한다. 즉 조사자가 통계학적으로 모수를 추정할 때, 모수는 어떠하다(혹은 어떠할 것이다)는 조사자의 주장을 말한다. 가설에는 귀무가설과 대립가설이 있으며, 귀무가설(歸無假說, null hypothesis)는 영가설(零假說)이라고도 한다. '귀무(歸無), 영(零), nul'이라는 용어가 아무것도 없다는 것을 의미하듯이, 귀무가설은 일반적으로 믿어지는 사실이라서 조사의 의미가 없다는 것으로 해석하면 된다. 즉 연구를 해서 어떠한 의미도 찾아내지 못한다는 의미이다. 조사보고서나 연구논문에서는 가설(hypothesis)의 H와 영(零, 0)의 의미로 H_0으로 표기한다. 대립가설(對立假說, antihypothesis)는 연구가설(硏究假說, research hypothesis)이라고도 한다. '대립(對立), 연구(硏究), anti-'이라는 용어에서 보듯이 '귀무' 즉, 공공연하게 사실로 받아들여지는 현상에 대해 '대립'이라는 가설을 설정하고 연구한다는 의미이다. 조사보고서나 연구 논문에서는 가설(hypothesis)의 H와 영(零, 0)과 반대가 된다는 의미로 H_1으로 표기한다.

2. 검정이 무엇인지 설명하고, 그 종류를 유의수준과 함께 설명하라.

귀무가설과 대립가설을 세웠다면 어떤 가설이 맞는가를 판단해야 한다. 우선 어떤 가설을 채택하기 이전에 가설을 기각할 수 있는 검정방법을 결정해야 한다. 가설검정의 출발점은 귀무가설이 되며, 귀무가설의 채택여부를 판단해서 귀무가설을 기각할 때 대립가설을 받아들이는 논리가 형성된다. 기각역의 판단기준은 양측검정과 단측검정으로 구분된다. 기각이 되는 기준은 조사결과가 α에 포함되면 기각하고 포함되지 않으면 채택을 하며, α를 유의수준이라 한다. 양측검정이란 조사하고자 하는 대립가설 즉, '사실과 다르다'라는 것을 검정하여, 귀무가설을 기각하고 대립가설을 채택하고자 하는 것이다. 그래프로 나타내면 귀무가설이 기각되는 영역이 +방향과 -방향에 양쪽에 있으므로 유의 수준 α는 각각 $\frac{\alpha}{2}$씩으로 구분되어 나타난다. 단측검정이란 대립가설을 조사목적에 따라 기각되어야 하는 영역이 적다(작다)는 것으로 수립할 수도 있고, 혹은 많다(크다)는 것으로도 수립할 수도 있다.

3. 통계적 판단을 하는 데 있어 유의수준이 어떠한 역할을 하는지 설명하라.

통계적 판단을 한다는 것은 모수를 추정한다는 의미이고, 추정은 틀릴 가능성이 있기 때문에 모수를 추정할 때는 항상 오류가능성(확률)을 제시한다. 통계적 의사결정에서 틀리지 않는 것이 목적이라면 추정된 모수의 틀릴 확률을 낮게 잡으면 되고, 의사결정의 오류에 상관없이 특정한 수치를 도출하는 것이 목적이라면 표본으로부터 계산한 계산결과를 정확하게 제시하면 된다. 통계학에서는 모수의 추정이 맞을 확률을 $1-\alpha$로 표시하며, α를 '유의수준(significance level)'이라 하며, 확률(probabilty)로 표시되므로 약자를 사용하여 p값(p-value)로 표시한다.

4. 통계적 오류에 대해 설명하고, 이러한 오류를 피하기 위해서는 어떻게 해야 하는지 설명하라.

통계조사는 모집단을 대상으로 하지 않기 때문에 필연적으로 틀릴 수 있다는 한계를 내재하고 있다고 했다. 이러한 한계를 통계적 오류라 하는데, 이는 아무리 유의수준으로 가설을 검정했다고 해서 그 결과가 반드시 맞는 것을 의미하지는 않는다. 이처럼 통계학을 이용해서 모수를 추정하더라도 실제와는 다른 결론에 도달하는 것을 오류라 한다. 통계학에서는 이러한 통계적 오류는 1종 오류(type Ⅰ eror)와 2종 오류(type Ⅱ eror)로 구분한다.

1종 오류(type Ⅰ eror) : 귀무가설을 채택해야 함에도 귀무가설을 기각하는 경우
2종 오류(type Ⅱ eror) : 귀무가설을 기각해야 함에도 귀무가설을 채택하는 경우

귀무가설을 채택해야 하지만 귀무가설을 기각하는 경우의 확률은 로 표시하고 유의수준이라 하지만, 귀무가설을 기각해야 함에도 귀무가설을 채택하는 경우의 확률은 β로 표시한다. 그러므로 기각해야 할 귀무가설을 기각하는 확률은 $1-\beta$가 되고, 이는 2종 오류가 발생하지 않을 확률이며 가설의 검정력(power of hypothesis testing)이라 한다. 1종 오류를 피하려면 귀무가설을 채택하기 위한 채택역을 넓히면 되고 기각역을 줄이면 된다. 즉, α의 값을 줄이면 1종 오류를 피할 수 있다. 그러나 이런 경우 기각해야 할 귀무가설을 채택하는 2종 오류에서 자유로울 수 없다. 때문에 연구자는 1종 오류와 2종 오류를 피할 수 있는 적절한 유의수준을 선택해야만 한다.

5. 전기 자동차에 들어가는 배터리 기술이 발달하면서 점점 운행 거리가 늘어나고 있다. 제조사에서 주장하는 전기 자동차의 운행 거리가 완충 시 600km라는 광고를 확인하기 위해 직접 조사를 하고자 한다. 표본 121개를 선택하여 직접 운행 거리를 측정했다. 그 결과 평균 운행거리는 599.2km였고, 표준편차는 25km였다. 이때 가설을 수립하고, 유의수준 0.05에서 양측검정과 좌측검정을 실시하라.

● 양측검정

H_0 : 전기자동차의 운행거리는 $600km$이다.
H_1 : 전기자동차의 운행거리는 $600km$가 아니다.

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$이며, $|z| > 1.96$이면 귀무가설을 기각한다.

$z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$에 대입하면, $\dfrac{599.1-600}{25/\sqrt{121}} = -0.369$

$|z| < 1.96$이므로 '전기자동차의 운행거리는 600km이다.'라는 귀무가설(H_0)을 기각할 수 없다.

● 좌측검정

H_0 : 전기자동차의 운행거리는 $600km$ 이상이다.
H_1 : 전기자동차의 운행거리는 $600km$보다 적다.

$-z_{\alpha} = -z_{0.5} = -1.96$이며, $z < -1.96$이면 귀무가설을 기각한다.

$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ 대입하면, $\frac{599.1-600}{25/\sqrt{121}}=-0.369$

계산된 검정통계량이 -0.369로 $z>-1.96$이므로 '전기자동차의 운행거리는 600km이다.'라는 귀무가설(H_0)을 기각할 수 없다.

6. [컴퓨터 실습] 정부 발표에 의하면 국내의 1인당 스마트폰 데이터 사용량이 월간 10Gb를 넘었다고 한다. 연습문제.xlsx의 '7장_1인당 월간 데이터 사용량' 시트는 스마트폰 사용자 30명을 대상으로 월간 데이터 사용량을 측정하여 적은 것이다. 모집단의 분포는 정규분포라고 가정한다. 이때 가설을 수립하고, 유의수준 0.05에서 양측검정과 좌측검정을 실시하라.

● 양측검정

H_0 : 1인당 월간 데이터 사용량은 $10Gb$이다.
H_1 : 1인당 월간 데이터 사용량은 $10Gb$이 아니다.

통계량은 $\bar{x}=10.4$, $s \fallingdotseq 3.014618$, $\sqrt{n} \fallingdotseq 5.477226$으로 계산된다.

양측에서의 임계치 $z=\pm 1.96$

검정통계량 $z=0.660\cdots$

검정통계량이 $z=\pm 1.96$의 범위에 있지 않으므로, 귀무가설을 기각하고 대립가설을 채택한다. 즉 1인당 월간 데이터 사용량은 10Gb가 아니다. 내재적 의미는 우측 영역으로 벗어나 있으므로 10Gb보다 많다고 할 수 있다.

● 단측검정

H_0 : 1인당 월간 데이터 사용량은 $10Gb$이상이다.
H_1 : 1인당 월간 데이터 사용량은 $10Gb$미만이다.

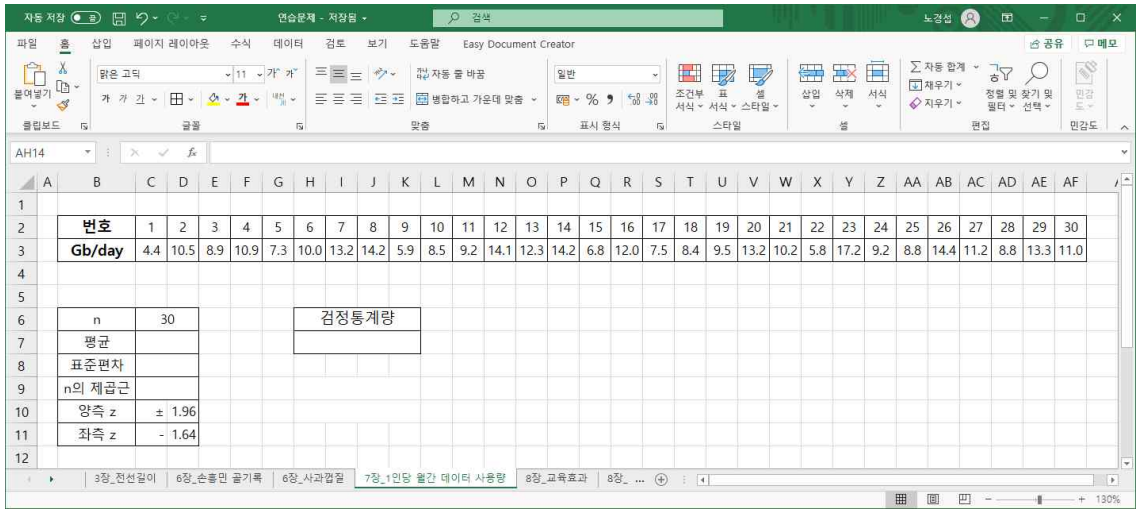
통계량은 $\bar{x}=10.4$, $s \fallingdotseq 3.014618$, $\sqrt{n} \fallingdotseq 5.477226$으로 계산된다.

좌측에서의 임계치 $z=-1.64$

검정통계량 $z=0.660\cdots$

검정통계량이 z 1.64의 범위에 있지 않으므로, 귀무가설을 기각하고 대립가설을 채택한다. 즉 1인당 데이터 사용량은 10Gb이상이다.

● Excel 풀이

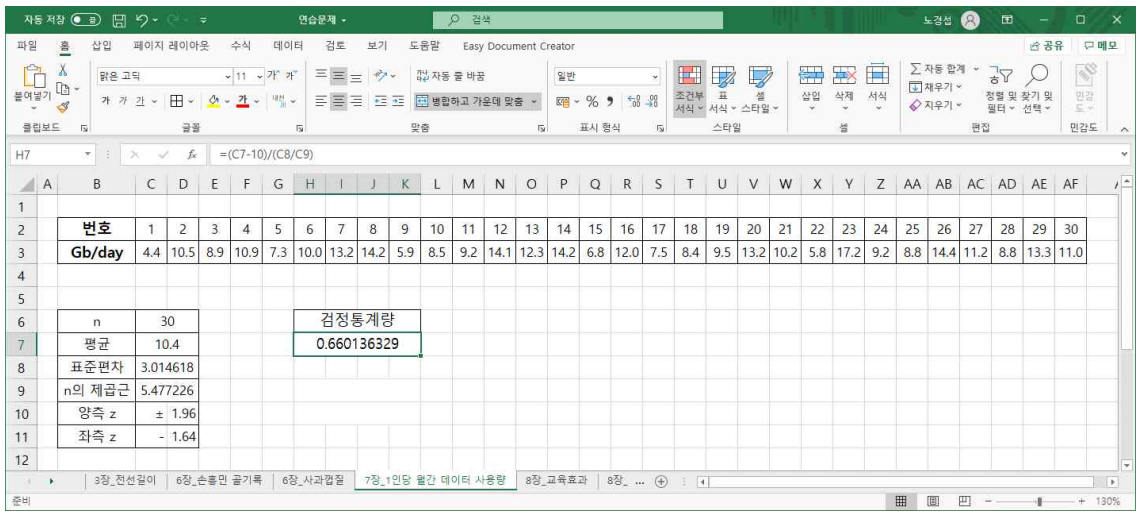


번호	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Gb/day	4.4	10.5	8.9	10.9	7.3	10.0	13.2	14.2	5.9	8.5	9.2	14.1	12.3	14.2	6.8	12.0	7.5	8.4	9.5	13.2	10.2	5.8	17.2	9.2	8.8	14.4	11.2	8.8	13.3	11.0

n	30
평균	
표준편차	
n의 제곱근	
양측 z	± 1.96
좌측 z	- 1.64

검정통계량

'7장_1인당 월간 데이터 사용량' 시트를 연다.



번호	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Gb/day	4.4	10.5	8.9	10.9	7.3	10.0	13.2	14.2	5.9	8.5	9.2	14.1	12.3	14.2	6.8	12.0	7.5	8.4	9.5	13.2	10.2	5.8	17.2	9.2	8.8	14.4	11.2	8.8	13.3	11.0

n	30
평균	10.4
표준편차	3.014618
n의 제곱근	5.477226
양측 z	± 1.96
좌측 z	- 1.64

검정통계량
0.660136329

통계량 계산의 편의를 위해 빈 칸의 '평균', '표준편차', 'n의 제곱근'을 구한다.
H7셀에 공식에 맞는 '=(C7-10)/(C8/C9)'을 입력하여 검정통계량을 구한다.

7. [컴퓨터 실습] 휴대전화 배터리의 표본 100개를 선택하여 배터리의 작동 시간을 확인했더니, 평균 작동 시간은 5.1시간이었고, 표준편차가 50분이었다. 유의수준이 0.05일 때 귀무가설의 기각 여부를 Excel을 이용하여 판단하라.

$H_0 : \mu = 5.1$, $H_1 : \neq 5.1$

$n = 100$, $\mu = 5$, $\sigma = 50$, $\alpha = 0.05$, $\bar{x} = 5.1$이므로

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{5.1 - 5}{50/\sqrt{100}} = 0.02$$

8. [컴퓨터 실습] B보험회사에서는 재무 컨설턴트가 월간 목표인 106%에 미달하면, 영업 성과의 향상을 위해 영업 전략교육을 진행하기로 하였다. 무작위로 재무 컨설턴트 150명을 표본으로 선택하여 평균 달성 목표를 확인했더니 99%라는 결과를 얻었다. 영업 전략교육을 실시해야 할지 말아야 할지에 대해 유의수준 0.05에서 Excel을 이용하여 검정하라.

$H_0 : p \leq 1.06$, $H_1 : < 1.06$

$n = 150$, $\hat{p} = 0.99$, $p = 1.06$, $\alpha = 0.05$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{150}}} = \frac{0.99 - 0.06}{\sqrt{\frac{1.06(1-1.06)}{150}}} = 0.0205$$

$\alpha = 0.05$인 좌측검정이므로, 표준정규분포표 $P(0 \leq Z \leq z)$에서 $z_{0.05} = 1.64$이다. 좌측검정이므로 $z_{0.05} = -1.64$이며, $z = 0.0205$는 좌측 기각역에 속한다.

영업성과 향상을 위한 영업전략교육을 실시하지 않아도 된다.

Chapter 08 연습문제

1. 평균 차이를 검정할 때, 표본이 1개인 경우 측정 횟수에 따른 검정 방법과 표본이 2개인 경우 검정 방법이 어떻게 다른지 설명하라.

평균차이 검정에서 표본이 1개인 경우는 측정횟수에 따라 구분되는데 1회의 측정으로 검정하는 경우는 일표본 t검정이라 하고, 사전/사후와 같이 2회 측정이 되는 경우는 대응표본 t검정이라 한다. 각기 다른 2개의 표본을 기준으로 평균을 비교하는 경우는 독립표본 t검정이라 한다.

2. 대응표본의 개념에 대해 설명하고, 주로 어떤 경우에 활용되는지 설명하라.

두 개의 모집단으로부터 표본을 추출할 때, 표본을 구성하는 각각의 인자에 대해 짝을 지어 서로 연관이 되도록 구성한 표본을 대응표본(paired sample)이라 한다. 대응표본은 주로 교육, 광고, 약품 등의 분야에서 사전(ex-ante)/사후(ex-post)를 비교하여 효과를 평가하는 검정방법에 주로 사용된다. 두 개의 모집단에서 짝을 맞추어 추출한 2개의 표본이지만, 이 두 개의 표본은 각각 짝을 맞추어 비교하고 평균차이를 비교한다는 것을 기준으로 보면 대상이 같으므로 하나의 표본에 대한 검정이라 생각할 수 있다.

3. [컴퓨터 실습] A제약회사는 암기에 도움이 되는 청명환을 개발했다. 이 제약회사는 청명환을 복용하면 암기력이 20% 증가하여 학습에 도움이 된다고 주장한다. 연습문제.xlsx의 '8장_청명환' 시트는 청명환을 복용한 사람 25명에 대해 청명환을 복용하기 전과 후의 암기능력을 조사한 자료다. 청명환의 효과가 있는지 가설을 수립하고 유의수준 5%로 검정하라.

