2023학년도 2학기 출석과제물

|  |  |  |  |
| --- | --- | --- | --- |
|  | **교과목명** | : | 통계로세상읽기 |
|  | **학번** | : | 202234-366307 |
|  | **성명** | : | 최문성 |
|  | **연락처** | : | 010-8010-6050 |

1. 국가통계(공식통계)의 이용을 개인, 기업, 정부의 측면에서 설명하라.

국가 통계란 정부 및 정부 기관에서 특정한 목적을 가지고 활용할 수 있는 정보를 얻기위해 작성한 통계이고 통계법에서 지정한 통계로 한정된다. 공식통계 또한 국가통계와 비슷하지만 통계법상에서는 통제작성기관이 특정 목적을 가지고 직접 또는 다른 기관에 위임 및 위탁하여 작성하는 지정통계와 일반통계를 통