H_0 : 명환의 암기효과가 나타나지 않았다.
H_1 : 청명환의 암기효과가 나타났다.

차이의 평균$(\bar{d}_x) = \dfrac{\sum(x_i - x_j)}{n}$

분산$(s_d^2) = \dfrac{\sum(d-\bar{d}_x)^2}{n-1}$, 표준편차$(s_d) = \sqrt{\dfrac{\sum(d-\bar{d}_x)^2}{n-1}}$, 표준오차$(\bar{s}_d) = \dfrac{s_d}{\sqrt{n}}$

각 공식에 대입해 보면

차이의 평균$= \dfrac{(84-90)+(94-95)+\cdots+(94-96)}{25} = -6.68$

차이의 분산$= \dfrac{(-6-(-6.68))^2+(-1-(-6.68))^2+\cdots+(-2-(-6.68))^2}{25-1} = 11.56$

차이의 표준편차$= \sqrt{\dfrac{(-6-(-6.68))^2+(-1-(-6.68))^2+\cdots+(-2-(-6.68))^2}{25-1}} = 3.4$

차이의 표준오차$= \dfrac{3.4}{\sqrt{25}} = 0.141666667$

양측검정을 하는 것이므로 유의수준은 0.05이며 자유도는 24인 $t_{(24,\frac{\alpha}{2})} = 2.0639$이다.

양측검정의 신뢰구간은 $\bar{d}_x - 2.0639 \times \frac{3.4}{\sqrt{25}} \le d_x \le \bar{d}_x + 2.0639 \times \frac{3.4}{\sqrt{25}}$

$$-6.972 \le d_x \le -6.388$$

95%의 신뢰구간에서의 검정통계량 $t = \frac{\bar{d}_x}{s_d/\sqrt{n}} = \frac{-6.68}{0.141666667} \fallingdotseq -47.153$

-47.153은 신뢰구간의 범위에 포함되지 않는다.
즉 귀무가설을 기각하고 대립가설을 채택한다. 청명환의 효과가 있다.

4. 독립표본의 개념에 대해 설명하고, 주로 어떤 경우에 활용되는지 설명하라.

두 개의 모집단으로부터 표본을 추출할 때, 표본을 구성하는 각각의 인자에 상관없이 추출하여 구성된 표본을 독립표본(independent sample)이라 한다. 두 개의 모집단으로부터 추출한 표본을 서로 비교하는 경우이므로 이 두 표본은 서로 독립되어 어떠한 연관성이 전혀 없다. 평균차이를 비교한다는 것을 기준으로 보면 한 번의 추출을 각각 진행하여 두 개의 표본을 만들고, 이들을 서로 비교하여 검정하는 방법이다. 독립표본 t검정은 서로 다른 모집단으로부터 표본을 추출해서 서로 평균을 비교하여 차이를 검정하는 방법이다. 그런데 비교하기 전에 먼저 파악해야 할 것이 서로 다른 모집단으로부터 구성된 표본을 비교하는 분석방법이므로 표본a와 표본b의 분포를 먼저 알아야 한다. 서로 평균을 비교해 보는 것이지만 평균으로는 표본의 특성을 나타내는데 부족하기에 분산을 알아야 하고, 표본 a와 표본 b의 분포는 분산이 같은 경우와 다른 경우로 나누어 생각해야 한다.

5. [컴퓨터 실습] 국내 1, 2위 시장을 점유하는 타이어 제조업체들이 서로 자사 제품이 최고라는 광고를 한다. 연습문제.xlsx 파일의 '8장_타이어 주행거리' 시트는 동급의 타이어를 A사와 B사에서 각각 25개씩 표본 추출해서 주행거리를 측정한 자료다. A사와 B사의 동급 타이어의 주행거리에 차이가 있는지 유의수준 5% 이내에서 검정하라.

H_0 : A 와 B사의 타이어의 주행거리에 차이가 없다.
H_1 : A사와 B사의 타이어의 주행거리에 차이가 있다.

$\bar{x}_A = 58{,}640$, $\bar{x}_B = 55{,}400$, $s_A^2 = 51{,}420{,}000$, $s_B^2 = 38{,}166{,}667$

$$s_p^2 = \frac{(n_A-1)\times s_A^2 + (n_B-1)\times s_B^2}{(n_A-1)-(n_B-1)} = \frac{(25-1)\times 51{,}420{,}000 + (25-1)\times(25-1)\times 38{,}166{,}667}{50-2} \fallingdotseq 44{,}703{,}333$$

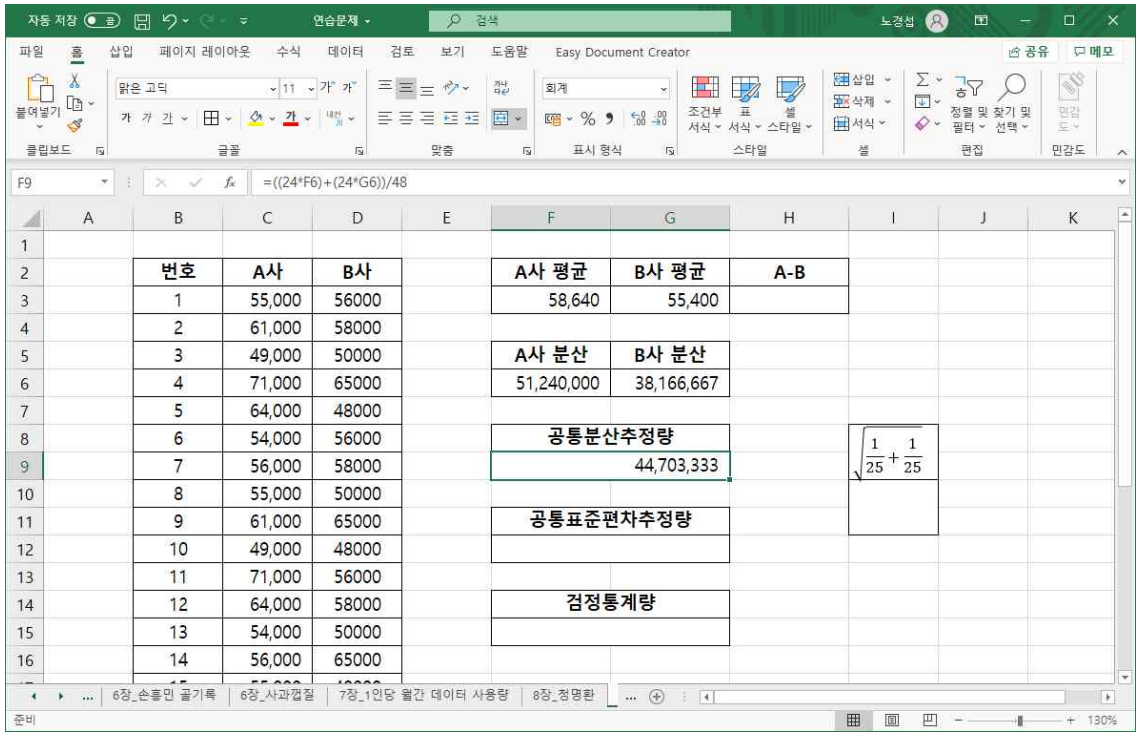
$s_p = \sqrt{44{,}703{,}333} \fallingdotseq 6{,}686$

유의수준 0.05이며 자유도가 24인 $t_{(24,\frac{\alpha}{2})} = 2.064$이다.

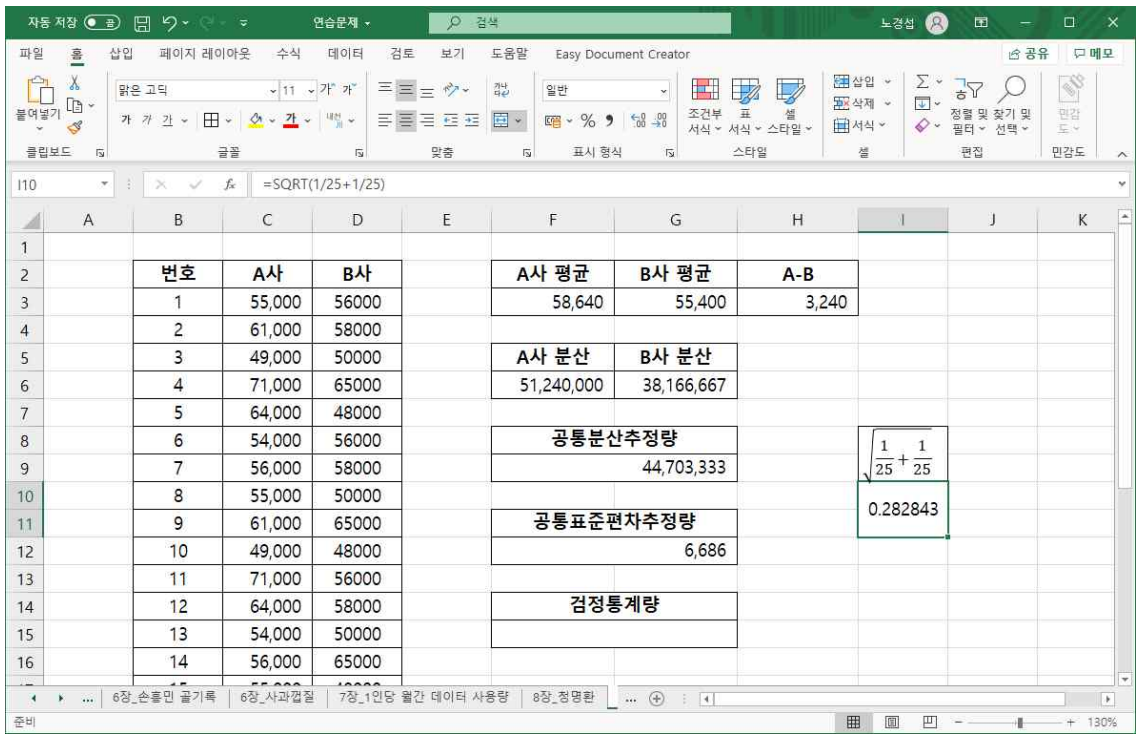
$$t_{(n_A+n_B-2,\frac{\alpha}{2})} = \frac{(\bar{x}_A - \bar{x}_B)}{s_p\sqrt{(\frac{1}{n_A}+\frac{1}{n_B})}} = \frac{58{,}640-55{,}400}{6{,}686\times\sqrt{\frac{1}{25}+\frac{1}{25}}} \fallingdotseq 1.713$$

$t = 2.064$이므로 1.713은 채택역에 속하므로 귀무가설을 기각할 수 없다. 따라서 A사와 B사의 타이어 주행거리는 차이가 있다고 할 수 없다.

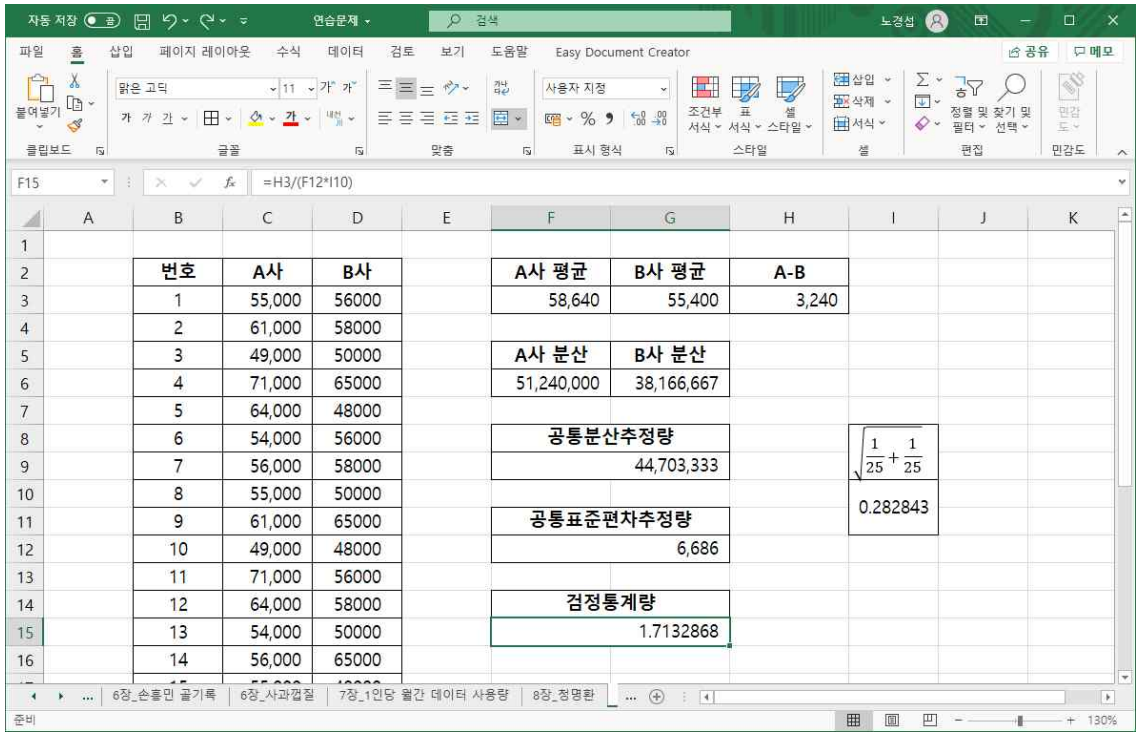
● Excel 풀이



A사와 B사의 평균과 분산을 구하고, 공통분산추정량을 구하기 위해 '=((24*F6)+(24*G6))/48'을 입력한다.



검정통계량 계산의 편의를 위해 A사 평균과 B사 평균의 차를 구하고, $\sqrt{\frac{1}{25}+\frac{1}{25}}$ 를 구한다.



검정통계량 계산을 위해 '=H3/(F12*I10)'을 입력하여 검정통계량을 계산한다.

6. [컴퓨터 실습] A방송국에서는 미니 시리즈 16부작 드라마 2편을 방송하였다. a작품은 호화 캐스팅, 현지 로케, 해외 스타 출연 등에 200억 원의 제작비를 들였고, b작품은 탄탄한 구성, 치밀한 준비, 무대 경험이 풍부한 무명 연기자들 섭외에 30억 원을 들였다. 연습문제.xlsx의 '8장_드라마 시청률' 시트는 25개 가구를 기준으로 16주 동안 두 작품의 시청률을 측정한 자료다. 두 드라마의 시청률에 차이가 있는지 확인하라.

$H_0 : p_a - p_b = 0$ 작비에 따라 a작품과 b작품의 시청률에 차이가 없다.
$H_1 : p_a - p_b \neq 0 \Rightarrow$ 제작비에 따라 a작품과 b작품의 시청률에 차이가 있다.

a작품과 b작품의 평균 시청률 $\bar{x}_a$, $\bar{x}_b$를 구하면, $\bar{x}_a \fallingdotseq 2.394$, $\bar{x}_b \fallingdotseq 2.588$이고

비율 $\hat{p}_a$, $\hat{p}_b$를 구해보면, $\hat{p}_a \fallingdotseq 0.150$, $\hat{p}_b \fallingdotseq 0.162$이다.

95%에서의 $z = 1.96$이며,

모집단 비율 차이에서 설명되는 표준오차는

$$s(\hat{p}_a - \hat{p}_b) = \sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a} + \frac{\hat{p}_b(1-\hat{p}_b)}{n_b}}$$
$$= \sqrt{\frac{0.150(1-0.150)}{16} + \frac{0.162(1-0.162)}{16}} \fallingdotseq 0.128$$

그러므로 95%에서의 신뢰구간은

$$(\hat{p}_a-\hat{p}_b)-z_{\frac{\alpha}{2}}\times\sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a}+\frac{\hat{p}_b(1-\hat{p}_b)}{n_b}}\le p_a-p_b\le(\hat{p}_a-\hat{p}_b)+z_{\frac{\alpha}{2}}\times\sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a}+\frac{\hat{p}_b(1-\hat{p}_b)}{n_b}}$$

직접 대입해보면

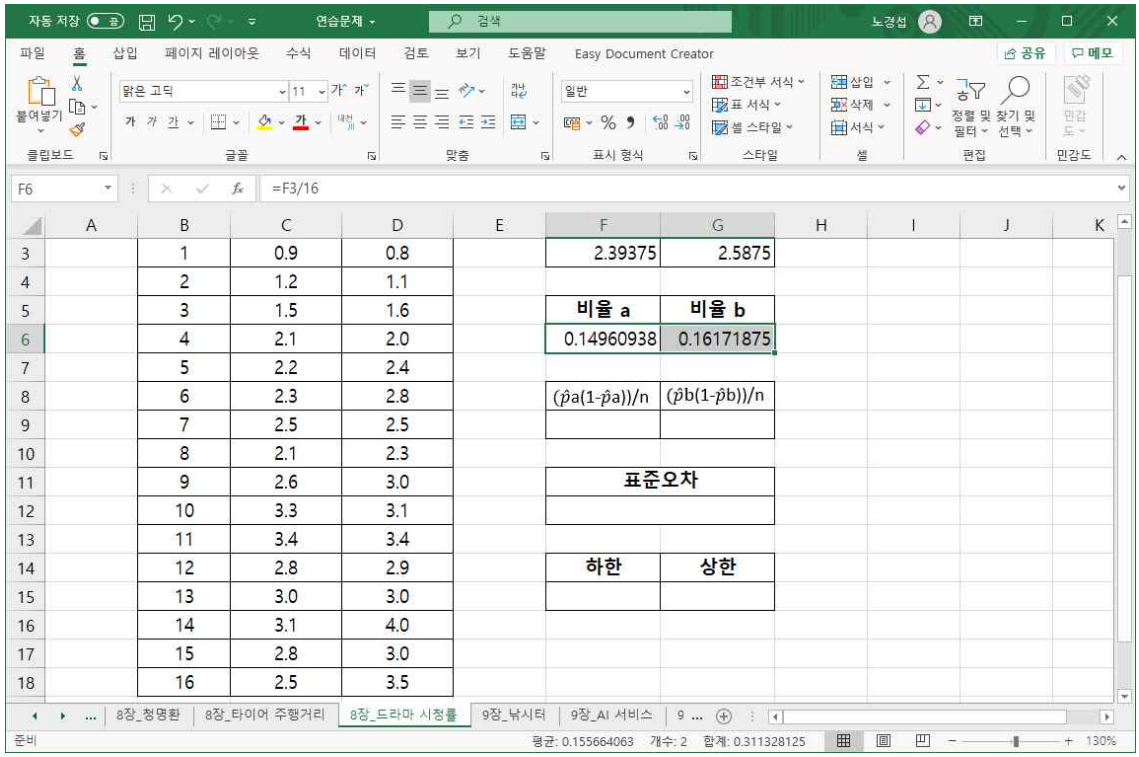
$$(0.150-0.162)-1.96\times\sqrt{\frac{0.150(1-0.150)}{16}+\frac{0.162(1-0.162)}{16}}$$

$$\le p_a-p_b\le$$

$$(0.150-0.162)+1.96\times\sqrt{\frac{0.150(1-0.150)}{16}+\frac{0.162(1-0.162)}{16}}$$

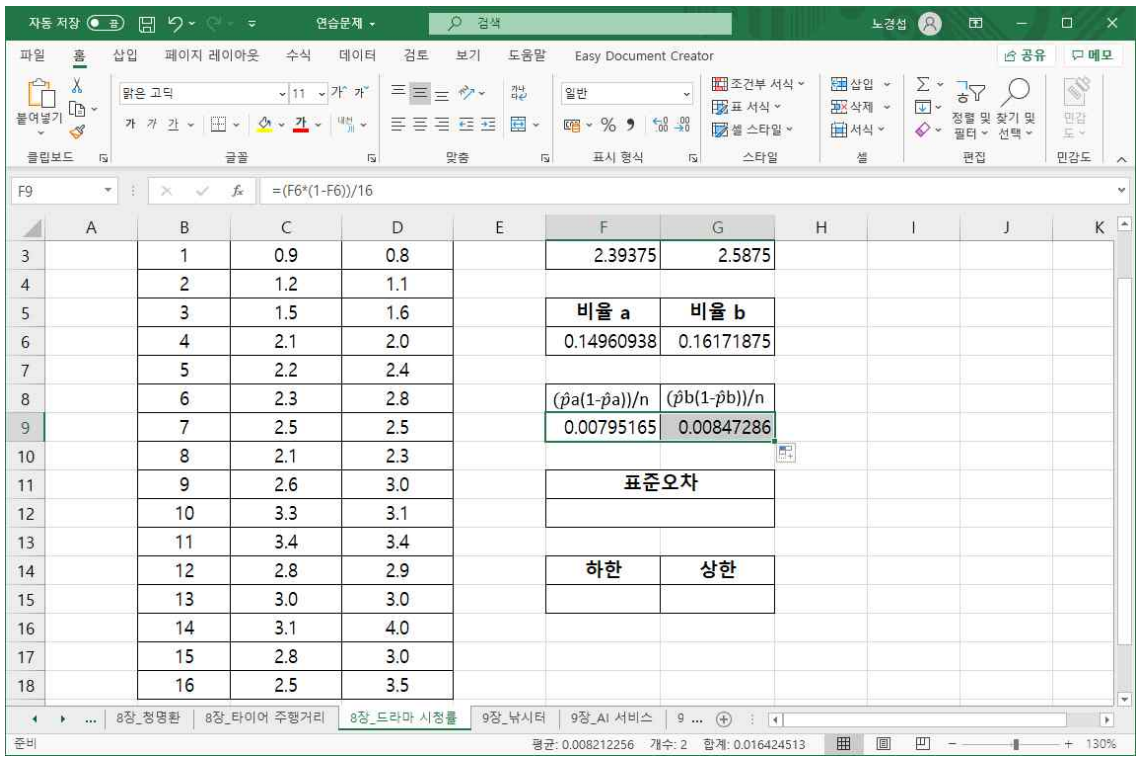
$$-0.263\le p_a-p_b\le 0.239$$

● Excel 풀이



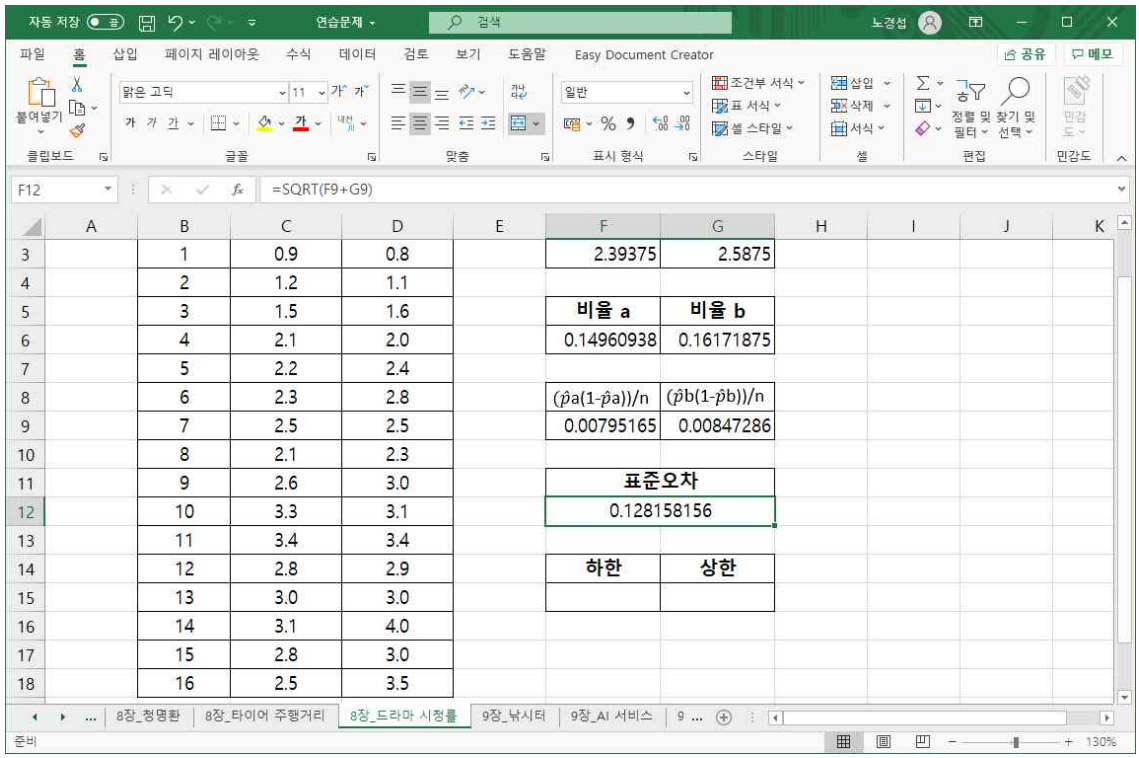
	A	B	C	D	E	F	G	H	I	J	K
3		1	0.9	0.8		2.39375	2.5875				
4		2	1.2	1.1							
5		3	1.5	1.6		비율 a	비율 b				
6		4	2.1	2.0		0.14960938	0.16171875				
7		5	2.2	2.4							
8		6	2.3	2.8		$(\hat{p}a(1-\hat{p}a))/n$	$(\hat{p}b(1-\hat{p}b))/n$				
9		7	2.5	2.5							
10		8	2.1	2.3							
11		9	2.6	3.0		표준오차					
12		10	3.3	3.1							
13		11	3.4	3.4							
14		12	2.8	2.9		하한	상한				
15		13	3.0	3.0							
16		14	3.1	4.0							
17		15	2.8	3.0							
18		16	2.5	3.5							

작품a와 작품b의 평균시청률과 전체에 대한 비율을 구한다.



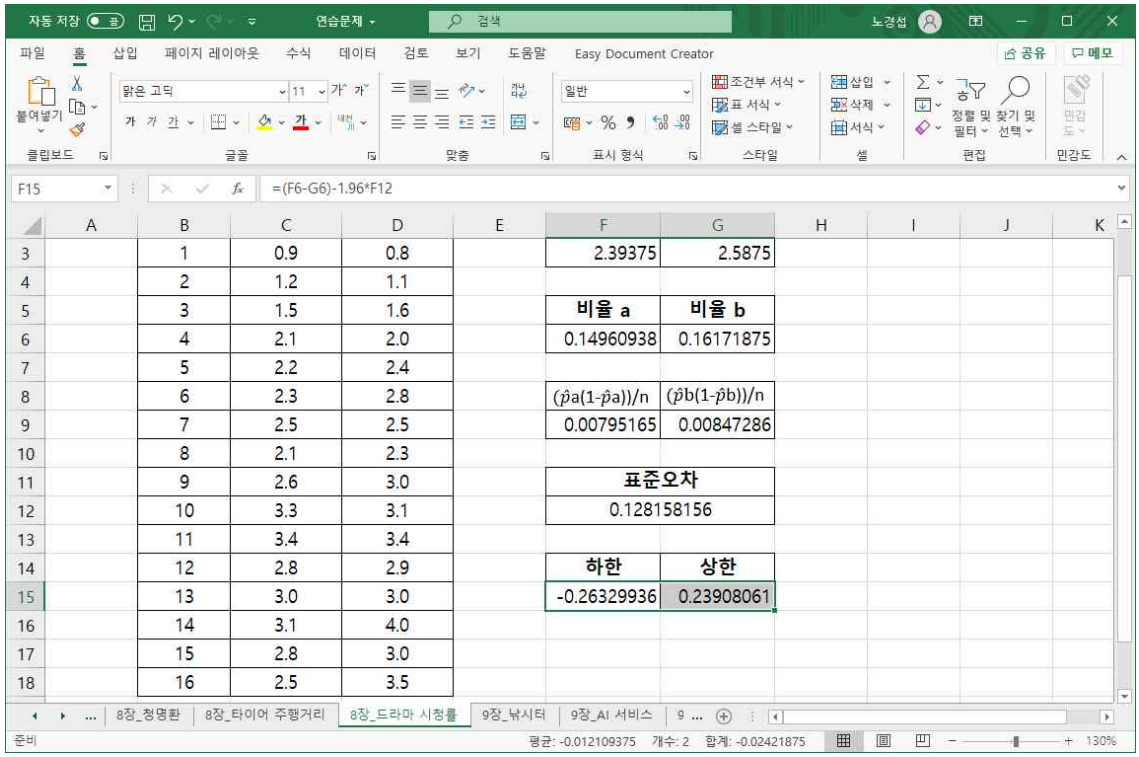
	A	B	C	D	E	F	G	H	I	J	K
3		1	0.9	0.8		2.39375	2.5875				
4		2	1.2	1.1							
5		3	1.5	1.6		비율 a	비율 b				
6		4	2.1	2.0		0.14960938	0.16171875				
7		5	2.2	2.4							
8		6	2.3	2.8		$(\hat{p}a(1-\hat{p}a))/n$	$(\hat{p}b(1-\hat{p}b))/n$				
9		7	2.5	2.5		0.00795165	0.00847286				
10		8	2.1	2.3							
11		9	2.6	3.0		표준오차					
12		10	3.3	3.1							
13		11	3.4	3.4							
14		12	2.8	2.9		하한	상한				
15		13	3.0	3.0							
16		14	3.1	4.0							
17		15	2.8	3.0							
18		16	2.5	3.5							

계산의 실수를 막고 편하게 하기 위해 $\frac{\hat{p}_a \times (1-\hat{p}_a)}{n_a}$와 $\frac{\hat{p}_b \times (1-\hat{p}_b)}{n_b}$를 먼저 계산한다.



	A	B	C	D	E	F	G	H	I	J	K
3		1	0.9	0.8		2.39375	2.5875				
4		2	1.2	1.1							
5		3	1.5	1.6		비율 a	비율 b				
6		4	2.1	2.0		0.14960938	0.16171875				
7		5	2.2	2.4							
8		6	2.3	2.8		(p̂a(1-p̂a))/n	(p̂b(1-p̂b))/n				
9		7	2.5	2.5		0.00795165	0.00847286				
10		8	2.1	2.3							
11		9	2.6	3.0		표준오차					
12		10	3.3	3.1		0.128158156					
13		11	3.4	3.4							
14		12	2.8	2.9		하한	상한				
15		13	3.0	3.0							
16		14	3.1	4.0							
17		15	2.8	3.0							
18		16	2.5	3.5							

$\frac{\hat{p}_a \times (1-\hat{p}_a)}{n_a}$와 $\frac{\hat{p}_b \times (1-\hat{p}_b)}{n_b}$를 이용하여 표준오차를 계산한다.



F15: =(F6-G6)-1.96*F12

	A	B	C	D	E	F	G	H	I	J	K
3		1	0.9	0.8		2.39375	2.5875				
4		2	1.2	1.1							
5		3	1.5	1.6		비율 a	비율 b				
6		4	2.1	2.0		0.14960938	0.16171875				
7		5	2.2	2.4							
8		6	2.3	2.8		(p̂a(1-p̂a))/n	(p̂b(1-p̂b))/n				
9		7	2.5	2.5		0.00795165	0.00847286				
10		8	2.1	2.3							
11		9	2.6	3.0		표준오차					
12		10	3.3	3.1		0.128158156					
13		11	3.4	3.4							
14		12	2.8	2.9		하한	상한				
15		13	3.0	3.0		-0.26329936	0.23908061				
16		14	3.1	4.0							
17		15	2.8	3.0							
18		16	2.5	3.5							

계산된 표준오차를 이용하여 신뢰구간의 하한과 상한을 구한다.

7. [컴퓨터 실습] [연습문제 6]을 공통 비율을 이용하여 검정하라.

$H_0 : p_a - p_b = 0$, $H_1 : p_a - p_b \neq 0$ 의 가설을 검정하기 위해, $H_0 : p_a = p_b$ 이므로 공통비율을 이용한다.

$$\hat{p}_{ab} = \frac{a\text{작품 표본수} \times a\text{작품 비율} + b\text{작품 표본수} \times b\text{작품 비율}}{\text{체 표본의 수}}$$

$$= \frac{16 \times 0.150 + 16 \times 0.162}{32}$$

$$= 0.156$$

$$z = \frac{(0.150 - 0.162)}{\sqrt{0.156(1-0.156) \times (\frac{1}{16} + \frac{1}{16}}} = 0.104$$

$\alpha = 0.05$에서의 양측검정 기각역은 $|\pm 1.96| > |z|$이므로, $H_0 : p_a - p_b = 0$을 채택한다.

Chapter 09 연습문제

1. 분산분석의 개념에 대해 설명하고, t검정과 어떤 차이가 있는지 설명하라.

t검정이 1개의 집단 혹은 2개의 집단에 대해 평균차이를 비교하여 차이를 검정하지만, 분산분석(analysis of variance ; ANOVA)은 3개 이상의 집단에 대한 평균차이를 검증하는 분석방법이며, 특성에 대한 산포의 제곱합을 요인별 제곱합으로 분해한 후 영향요인을 찾아내는 방법이다. t검정에서는 직접적으로 집단 2개에 대한 차이를 비교하였지만, 3개 이상의 집단을 비교하는 경우는 직접 비교하는 방법이 상당히 복잡해진다. 때문에 분산분석을 사용하는 것이 편리하며, 집단 간의 분산과 집단 내의 분산을 확인하여 모집단의 특성을 찾아내기에 적합하다. 3개 이상의 집단에 대한 평균차이의 검정을 위해서 분산을 비교하는 분석방법이다.

2. 분산분석의 종류와 차이점에 대해 설명하라.

일원분산분석이란 단 한가지의 요인을 기준으로 집단 간의 차이를 조사하는 것이다. 예를 들면, 각기 다른 편의점 3개 이상을 대상으로 고객의 만족도를 조사하는 경우이다. 이원분산분석이란 두 가지의 요인을 기준으로 집단 간의 차이를 조사하는 것이다. 예를 들면, 3종의 편의점을 위치에 따라 나누고 만족도를 조사하는 경우이다. 다원분산분석이란 세 가지 이상의 요인을 기준으로 집단 간의 차이를 조사하는 것이다.

3. 분산분석을 실시하기 위해 필요한 가정을 나열하고, 이에 대해 설명하라.

❶ 각 모집단은 정규분포이다.

표본으로 모집단 간의 평균차이를 추정하는 것이므로 모집단은 정규분포를 구성해야 하고, 정규분포이지만 집단 간 평균은 서로 다를 수 있다.

❷ 집단 간 분산은 서로 동일해야 한다.

집단 간 비교에 있어서 분산이 동일하지 않으면 Chapter 09에서 등분산인 경우에 t검정을 하는 경우와 같이 분산분석 역시 집단 간의 평균비교이므로 분산이 다르면 집단 간의 평균 차이를 구별하기 쉽지 않다. 분산분석은 집단이 3개 이상의 경우에 사용하는 분석 방법이므로 분산이 다르면 계산이 어려워진다.

❸ 각 표본들은 독립적으로 추출되어야 한다.

표본을 구성하는 과정에서 각각의 표본들은 모두 독립적으로 구성되어야 한다. 즉 표본을 구성하는 과정에서 어느 집단이 다른 집단에 영향을 주지 않아야 한다.

❹ 각 표본의 크기는 적절해야 한다.

분석을 진행하기 위해서는 표본의 크기가 충분해야 한다. 이는 통계학 전체에 적용되는 기준이 되는 '충분성'과 연결되는 가정이며, 만약 집단 간 표본크기의 차이가 많이 나는 경우라면 편의가 발생할 수 있기에 적절한 연구조사가 되기 힘들지만, 집단들의 표본이 충분하게 추출되었다면 분산분석을 실시할 때 표본의 개수에 상관없이 분석을 진행할 수 있다.

4. 국내 유명 낚시터 3곳에서 하루 3시간씩 낚시를 하여 각각의 낚시터에서 잡힌 물고기 수에 차이가 있는지 알아보고자 한다. 그 결과가 다음과 같을 때, $\alpha = 0.05$에서 검정하라.

횟수	A낚시터	B낚시터	C낚시터
1	7	5	3
2	5	8	4
3	8	10	3
4	4	4	7
5	6	8	5
6	2	6	3
7	3	7	3
8	5	5	2
9	3	8	4
10	–	–	5

H_0 : 시터 별 3시간 동안 잡힌 물고기 마리수는 차이가 없다.
H_1 : 낚시터 별 3시간 동안 잡힌 물고기 마리수는 차이가 있다.

$\overline{x} = 3.693$, $\overline{x_1}$ 평균 $= 4.711$, $\overline{x_2}$ 평균 $= 6.778$, $\overline{x_3}$ 평균 $= 3.880$으로

총편차 $= (7-3.693)^2 + (5-3.693)^2 + (8-3.693)^2 + \cdots + (3-3.693)^2 = 123.66$

집단간편차 $= 9 \times (4.711-3.693)^2 + 9 \times (6.778-3.693)^2 + 10 \times (3.880-3.693)^2 = 41.56$

집단내편차 $SSE_1 = (7-4.711)^2 + (5-4.711)^2 + \cdots + (3-4.711)^2 = 33.549$

$SSE_2 = (5-6.778)^2 + (8-6.778)^2 + \cdots + (8-6.778)^2 = 29.556$

$SSE_3 = (3-3.880)^2 + (4-3.880)^2 + \cdots + (5-3.444)^2 = 18.996$

$SSE_1 + SSE_2 + SSE_3 = 82.100$

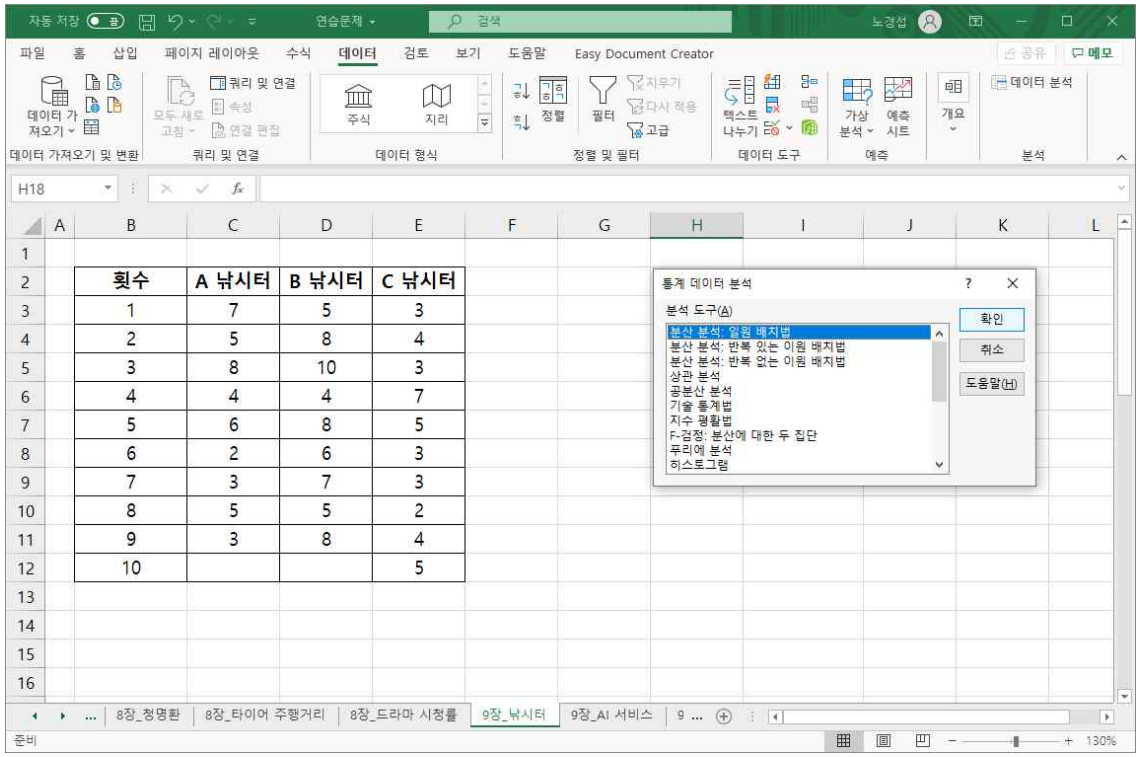
총편차 = 집단간편차 + 집단내편차 $SST = SSB + SSW \Rightarrow 123.66 = 41.56 + 82.10$

집단간분산 $\Rightarrow MSB = \dfrac{41.56}{2} = 20.78$, 집단내분산 $\Rightarrow MSW = \dfrac{82.10}{25} = 3.28$

$\therefore F = \dfrac{\text{집단간}}{\text{집단내}} = \dfrac{20.78}{3.28} = 6.33$

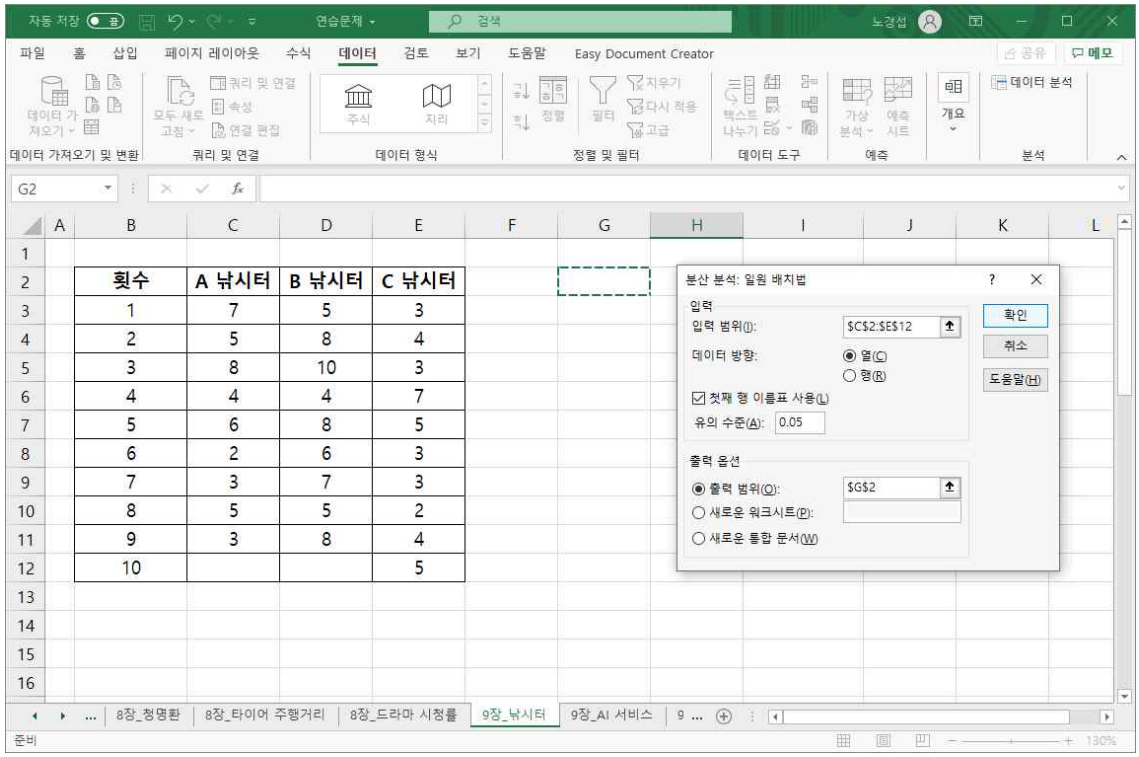
따라서 6.33배의 차이가 있다는 것을 알 수 있었으며, 분자의 자유도는 2, 분모의 자유도는 21에 해당하는 F분포표의 값은 3.47이므로 6.33은 3.47보다 크다. 즉, 6.33가 나오게 될 확률은 5%보다 더 작아진다. 그리하여 귀무가설을 기각하고 대립가설을 채택하여 낚시터 별로 3시간동안 잡힌 물고기 양은 차이가 있는 것으로 나타났다.

5. [컴퓨터 실습] 연습문제.xlsx의 '9장_낚시터' 시트를 이용하여 [연습문제 4]를 구하라.



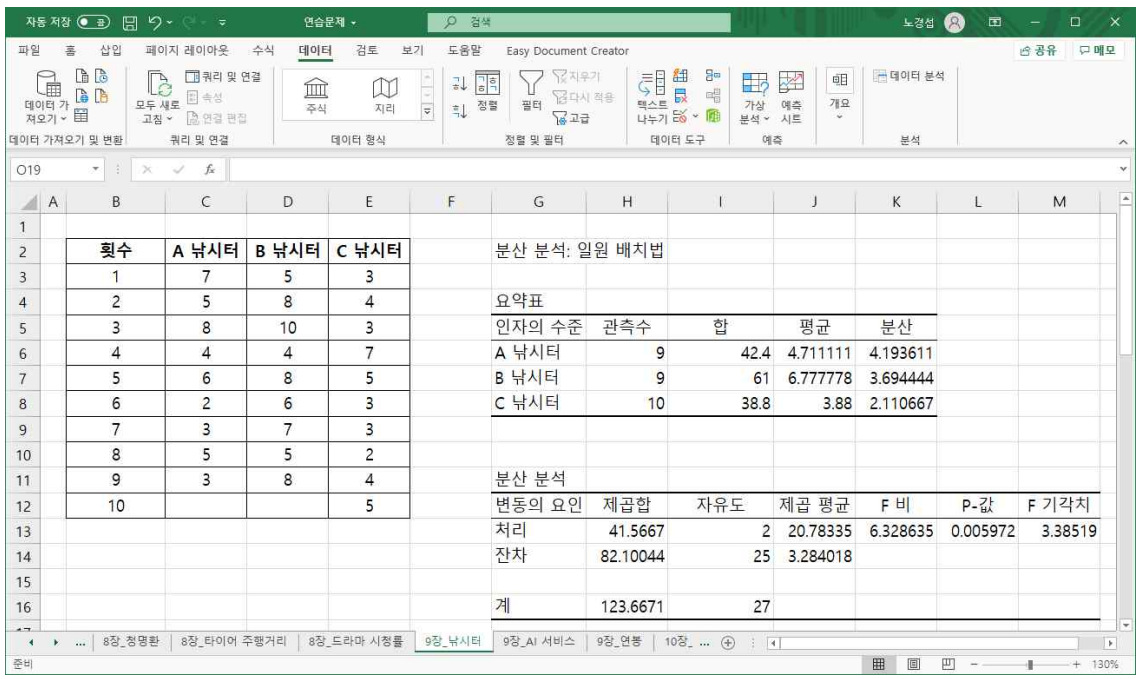
횟수	A 낚시터	B 낚시터	C 낚시터
1	7	5	3
2	5	8	4
3	8	10	3
4	4	4	7
5	6	8	5
6	2	6	3
7	3	7	3
8	5	5	2
9	3	8	4
10			5

[데이터] 탭으로 이동하여 [데이터 분석] 메뉴를 클릭하고, 창이 열리면 '분산분석: 일원배치법'을 선택한다.



횟수	A 낚시터	B 낚시터	C 낚시터
1	7	5	3
2	5	8	4
3	8	10	3
4	4	4	7
5	6	8	5
6	2	6	3
7	3	7	3
8	5	5	2
9	3	8	4
10			5

입력 범위를 선택하고, 출력 범위를 G2셀로 선택한다.



횟수	A 낚시터	B 낚시터	C 낚시터
1	7	5	3
2	5	8	4
3	8	10	3
4	4	4	7
5	6	8	5
6	2	6	3
7	3	7	3
8	5	5	2
9	3	8	4
10			5

분산 분석: 일원 배치법

요약표

인자의 수준	관측수	합	평균	분산
A 낚시터	9	42.4	4.711111	4.193611
B 낚시터	9	61	6.777778	3.694444
C 낚시터	10	38.8	3.88	2.110667

분산 분석

변동의 요인	제곱합	자유도	제곱 평균	F 비	P-값	F 기각치
처리	41.5667	2	20.78335	6.328635	0.005972	3.38519
잔차	82.10044	25	3.284018			
계	123.6671	27				

출력 결과가 완성되었다.

6. 인공지능 스피커로 서비스를 제공하는 A, B, C의 3개 회사를 기준으로 서울, 대전, 광주, 부산에 거주하는 소비자들의 서비스 품질 만족도를 조사한 결과가 다음과 같다. 통신사별, 지역별로 소비자들이 느끼는 품질의 만족도에 차이가 있는지 $\alpha = 0.05$에서 검정하라.

통신사 별 H_0^i : 신사별 소비자 만족도에 차이가 없다.
H_1^i : 통신사별 소비자 만족도에 차이가 있다.

상권 별 H_0^j : 상권별 소비자 만족도에 차이가 없다.
H_1^j : 상권별 소비자 만족도에 차이가 있다.

상호작용 별 H_0^{ij} : 상호작용과 소비자 만족도에 차이가 없다.
H_1^{ij} : 상호작용과 소비자 만족도에 차이가 있다.

$$SST = (8-6.479)^2+(7-6.479)^2+(9-6.479)^2+\cdots+(5-6.479)^2 = 115.979$$
$$SSB_i = 16\times(7.313-6.479)^2+16\times(6.625-6.479)^2+16\times(5.500-6.479)^2 = 28.800$$
$$SSB_j = 16\times(3.111-2.926)^2+16\times(2.778-2.926)^2+16\times(2.889-2.926)^2 = 4.975$$
$$SSB_{ij} = 4\times(8.000-7.313-6.000+6.479)^2+4\times(6.000-7.313-6.333+6.479)^2$$
$$+4\times(7.250-7.313-6.917+6.479)^2+\cdots+4\times(7.750-5.500-6.500+6.479)^2$$
$$+4\times(5.000-5.500-6.500+6.479)^2 = 45.708$$
$$SSW = SST-SSB_i-SSB_j-SSB_{ij} = 36.424$$

종류	A 텔레콤	B 텔레콤	C 텔레콤	지역전체평균
서울	8	7	3	6.167
	7	5	6	
	9	8	4	
	8	6	3	
지역텔레콤평균	8.000	6.500	4.000	
대전	5	9	7	6.333
	7	8	6	
	6	8	5	
	6	6	3	
지역텔레콤평균	6.000	7.750	5.250	
광주	7	7	7	6.917
	8	5	7	
	8	5	8	
	6	6	9	
지역텔레콤평균	7.250	5.750	7.750	
부산	8	6	5	6.500
	9	7	5	
	7	7	5	
	8	6	5	
지역텔레콤평균	8.000	6.500	5.000	
텔레콤평균	7.313	6.625	5.500	6.479
SST	115.979			
SSBi	28.800			
SSBj	4.975			
SSBij	45.708			

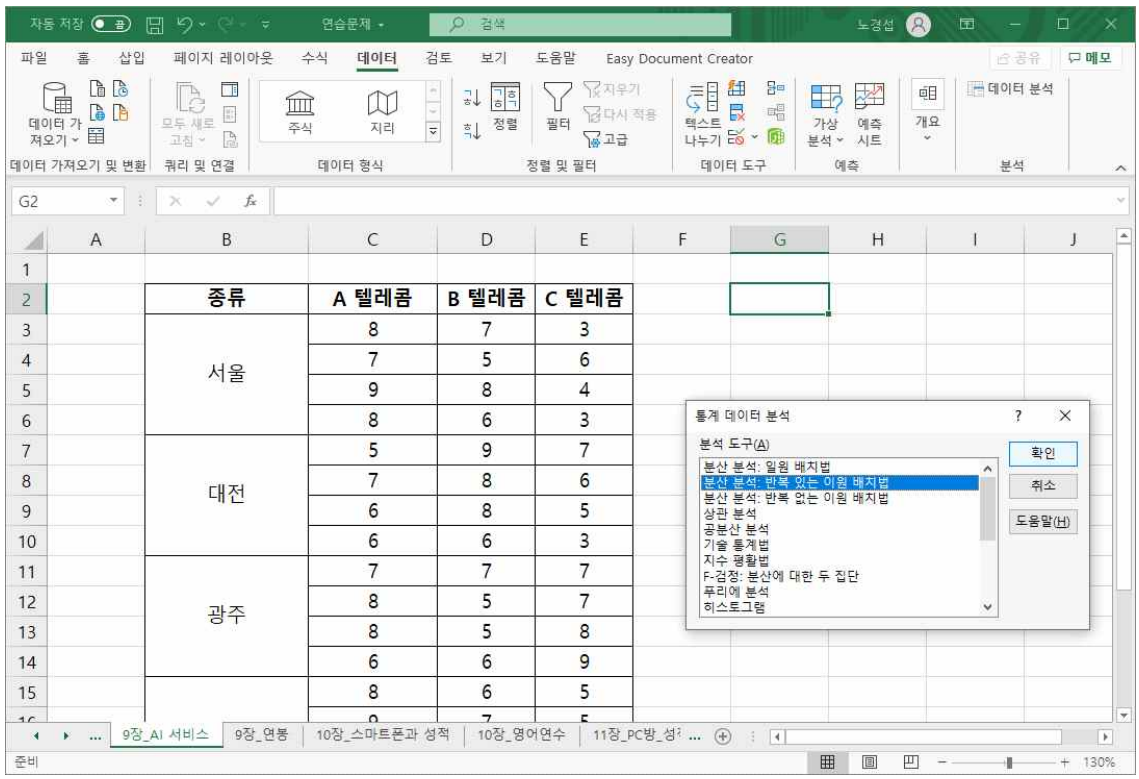
1)

1) 계산하면서 참고할 것.

구분	편차(제곱합)	자유도	평균제곱	분산비율 F
i j ij 상호작용 집단 내	$SSB_i = 28.800$ $SSB_j = 4.975$ $SSB_{ij} = 45.708$ $SSW = 36.424$	$i-1=2$ $j-1=3$ $(i-1)(j-1)=6$ $i \times j \times (k-1)=36$	$MSB_i = 14.400$ $MSB_j = 1.658$ $MSB_{ij} = 7.618$ $MSW = 1.012$	$MSB_i/MSW = 14.229$ $MSB_j/MSW = 1.665$ $MSB_{ij}/MSW = 7.083$
합계	$SST = 115.979$	$n-1=35$		

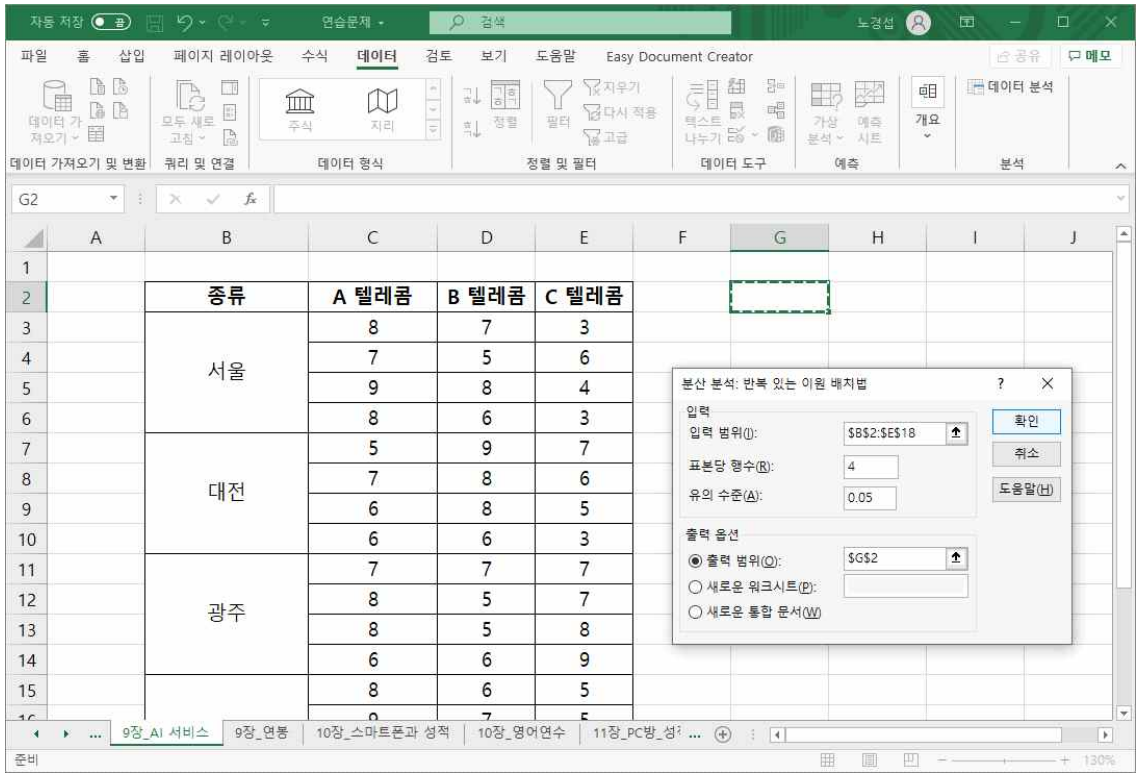
MSB_i, MSB_j, MSB_{ij}의 자유도인 2, 2, 4를 기준으로 F분포표의 값은 3,55, 3,55, 2.93이다. $MSB_j/MSW = 1.844$는 3.55를 넘어서지 못하므로 '상권별 소비자의 만족도에 따른 차이가 없다.'는 귀무가설을 기각하지 못한다.

7. [컴퓨터 실습] 연습문제.xlsx의 '9장_AI 서비스' 시트를 이용하여 [연습문제 6]을 구하라.



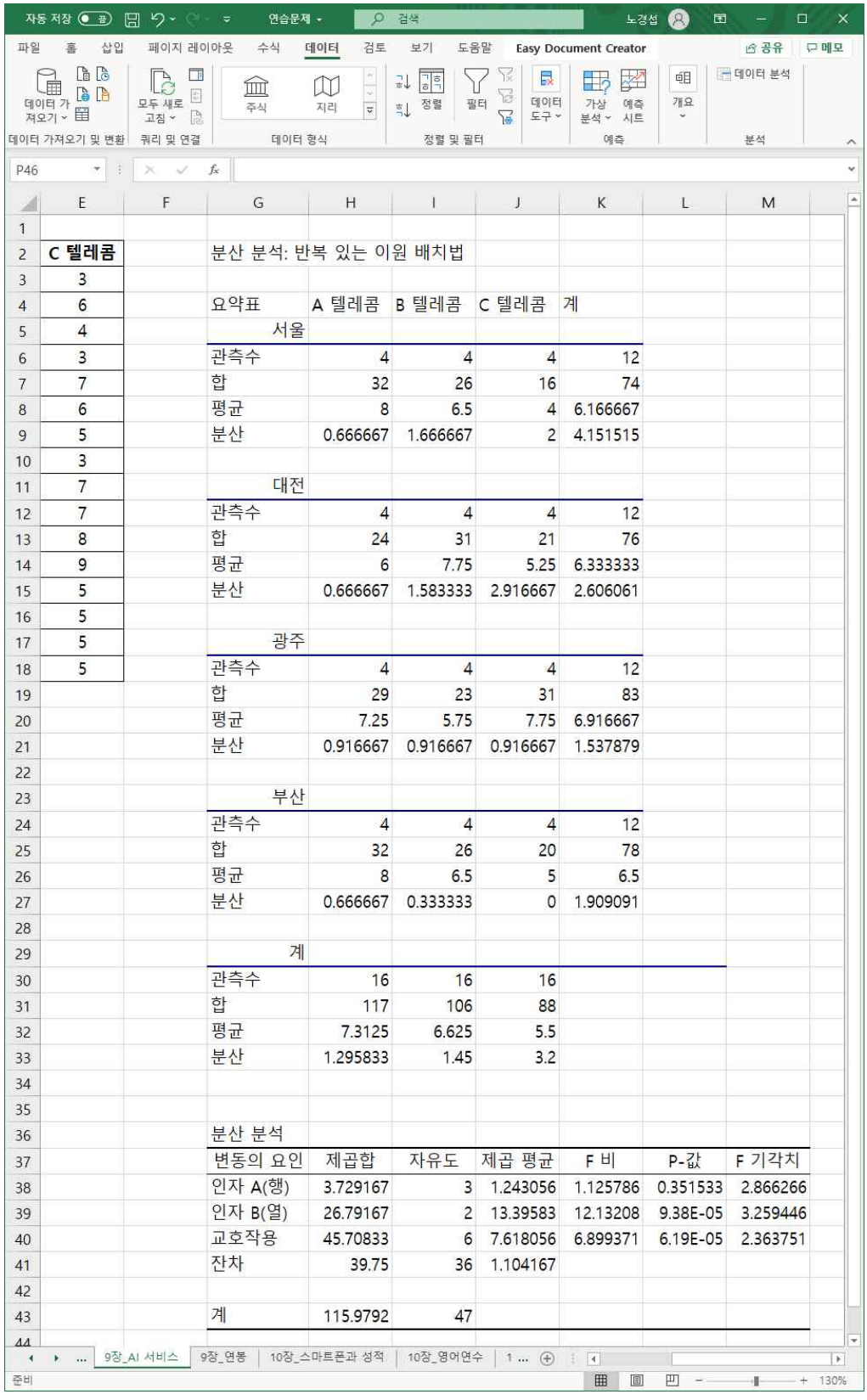
종류	A 텔레콤	B 텔레콤	C 텔레콤
서울	8	7	3
	7	5	6
	9	8	4
	8	6	3
대전	5	9	7
	7	8	6
	6	8	5
	6	6	3
광주	7	7	7
	8	5	7
	8	5	8
	6	6	9
	8	6	5

[데이터 탭]의 [데이터 분석] 메뉴를 클릭한다.



종류	A 텔레콤	B 텔레콤	C 텔레콤
서울	8	7	3
	7	5	6
	9	8	4
	8	6	3
대전	5	9	7
	7	8	6
	6	8	5
	6	6	3
광주	7	7	7
	8	5	7
	8	5	8
	6	6	9
	8	6	5

입력 범위를 입력하고 표본당 행수는 4, 출력 범위를 G2셀로 설정한다.



C 텔레콤
3
6
4
3
7
6
5
3
7
7
8
9
5
5
5
5

분산 분석: 반복 있는 이원 배치법

요약표	A 텔레콤	B 텔레콤	C 텔레콤	계
서울				
관측수	4	4	4	12
합	32	26	16	74
평균	8	6.5	4	6.166667
분산	0.666667	1.666667	2	4.151515
대전				
관측수	4	4	4	12
합	24	31	21	76
평균	6	7.75	5.25	6.333333
분산	0.666667	1.583333	2.916667	2.606061
광주				
관측수	4	4	4	12
합	29	23	31	83
평균	7.25	5.75	7.75	6.916667
분산	0.916667	0.916667	0.916667	1.537879
부산				
관측수	4	4	4	12
합	32	26	20	78
평균	8	6.5	5	6.5
분산	0.666667	0.333333	0	1.909091
계				
관측수	16	16	16	
합	117	106	88	
평균	7.3125	6.625	5.5	
분산	1.295833	1.45	3.2	

분산 분석

변동의 요인	제곱합	자유도	제곱 평균	F 비	P-값	F 기각치
인자 A(행)	3.729167	3	1.243056	1.125786	0.351533	2.866266
인자 B(열)	26.79167	2	13.39583	12.13208	9.38E-05	3.259446
교호작용	45.70833	6	7.618056	6.899371	6.19E-05	2.363751
잔차	39.75	36	1.104167			
계	115.9792	47				

분석 결과가 출력되었으나, 손풀이와 수치가 약간 다른 것을 확인할 수 있다. 손풀이에서는 소수점 4째 자리에서 반올림한 결과이므로 수치가 약간 다르다.

8. 일원 분산분석과 이원 분산분석을 진행하는 과정에서 집단 간, 집단 내, 상호작용 항에 관한 자유도를 구하는 방법을 설명하라.

1) 집단 내각 집단 별로 자유도를 합치면 각 집단 만큼의 수를 빼줘야 한다. 그러므로 n−i 가 된다.(i = 집단 수)

2) 집단 간추출된 표본 집단의 자유도 이므로 i−1 이 된다.

3) 상호작용각 집단의 자유도, 예를 들면, n−1, j−1 있다 하면 이 두 집단의 상호작용에 대한 자유도는 (n−1)(j−1)이다.

9. 이원 분산분석에서 상호작용 효과를 확인하는 이유는 무엇인지 설명하라.

이원분산분석의 일원분산분석과의 차이점은 독립변수가 2개라는 것이다. 독립변수가 2개이기 때문에 두 개의 독립변수가 동시에 작용하여 종속변수에 미치는 영향 즉, 상호작용 효과를 확인해야한다.

10. [컴퓨터 실습] 한국, 미국, 일본의 ICT 전문가들의 연봉을 비교해보기 위해 각 나라에서 무작위로 3명의 기업 임원급 ICT 전문가들의 연봉을 조사하였다. 연습문제.xlsx의 '9장_연봉' 시트를 이용해 각 나라별로 연봉 차이가 있는지 $\alpha = 0.05$에서 확인하라.

H_0 : 나라별 ICT전문가들의 연봉에 차이가 없다.

H_1 : 각 나라별 ICT전문가들의 연봉에 차이가 있다.

$\overline{x}$=345,555.56, $\overline{x_1}$평균=260,000, $\overline{x_2}$평균=433,333.33, $\overline{x_3}$평균=333,333.33으로

$$\text{총편차} = (150{,}000-345{,}555.56)^2+(350{,}000-345{,}555.56)^2+(8-3.693)^2 \\ +\cdots+(430{,}000-345{,}555.56)^2 = 108{,}222{,}222{,}222$$

$$\text{집단간편차} = (260.000-345{,}555.56)^2+(443{,}333.33-345{,}555.56)^2+(430{,}000-345{,}555.56)^2 \\ = 17{,}029{,}629{,}629.63$$

집단내편차

$$SSE_1 = (150{,}000-260{,}000)^2+(350{,}000-260{,}000)^2+(280{,}000-260{,}000)^2 = 20{,}600{,}000{,}000$$

$$SSE_2 = (340{,}000-443{,}333.33)^2+(540{,}000-443{,}333.33)^2+(450{,}000-443{,}333.33)^2 = 20{,}066{,}666{,}666.67$$

$$SSE_3 = (250{,}000-333{,}333.33)^2+(320{,}000-333{,}333.33)^2+(430{,}000-333{,}333.33)^2 = 16{,}466{,}666{,}666.67$$

$$SSE_1+SSE_2+SSE_3 = 57{,}133{,}333{,}333.33$$

총편차 = 집단간편차 + 집단내편차 $SST = SSB + SSW$

$$\Rightarrow 108{,}222{,}222{,}222 = 51{,}088{,}888{,}889 + 57{,}133{,}333{,}333$$

$$\text{단간분산} \quad MSB = \frac{51,088,888,888}{2} = 25,544,444,444$$

$$\text{집단내분산} \Rightarrow MSW = \frac{19,044,444,444}{6} = 9,522,222,222$$

$$\therefore F = \frac{\text{집단간}}{\text{집단내}} = \frac{25,544,444,444}{9,522,222,222} = 2.68$$

따라서 2.68배의 차이가 있다는 것을 알 수 있었으며, 분자의 자유도는 2, 분모의 자유도는 6에 해당하는 F분포표의 값은 5.14이므로 2.68은 5.14보다 작다. 그리하여 귀무가설을 채택하고 대립가설을 기각하여 국가별 ICT전문가들의 연봉은 차이가 없다고 한다.

Chapter 10 연습문제

1. 연관성 분석을 하는 이유와 그 종류에 대해 설명하라.

연관성 분석(association analysis)이란 조사대상에 대한 수집된 자료의 척도를 기준으로 변수들 간에 어느 정도 밀접한 관계가 있는지에 대해 판단하기 위한 분석을 의미한다. 변수를 구성하는 수집된 자료의 척도를 기준으로 연관성을 파악하는 것이기 때문에 척도에 따라 연관성을 파악하는 분석방법이 달라진다.

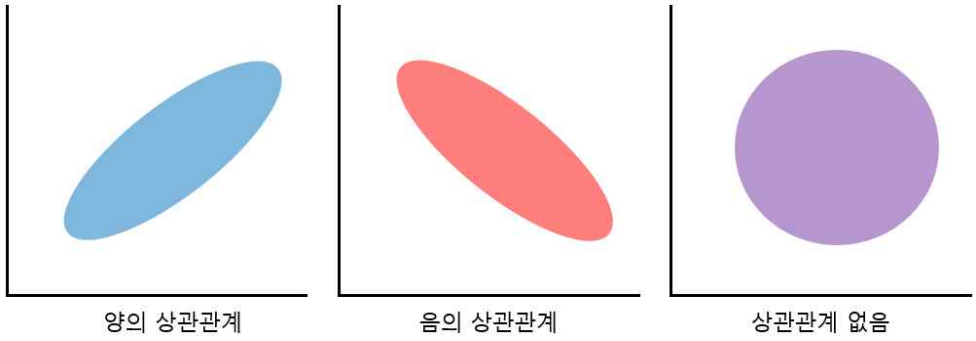
척도는 크게 범주형과 연속형으로 구분되며, 등간척도와 비율척도의 연속형 척도에 대한 연관성, 범주형 척도인 명목척도와 서열척도에 대한 연관성 분석방법이 각기 다르다.
상관분석은 주로 등간척도와 비율척도로 구성된 연속형 변수들 사이의 연관성을 분석하는 방법인데, 연속형 변수들에 대해서도 다른 변수들의 개입여부에 따라 편상관 분석과 피어슨 상관분석으로 구분되고, 명목척도와 서열척도로 구성된 범주형 변수의 경우 서열로 구성된 변수들은 스피어만 서열 상관분석을 하고, 명목척도와 서열척도로 구성된 경우는 교차분석으로 검정을 실시한다.

구분	사용 척도	분석 방법	기타 변수의 개입 여부
교차분석	명목척도, 서열척도	교차분석	
상관분석	서열척도	스피어만 서열 상관분석	
	등간척도, 비율척도	피어슨 상관분석	×
		편상관분석	○

2. 상관분석에 대해 설명하고, 산포의 종류에 따른 상관의 의미를 설명하라.

상관분석(correlation analysis)은 조사목적을 달성하기 위해 구성된 변수들 간의 연관관계를 분석하는 방법이다. 상관관계는 두 개의 변수를 기준으로 선형관계의 형태와 연관정도를 수치로 나타낸다. 키와 몸무게의 관계, 광고비와 매출액의 관계, 흡연과 수명의 관계 등의 연관성을 파악하는 경우가 이에 속한다.

그래프를 통한 산포도(scatter diagram)는 두 개의 변수를 각각 x축과 y축으로 구성하여 이들의 흩어진 정도를 표시한 것을 말한다. 산술적 계산방법을 사용하지 않고 도표로 표시하는 방법을 사용하는 이유는 측정치에 대한 산포도를 그림으로써 변수들 사이에 어떤 상관관계가 있는지 직관적으로 보고 판단하기 위해 사용한다.



위의 그림은 표본의 산포를 구분해서 양의 상관, 음의 상관, 상관없음의 3가지 형태로 나타낸 것이다.

3. 공분산에 대해 설명하고, 공분산으로 계산된 수치가 무엇을 의미하는지 설명하라.

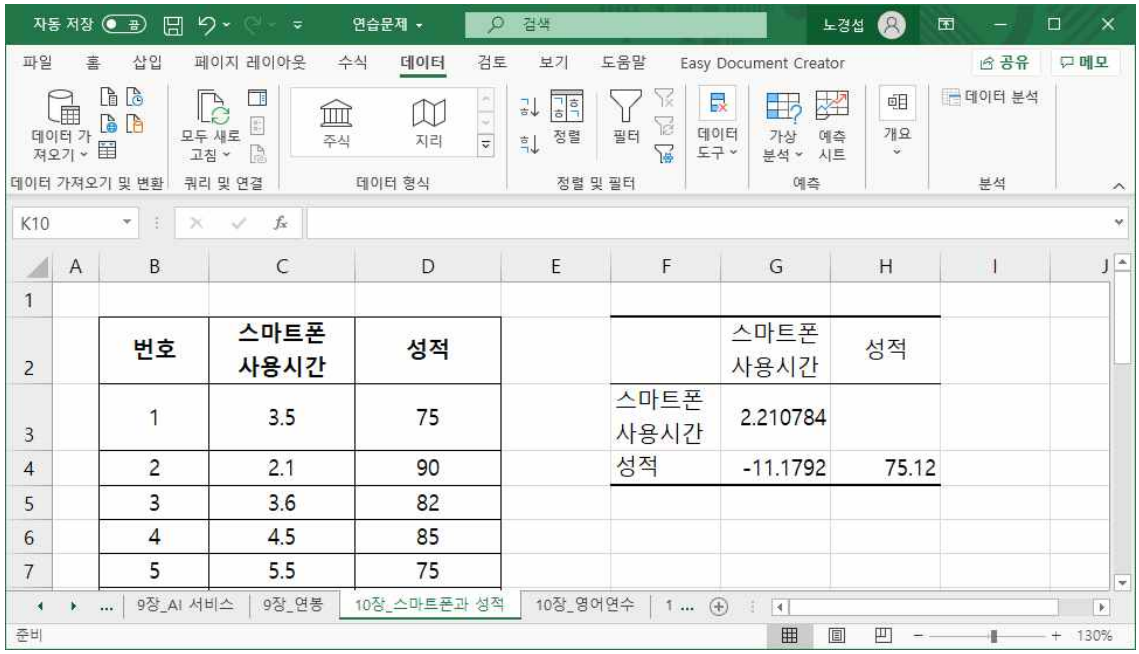
공분산(covariance)은 두 가지의 확률변수에 대한 흩어짐의 정도가 동일한 방향인 정(+)의 방향인가 혹은 반대방향인 음(−)의 방향인가를 나타내는 수치다. 즉, x가 변하면 y는 어떻게, 어느 정도로 변하는가와 같이 두 변수가 서로 변하는 정도를 수치로 나타낸 것이다. 확률변수 X에 대한 흩어짐의 정도를 산포도라 했고, 이를 분산으로 표시할 수 있다. 또 다른 확률변수 Y에 대한 흩어짐의 정도도 분산으로 표시할 수 있다. 이 각각의 분산에 대한 공통점이 공분산이며, 이에 대한 분석을 공분산분석이라 한다.

4. [컴퓨터 실습] 스마트폰 사용 시간과 성적 간의 관계를 조사하고자 한다. 임의로 지난 학기 통계학 수업을 듣는 25명을 대상으로 하루 평균 스마트폰 이용 시간과 성적에 대한 자료를 수집하였다. 연습문제.xlsx의 '10장_스마트폰과 성적' 시트를 이용하여 스마트폰의 사용 시간과 성적의 관계를 확인하기 위해 공분산을 구하라.

● 직접 계산

번호	스마트폰 사용시간	성적	스마트폰 사용시간 편차	성적 편차	스마트폰 사용시간 X 성적 편차
1	3.5	75	-0.20	-5.80	1.18
2	2.1	90	-1.60	9.20	-14.76
3	3.6	82	-0.10	1.20	-0.12
4	4.5	85	0.80	4.20	3.34
5	5.5	75	1.80	-5.80	-10.42
6	1.1	95	-2.60	14.20	-36.98
7	5.5	68	1.80	-12.80	-22.99
8	5.0	70	1.30	-10.80	-14.00
9	2.6	85	-1.10	4.20	-4.64
10	3.5	75	-0.20	-5.80	1.18
11	2.1	90	-1.60	9.20	-14.76
12	3.6	82	-0.10	1.20	-0.12
13	4.5	85	0.80	4.20	3.34
14	5.5	75	1.80	-5.80	-10.42
15	1.1	95	-2.60	14.20	-36.98
16	5.5	68	1.80	-12.80	-22.99
17	5.0	70	1.30	-10.80	-14.00
18	2.6	85	-1.10	4.20	-4.64
19	3.5	75	-0.20	-5.80	1.18
20	2.1	90	-1.60	9.20	-14.76
21	3.6	82	-0.10	1.20	-0.12
22	4.5	85	0.80	4.20	3.34
23	5.5	75	1.80	-5.80	-10.42
24	1.1	95	-2.60	14.20	-36.98
25	5.5	68	1.80	-12.80	-22.99
평균	3.70	80.80	편차곱의 총합		-279.48
표준편차	1.52	8.85	**공분산=(편차곱의 총합)/25**		**-11.18**

● 데이터 분석 도구 활용



	번호	스마트폰 사용시간	성적
1	1	3.5	75
2	2	2.1	90
3	3	3.6	82
4	4	4.5	85
5	5	5.5	75

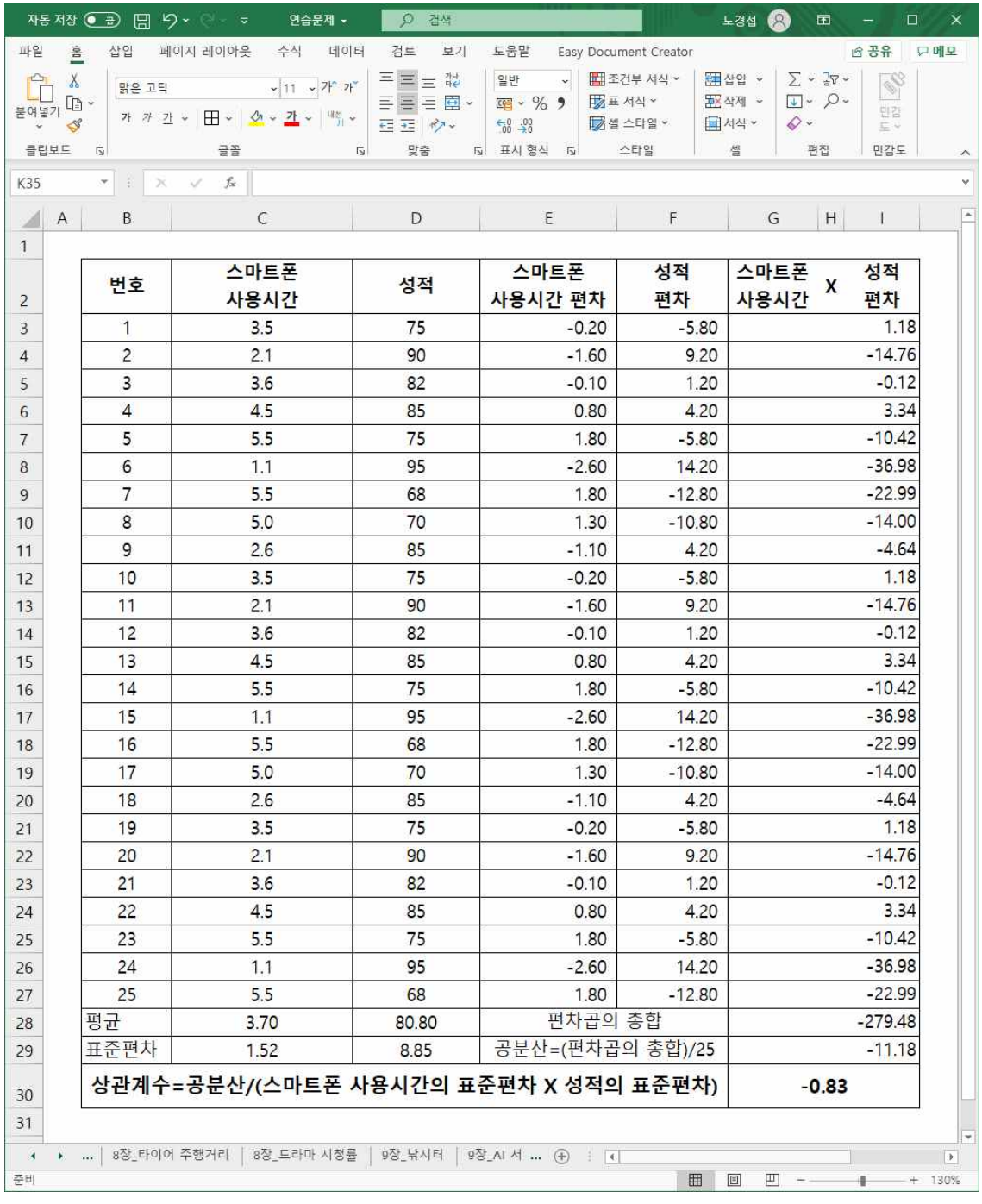
	스마트폰 사용시간	성적
스마트폰 사용시간	2.210784	
성적	-11.1792	75.12

5. 상관계수에 대해 설명하고, 상관계수의 특징을 설명하라.

상관계수(correlation coefficient)는 공분산을 표준화한 값이다. 공분산이 X와 Y값이 각각 양의 값끼리 대응하는 경우 공분산은 커지며, X와 Y의 값이 서로 양의 값과 음의 값끼리 대응하면 공분산은 작아지고, X의 작은 값과 Y의 큰 값이 대응하면 공분산은 작아지고, X와 Y가 일정한 규칙이 없이 대응하면 공분산은 0에 가까워진다고 했으나, 이러한 경우는 너무나 많으므로 단순히 숫자를 표시해서는 관계를 정확하게 파악하기 어렵다. 이런 공분산의 한계를 극복하기 위하여 '표준화'를 해야 하는데, 공분산을 구하고, X의 표준편차와 Y의 표준편차편차를 곱해서 나누어주면 표준화되며, 이처럼 표준화된 공분산 계수를 상관계수라 한다.

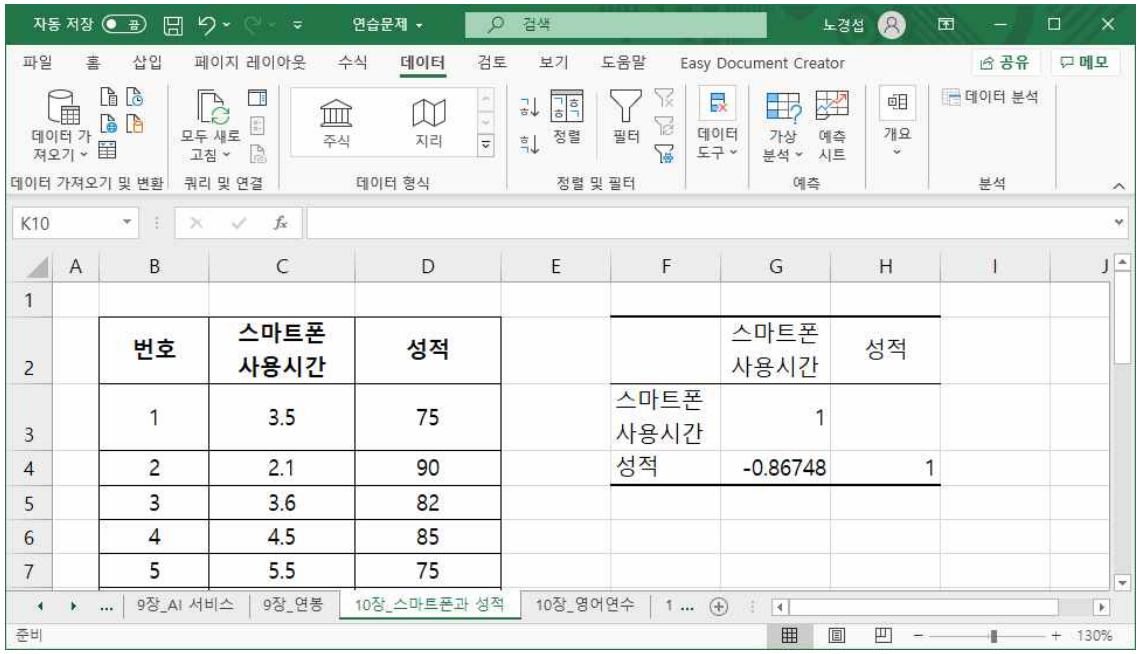
6. [컴퓨터 실습] 연습문제.xlsx의 '10장_스마트폰과 성적' 시트를 이용하여 상관계수를 구하라.

● 직접 계산



번호	스마트폰 사용시간	성적	스마트폰 사용시간 편차	성적 편차	스마트폰 사용시간 X 성적 편차
1	3.5	75	-0.20	-5.80	1.18
2	2.1	90	-1.60	9.20	-14.76
3	3.6	82	-0.10	1.20	-0.12
4	4.5	85	0.80	4.20	3.34
5	5.5	75	1.80	-5.80	-10.42
6	1.1	95	-2.60	14.20	-36.98
7	5.5	68	1.80	-12.80	-22.99
8	5.0	70	1.30	-10.80	-14.00
9	2.6	85	-1.10	4.20	-4.64
10	3.5	75	-0.20	-5.80	1.18
11	2.1	90	-1.60	9.20	-14.76
12	3.6	82	-0.10	1.20	-0.12
13	4.5	85	0.80	4.20	3.34
14	5.5	75	1.80	-5.80	-10.42
15	1.1	95	-2.60	14.20	-36.98
16	5.5	68	1.80	-12.80	-22.99
17	5.0	70	1.30	-10.80	-14.00
18	2.6	85	-1.10	4.20	-4.64
19	3.5	75	-0.20	-5.80	1.18
20	2.1	90	-1.60	9.20	-14.76
21	3.6	82	-0.10	1.20	-0.12
22	4.5	85	0.80	4.20	3.34
23	5.5	75	1.80	-5.80	-10.42
24	1.1	95	-2.60	14.20	-36.98
25	5.5	68	1.80	-12.80	-22.99
평균	3.70	80.80	편차곱의 총합		-279.48
표준편차	1.52	8.85	공분산=(편차곱의 총합)/25		-11.18
상관계수=공분산/(스마트폰 사용시간의 표준편차 X 성적의 표준편차)					-0.83

● 데이터 분석 도구 활용



번호	스마트폰 사용시간	성적
1	3.5	75
2	2.1	90
3	3.6	82
4	4.5	85
5	5.5	75

	스마트폰 사용시간	성적
스마트폰 사용시간	1	
성적	-0.86748	1

7. 교차분석과 카이제곱 검정의 관계를 설명하라.

교차분석(cross-tabulation analysis)이란 범주형으로 구성된 자료들 간의 연관관계를 확인하기 위하여 서로 교차표를 만들어 관계를 확인하는 분석방법을 말한다. 행과 열에 범주형 변수를 구분하여 서로 연관성이 있는 빈도를 확인하는 교차표를 만들기 때문에 교차분석이라고 하며, 변수들의 빈도를 이용하여 상호연관성에 대한 판단을 할 때 검정통계량으로 χ^2의 통계량을 이용하기 때문에 카이제곱(χ^2)검정이라 한다.

8. 교차표에 나타나는 빈도에 대해 설명하고, 이때 활용하는 카이제곱 통계량에 대해 설명하라.

교차표(cross-tabulation)란 조사된 요인 2가지에 대해 가로와 세로로 배열하여 교차되는 항목에 대한 빈도를 나타낸 표를 말한다. 이 때 2가지 요인에 대해 직접 수집한 데이터를 기준으로 빈도를 입력해야 하는데, 이를 관측빈도(observed frequency)라 한다.

기대빈도(expected frequency)는 전체빈도 n에 대하여 행과 열의 합을 기준으로 보았을 때, 각 교차되는 셀에 몇 번의 빈도가 확인될 수 있을지를 예상하여 기대하는 기대값이다. 첫 번째 요인의 1에서의 관측빈도 O_{11}은 n_{11}이지만 첫 번째 요인의 1이면서 두 번째 요인의 1에 대한 기대빈도 E_{11}의 기대빈도를 계산한다면, 두 번째 요인의 1에 대한 합계가 n_{i1}이면서 첫 번째 요인 1에 포함되어야 하므로 각각 곱해서 전체의 수로 나누어주면 기대빈도가 계산된다.

$$\text{대빈도} E_{ij} = \frac{n_i \times n_j}{n}$$

카이제곱 통계량이란 관측빈도와 기대빈도 사이에 유의한 차이가 있는지를 확인하는 통계량을 의미하며, 교차표의 모든 셀에서 발생하는 차이를 더해주어야 한다. 이때 계산 결과가 음수로 나오게 된다면 각 셀을 상쇄하는 효과가 나타나므로 제곱을 해서 더해주어야 한다.

9. 교차분석에서 적합도 검정을 하는 이유를 설명하라.

표본으로부터 양자택일의 빈도를 조사할 때, 조사를 마치기 전까지는 결과를 그 누구도 예상할 수 없다. 만약 선호의 차이가 99% : 1%처럼 분명하다면 결과를 예상할 수 있으므로 굳이 조사를 할 필요성이 없지만, 기초정보가 없다면 양자택일의 경우에서는 기대빈도를 50%:50%로 예상할 수뿐이 없으며, 기대빈도와 관측빈도가 차이가 적으면 적을수록 적합한 기대라 할 수 있다. 실제로 조사를 실시하여 68%:32%로 나왔는데, 이러한 차이가 적합한 것인지를 판단하는 검정방법이 **적합도 검정(goodness of fit test)**이라 한다.

10. 대학에 입학한 A군에게 폴더블 스마트폰과 롤러블 스마트폰 중 하나를 선물하려 한다. 최근 바 타입에서 새로운 폼펙터로 유행이 옮겨 갔다고 판단했기에 어떤 것이 더 유익할지를 주위 사람들에게 물어서 조사하였다. 그 결과 다음 빈도표를 얻었을 때, 95% 수준으로 카이제곱 검정을 하라.

스마트폰	폴더블 스마트폰	롤러블 스마트폰	합계
	9	6	15

H_0 : 더블스마트폰에 대한선호도 = 롤러블스마트폰에 대한선호도
H_1 : 폴더블스마트폰에 대한선호도 ≠ 롤러블스마트폰에 대한선호도

기대빈도 15÷2=7.5, $d.f = 2-1 = 1$, 폴더블스마트폰, 롤러블스마트폰$\Rightarrow k = 2$

$$\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(9-7.5)^2}{7.5} + \frac{(6-7.5)^2}{7.5} = 0.3+0.3 = 0.6$$

χ^2분포표에서 $\alpha = 0.01, d.f = 1$의 임계치는 3.8414588이므로 0.6보다 크다. 기각역에 속하지 않아 귀무가설을 채택하므로 폴더블 스마트폰과 롤러블 스마트폰의 선호도는 같다고 판단한다.

11. 교차분석에서 독립성 검정을 하는 이유를 설명하라.

적합도 검정의 예에서는 단순화하기 위해 한가지의 범주만으로 계산했으나, 일반적으로 한가지만으로 검정을 하는 경우는 거의 없다. **독립성 검정(independence test)**이란 여러 가지 범주를 대상으로 각 범주들이 독립적인지를 판단하는 검정방법이다. 예를 들어, [표 10-9]에서는 지역을 1과 2로 구분하고 구매의사가 있음과 없음의 두 가지로 구분하여 2×2교차표를 구성하면 총 네 가지 범주로 구성되어 있다. 이 범주들이 서로 독립이라는 귀무가설을 설정하여 검정하는 검정방법을 말한다.

12. [컴퓨터 실습] 12개월 동안 영어를 배우기 위해 미주 지역이나 영어를 사용하는 동남아시아 지역으로 연수를 가는 학생들을 대상으로 조사를 진행하고자 한다. 연수 프로그램은 12개월 중 6개월은 한국에서 준비하다 6개월을 현지에서 머무는 프로그램과, 현지에서 12개월 모두를 보내는 프로그램으로 구분되어 있다. 연습문제.xlsx의 '10장_영어연수' 시트를 이용하여 독립성 검정을 위해 95% 수준으로 카이제곱 검정을 하라.

H_0 : 어권과 공부방법은 독립적이다.
H_1 : 영어권과 공부방법은 독립적이지 않다.

독립성 검정의 $d.f = (R-1)(C-1) = (2-1)(2-1) = 1$

미주지역*미리공부의 기대빈도 : $(106 \times 134)/200 = 71$

동남아*미리공부의 기대빈도 : $(106 \times 66)/200 = 35$

미주지역*현지공부의 기대빈도 : $(94 \times 134)/200 = 63$

동남아*현지공부의 기대빈도 : $(66 \times 94)/200 = 31$

구 분		공부 방법		행의 합계
		미리공부, 6개월	현지에서 12개월	
영어권	미주지역 기대빈도	85 71	49 63	134
	동남아 기대빈도	21 35	45 31	66
합계		106	94	200

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(85-71)^2}{71} + \frac{(49-63)^2}{63} + \frac{(21-35)^2}{35} + \frac{(45-31)^2}{31}$$
$$= 2.76 + 3.11 + 5.6 + 6.32$$
$$= 17.79$$

χ^2분포표에서 $\alpha = 0.01, d.f = 1$의 임계치 $3.841459 < 17.79$

영어권과 공부방법은 서로 독립적이라는 귀무가설을 기각하며, 영어권에 따라 공부방법의 차이가 있다는 대립가설을 채택한다.

Chapter 11 연습문제

1. 회귀분석의 개념을 상관분석과 비교하여 설명하고, 단순회귀분석을 하는 이유에 대해 설명하라.

두 개의 변수 X와 Y의 관계에 대한 분석은 상관관계분석으로 했으며, 양이나 음의 상관 혹은 상관없음의 3가지 형태로 구분했다. 회귀분석은 독립변수를 원인으로 한 종속변수의 변화에 대한 인과관계를 파악할 수 있도록 한다. **단순회귀분석(simple regression analysis)**은 하나의 변수가 다른 변수에 미치는 영향을 회귀식(방정식)인 X와 Y로 구성하고, X를 통해 Y의 인과관계에 대한 정보를 찾아가는 분석법이다. 단순회귀분석으로 X와 Y는 회귀식이라는 수학적인 방정식으로 표현하고, 회귀식을 이용하여 변수 X를 원인으로 Y가 어떤 관계가 있는지 추정하는 것을 목적으로 한다.

2. 모수를 정확하게 추정할 수 있도록 하는 최소자승법에 대해 설명하라.

표본으로부터 도출한 회귀직선은 모회귀직선에 최대한 가깝게 추정되어야 하지만, 최대한 이상적인 회귀직선을 계산해 낸다 하더라도 잔차가 발생하게 될 수 밖에 없으므로, 최소자승법에서는 잔차를 제곱하여 모두 더한 제곱합(sum of squares)을 최소가 되도록 하는 함수를 구하는 것이다.

잔차의 제곱합을 나타내는 $\sum\hat{\epsilon_i^2}$을 최소로 하는 방법을 최소자승법(method of least squares) 혹은 최소제곱법(method of least squares)이라 한다. 즉 확률분포곡선을 기준으로 측정치를 확인하며 오차제곱의 합이 최소가 되도록 하는 확률을 계산하는 것이다.

3. 최대우도법을 최소자승법과 비교하여 설명하라.

최소자승법은 측정치에서의 오차제곱합을 최소로 하는 회귀선을 계산하지만, 최대우도법(maximum likelihood method)은 어떤 하나의 함수가 최대의 모수를 포함하는 함수로 접근되도록 하는 방법이다.

최대우도법에서의 우도(likehood)는 회귀(방정)식이 측정치를 가장 잘 나타낼 수 있도록 그 가능성(우도함수)을 최대로 끌어 올리는 개념이다. 가능성은 확률로 생각할 수 있다. 그러나 우리가 Chapter 04에서 학습한 확률은 정해진 것이었고, 이 부분에서 가능성이 확률과 다른 점은 우도함수가 측정치를 최대한 포함하도록 변화를 주면서 최적의 회귀방정식을 찾아 간다는 것이다.

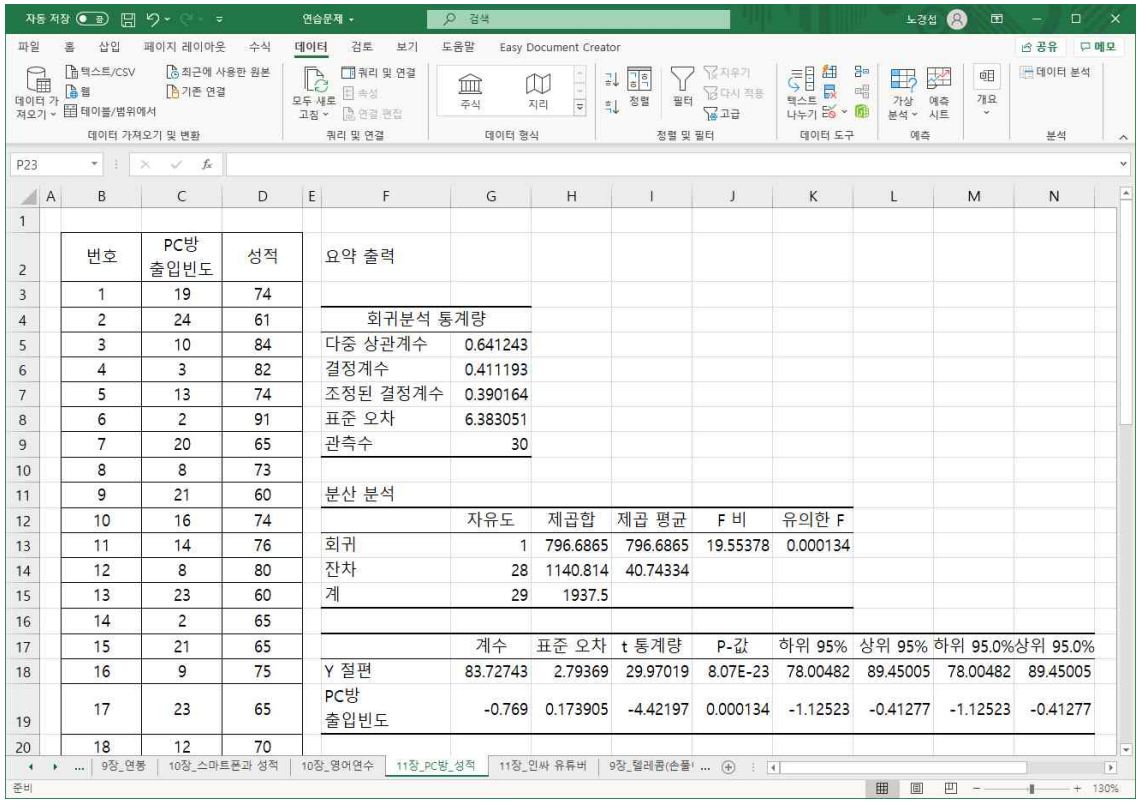
최소자승법(method of least squares)은 확률분포곡선을 기준으로 측정치에 대한 확인을 통해 오차를 최소가 되도록 확률을 계산한다.

최대우도법(maximum likelihood method)은 측정치를 기준으로 확률분포곡선으로 측정치를 가장 잘 포함하도록 함수를 최대한 우도화(우도, likehood)하면서 확률을 찾아 간다.

오차를 최소로 회귀식을 찾는 최소자승법에서와 같이 최대우도법 역시 확률표본에 가장

잘 맞는(최대한 우도화하는) 우도함수를 찾기 때문에 결국 최소자승법과 최대우도법의 추정치는 같게 된다.

4. [컴퓨터 실습] 한 달 간 PC방에 가는 빈도가 많을수록 성적에 어느 정도의 영향을 미치는지 평균 점수를 같이 조사하였다. 연습문제.xlsx의 '11장_PC방_성적' 시트를 이용하여 친절도가 재구매에 어떤 영향을 미치는지 회귀분석을 실시하라.



번호	PC방 출입빈도	성적
1	19	74
2	24	61
3	10	84
4	3	82
5	13	74
6	2	91
7	20	65
8	8	73
9	21	60
10	16	74
11	14	76
12	8	80
13	23	60
14	2	65
15	21	65
16	9	75
17	23	65
18	12	70

요약 출력

회귀분석 통계량	
다중 상관계수	0.641243
결정계수	0.411193
조정된 결정계수	0.390164
표준 오차	6.383051
관측수	30

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	796.6865	796.6865	19.55378	0.000134
잔차	28	1140.814	40.74334		
계	29	1937.5			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	83.72743	2.79369	29.97019	8.07E-23	78.00482	89.45005	78.00482	89.45005
PC방 출입빈도	-0.769	0.173905	-4.42197	0.000134	-1.12523	-0.41277	-1.12523	-0.41277

절편 → 83.72743
PC방 출입 빈도에 대한 계수 → −0.769

그러므로 회귀식은 ' 적 = 절편 + 기울기 • PC방출입빈도'는
$Y = 83.72743 - 0.769 \bullet X$로 도출되었다.

5. 회귀분석에서 적합도 검정을 하는 이유와 분산분석을 실시하는 이유를 설명하라.

적합도 검정(goodness−of−fit test)이란 도출한 회귀선 $Y = \beta_0 + \beta_1 \bullet X_1$가 표본 측정치를 어느 정도 잘 설명하고 있는지를 확인하는 것이다. 회귀선이 측정 자료와 100% 일치하면 모든 측정치가 회귀선 위에 나타나게 되며, 어느 정도 선형의 관계가 있다면 0보다는 크고 100%보다는 작은 설명력, 전혀 선형관계가 나타나지 않으면 0%의 설명력을 가지게 될 것이다. 추정된 회귀선이 어느 정도 측정치들과 일치하는 지의 정도를 표본에 대한 회귀선의 설명력(R^2)이라 표현하며, 0~1까지의 숫자로 나타내고 몇 %의 설명력을 가진다는 확률로 표현한다.

회귀분석에서 분산분석을 사용하는 이유는 총편차를 분해하는 과정이 분산분석과 동일하기 때문이다. 각 제곱합 SST, SSR, SSE를 구하고 나면 분산분석에서의 평균제곱을 구하기 위해 각각의 자유도를 알아야 한다.

SST 자유도 : $n-1$, SSR의 자유도 : 1, SSE의 자유도 : $n-2$ 이며, 분산비율 F값을 알기 위해서 총제곱합의 구성부분인 오차제곱합과 회귀제곱합을 각각의 자유도로 나누면 평균오차제곱(mean square of error : MSE) $= \frac{SSE}{n-2}$, 평균회귀제곱(mean square of regression : MSR) $= \frac{SSR}{1}$ 이다. 회귀와 잔차에 대한 분산비율 $F = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE}$로 계산한다.

6. [컴퓨터 실습] 연습문제.xlsx의 '11장_PC방_성적' 시트를 이용하여 결정계수와 분산비율을 구하라.

β_0, β_1를 계산하기 위하여 $\beta_0 = \overline{Y} - \beta_1 \overline{X}$, $\beta_1 = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}$ 에 대입해보면,

x_i의 평균 $\overline{x} = 14.6$ 이므로

$$\beta_1 = \frac{[(19-14.6)(74-72.5)+(24-14.6)(61-72.5)+\cdots+(14-14.6)(82-72.5)]}{[(19-14.6)^2+(24-14.6)^2+\cdots+(14-14.6)^2]} = -0.769$$

$$\beta_0 = 72.5 - 0.769 \bullet 14.6 = 83.72743$$

회귀식 $\hat{Y} = 83.72743 - 0.769 \times X_i$

총제곱합(SST) : $\sum(Y_i - \overline{Y})^2 = (74-72.5)^2+(61-72.5)^2+\cdots+(82-72.5)^2 = 1{,}937.5$

회귀제곱합(SSR) :

$\sum(Y_i - \hat{Y}_i)^2 = (74-69.116)^2+(61-65.271)^2+\cdots+(82-72.961)^2 \fallingdotseq 1{,}140.814$

잔차제곱합(SSE) :

$\sum(\hat{Y}_i - \overline{Y})^2 = (69.116-72.5)^2+(65.271-72.5)^2+\cdots+(72.961-72.5)^2 \fallingdotseq 796.682$

$\therefore R^2 = \frac{SSE}{SST} = \frac{769.682}{1{,}937.5} \fallingdotseq 0.411 \Rightarrow$ 회귀식은 는 41.112%의 설명력을 가지고 있다.

분산비(F)는 $\frac{SSR/1}{SSE/(n-2)}$이므로 $\frac{SSR/1}{SSE/(n-2)} = \frac{1140.814/1}{796.682/28} \fallingdotseq 40.095$ 로 확인되었다.

7. 회귀분석을 진행할 때 분산분석표를 작성하는 이유는 무엇인지 설명하라.

회귀식, R^2, 분산비 F를 도출하는 분산분석표를 작성한 이유는 귀식 $\hat{Y} = 83.72743 - 0.769 \times X_i$에 대하여 F 검정을 통해 회귀식의 유의성을 검정해야 하기 때문이다. 회귀식의 유의성을 검정하기 위해 회귀와 잔차에 대한 분산비율 F값을 계산해서 F분포표의 값보다 더 크다면 귀무가설을 기각하고 대립가설을 채택한다. 분산비율 F는 평균회귀제곱(MSR)을 평균오차제곱(MSE)으로 나눈 값으로 표준오차보다 회귀식으로 설명되는 부분이 어느 정도 더 많은지를 나타내는 수치다. 때문에 일반적인 분산의 분포를 나타내는 F값보다 더 높다는 의미는 회귀식으로 설명할 수 있는 부분이 더 많다는 의미가 되므로 회귀식이 유의하다는 판단을 내릴 수 있는 근거가 된다.

8. [컴퓨터 실습] [연습문제 6]에서 계산한 분산분석 결과를 이용하여 회귀식이 유의한지 $\alpha = 0.05$에서 검정하라.

H_0 : 회귀식이 유의하지 않다. H_1 : 회귀식이 유의하다.

분산비 F=40.095 이므로, F검정을 실시한 임계치보다 크다면 귀무가설을 기각하게 된다. F분포표에서 $F_{1,28} = 4.20$이므로, 귀무가설을 기각하고 대립가설을 채택한다.

회귀식은 유의하다.

9. [컴퓨터 실습] [연습문제 8]의 결과에서 회귀식의 계수에 대한 유의성을 판단하라.

β_0 , β_1의 유의성을 검정하기 위하여 β_0 , β_1에 대한 가설을 먼저 수립해야 하고, 가설은 $H_0 : \beta_0 = 0$, $H_1 : \beta_0 \neq 0$ 과 $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ 이 되어야 한다.

$\beta_0 = 83.72743$, $\beta_1 = 0.769$, t_{n-2}의 검정통계량$= \frac{\hat{\beta_0}}{s_{\hat{\beta_0}}}$과 $\frac{\hat{\beta_1}}{s_{\hat{\beta_1}}}$ 이므로

$s_{\hat{\beta_0}} = \sqrt{(\frac{1}{n} + \frac{\overline{X^2}}{\sum(X_i - \overline{X})^2})MSE}$ 와 $s_{\hat{\beta_1}} = \sqrt{MSE / \sum(X_i - \overline{X})^2}$ 을 이용하여 각각 대입한다. MSE는 잔차의 평균제곱인 28.453 이고, $\sum(X_i - \overline{X})^2 = 1347.2$이므로

$$s_{\hat{\beta_0}} = \sqrt{(\frac{1}{n} + \frac{\overline{X^2}}{\sum(X_i - \overline{X})^2})MSE} = \sqrt{(\frac{1}{30} + \frac{14.6^2}{1347.2})28.453} = \sqrt{4.535} \fallingdotseq 2.130$$

$$s_{\hat{\beta_1}} = \sqrt{MSE / \sum(X_i - \overline{X})^2} = \sqrt{\frac{28.453}{1347.2}} = 0.021$$

$\frac{\hat{\beta_0}}{s_{\hat{\beta_0}}} = \frac{83.72743}{2.130} = 39.309$, $|\pm t_{(\alpha/2,\ 0.025)}| = 2.048$ 이므로,

$\therefore$ 귀무가설을 기각하고 대립가설을 채택한다.

$\frac{\hat{\beta_1}}{s_{\hat{\beta_1}}} = \frac{-0.769}{0.021} = -36.619$, $|\pm t_{(\alpha/2,\ 0.025)}| = 2.048$ 이므로,

귀무가설을 기각하고 대립가설을 채택한다.

10. [컴퓨터 실습] [연습문제 4]에서 회귀식을 구성하는 계수의 95%의 신뢰구간을 구하라.

$\hat{\beta_0}$과 $\hat{\beta_1}$의 신뢰구간을 구하기 위하여, $\hat{Y} = 83.72743 - 0.769 \times X_i$ 를 이용하여

β_0 신뢰구간 : $\hat{\beta_0} - t_{(\alpha/2,\, n-2)} \cdot s_{\hat{\beta_0}} < \beta_0 < t_{(\alpha/2,\, n-2)} \cdot s_{\hat{\beta_0}}$ 에 대입하면

β_1의 신뢰구간 : $\hat{\beta_1} - t_{(\alpha/2,\, n-2)} \cdot s_{\hat{\beta_1}} < \beta_1 < t_{(\alpha/2,\, n-2)} \cdot s_{\hat{\beta_1}}$

$n = 30,\ t_{(0.025,\, 28)} = 2.048,\ MSE = \frac{796.682}{28} = 28.453$ 이므로

$\hat{\beta_0}$의 신뢰구간은 $83.72743 - 2.048 \times 2.130 < \beta_0 < 83.72743 + 2.048 \times 2.130$ 이므로

$79.365 < \beta_0 < 88.090$

$\hat{\beta_1}$의 신뢰구간은 $-0.769 - 2.048 \times 0.021 < \beta_1 - 0.769 + 2.048 \times 0.021$ 이므로

$-0.812 < \beta_1 \leftarrow 0.726$

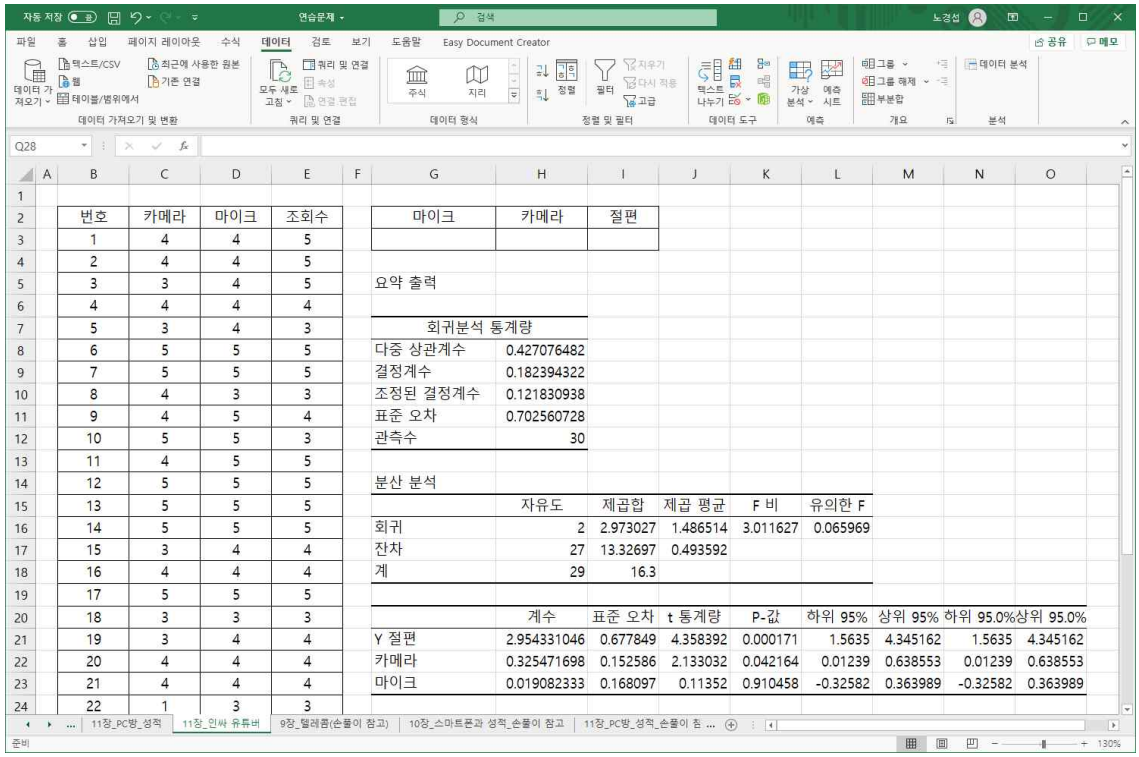
11. 단순회귀분석과 다중회귀분석의 차이점을 설명하고, 각각의 장단점에 대해 설명하라.

단순회귀분석은 광고비가 매출액에 영향을 미치는 것과 같은 사회현상을 분석할 때와 같이 단 하나의 독립변수가 종속변수에 영향이 있는지를 확인하는 방법이다. 회귀분석은 과거의 일정시점으로부터 현재까지의 자료를 바탕으로 미래를 예측할 수 있는 근거를 제시해주는데, 매출액에 단 하나의 변수인 광고비만 영향을 미친다고 단정할 수는 없다. 그래서 단일변수보다는 다중의 변수를 사용하는 회귀분석을 실시하는데 이를 다중회귀분석(multiple regression analysis)이라 한다.

단순회귀분석은 절대적으로 중요한 영향요소 하나만을 표현하는데 모든 상황을 단순화하여 표현하는데 유리하지만, 일반적인 현상을 설명하는 것에는 부족하다. 다중회귀분석은 일반적인 사회현상을 단순회귀분석보다 논리적으로 더 많은 변수로 설명하므로 이해하기는 편하지만 변수가 많아지면 계산하기가 복잡해지는 단점이 있으며, 조사자의 입장에서는 어떤 사회적 현상을 정확하게 설명하겠다는 압박을 느낄 수도 있기에 조사단계에서의 적절한 조사(연구)모형을 수립하는 것이 중요한다.

12. [컴퓨터 실습] 인싸 유튜버가 되기 위해 장비를 구입하고자 한다. 연습문제.xlsx의 '11장_인싸 유튜버' 시트를 보면 유튜브 조회수가 올라가는 정도를 1~5로 표시하였고, 카메라

와 마이크를 중요도에 따라 1~5로 구분하여 데이터를 수집했다. 독립변수를 카메라와 마이크로 하고 종속변수를 유튜브 조회수로 하는 회귀 모델에서 각각의 독립변수가 종속변수에 미치는 영향에 대한 계수를 계산하여 회귀식을 구하라.



번호	카메라	마이크	조회수
1	4	4	5
2	4	4	5
3	3	4	5
4	4	4	4
5	3	4	3
6	5	5	5
7	5	5	5
8	4	3	3
9	4	5	4
10	5	5	3
11	4	5	5
12	5	5	5
13	5	5	5
14	5	5	5
15	3	4	4
16	4	4	4
17	5	5	5
18	3	3	3
19	3	4	4
20	4	4	4
21	4	4	4
22	1	3	3

마이크	카메라	절편

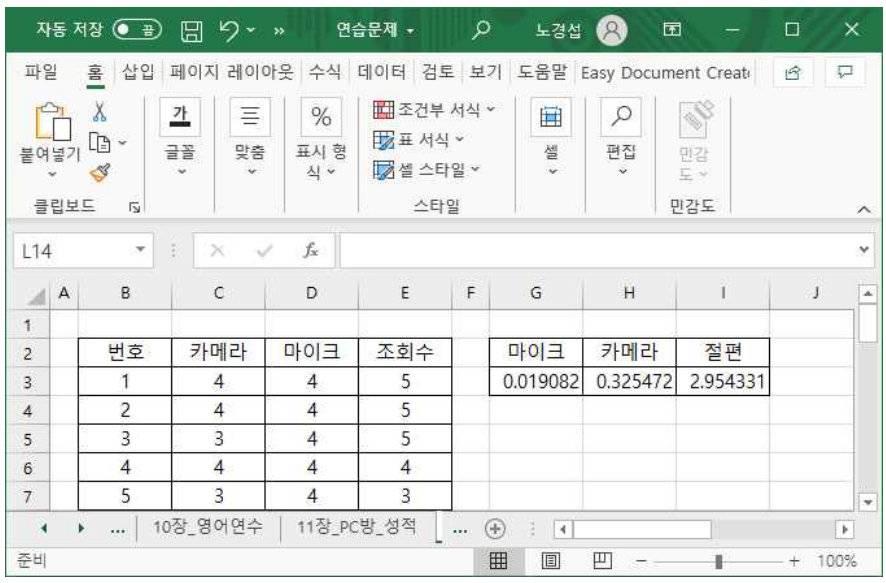
요약 출력

회귀분석 통계량	
다중 상관계수	0.427076482
결정계수	0.182394322
조정된 결정계수	0.121830938
표준 오차	0.702560728
관측수	30

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	2	2.973027	1.486514	3.011627	0.065969
잔차	27	13.32697	0.493592		
계	29	16.3			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	2.954331046	0.677849	4.358392	0.000171	1.5635	4.345162	1.5635	4.345162
카메라	0.325471698	0.152586	2.133032	0.042164	0.01239	0.638553	0.01239	0.638553
마이크	0.019082333	0.168097	0.11352	0.910458	-0.32582	0.363989	-0.32582	0.363989



번호	카메라	마이크	조회수
1	4	4	5
2	4	4	5
3	3	4	5
4	4	4	4
5	3	4	3

마이크	카메라	절편
0.019082	0.325472	2.954331

함수 '=LINEST(E3:E32,C3:D32)'을 이용하면

$$\hat{Y}=\hat{\beta}_n X_n+\hat{\beta}_{n-1}X_{n-1}+\hat{\beta}_{n-2}X_{n-2}+\cdots+\hat{\beta}_1 X_1+\quad \text{편}$$

위와 같은 형태로 계수들이 나열된다.

13. [컴퓨터 실습] [연습문제 12]에서 적합도 검정을 위한 R^2과 수정된 R^2을 계산하고, 분산분석을 활용하여 분산비율을 구하라.

회귀식 $\hat{Y}=2.954+0.325 \bullet X_1+0.019 \bullet X_2$ 이므로

$\hat{\beta_0}=2.954,\ \hat{\beta_1}=0.325,\ \hat{\beta_2}=0.019,$

총제곱합(SST) : $\sum(Y_i-\overline{Y})^2=(5-4.3)^2+(5-4.3)^2+\cdots+(4-4.3)^2=16.3$

회귀제곱합(SSR) : $\sum(Y_i-\hat{Y_i})^2=(5-4.33)^2+(5-4.33)^2+\cdots+(4-4.33)^2 \fallingdotseq 13.327$

잔차제곱합(SSE) : $\sum(\hat{Y_i}-\overline{Y})^2=(4.33-4.3)^2+(4.33-4.3)^2+\cdots+(4.33-4.3)^2 \fallingdotseq 2.964$

$$R^2=\frac{SSR}{SST}=\frac{13.327}{16.3} \fallingdotseq 0.818$$

그러나 R^2을 불편추정량으로 변경하기 위하여 SSE SST 를 자유도로 나눈 값을 이용하여, 수정된R^2을 계산해야 한다.

수정된$R^2=1-\frac{2.964/(30-2-1)}{16.3/(30-1)} \fallingdotseq 1-\frac{0.110}{0.562}=0.804$

회귀식은 29.8%의 설명력을 가지고 있다.

분산비(F)는 $\frac{SSR/i}{SSE/(n-i-1)}$이므로 $\frac{SSR/i}{SSE/(n-i-1)}=\frac{13.327/2}{2.964/29} \fallingdotseq 65.196$ 로 확인되었다.

14. [컴퓨터 실습] [연습문제 13]의 결과에서 회귀식의 계수에 대한 유의성을 $\alpha=0.05$에서 검정하라.

회귀식의 유의성을 판단하기 위한 가설은

H_0 : 회귀식이 유의하지 않다.

H_1 : 회귀식이 유의하다.

분산비 F=65.196이므로, F검정을 실시한 임계치보다 크다면 귀무가설을 기각하게 된다. F분포표에서 $F_{2,27}=3.35$이므로, 귀무가설을 기각하고 대립가설을 채택한다.

∴ 회귀식은 유의하다.

15. [컴퓨터 실습] [연습문제 12]에서 회귀식을 구성하는 계수의 95% 신뢰구간을 구하라.

$\hat{Y}= 2.954 + 0.325 \cdot X_1 + 0.019 \cdot X_2$

$\hat{\beta_0}= 2.954$, $\hat{\beta_1}= 0.325$, $\hat{\beta_2}= 0.019$의 95%에 대한 신뢰구간을 구하기 위한 $t_{(\alpha/2\,,\,n-i-1)} = 2.052$ 이고, 표준오차 $s_{\hat{\beta_0}} = 0.140$, $s_{\hat{\beta_1}} = 0.064$, $s_{\hat{\beta_2}} = 0.071$ 이므로 이를 $\hat{\beta_i} - t_{(\alpha/2,\,n-i-1)} \cdot s_{\hat{\beta_i}} < \beta_i < t_{(\alpha/2,\,n-i-1)} \cdot s_{\hat{\beta_i}}$ 에 대입하면,

$\hat{\beta_0}= 2.954 - 2.052 \cdot 0.140 < \beta_0 < 2.954 + 2.052 \cdot 0.140 \qquad 2.667 < \beta_0 < 3.241$

$\hat{\beta_1}= 0.325 - 2.052 \cdot 0.064 < \beta_1 < 0.325 + 2.052 \cdot 0.064 \Rightarrow 0.194 < \beta_1 < 0.456$

$\hat{\beta_2}= 0.019 - 2.052 \cdot 0.017 < \beta_2 < 0.019 + 2.052 \cdot 0.017 \Rightarrow -0.127 < \beta_2 < 0.165$

이와 같이 신뢰구간을 계산할 수 있다.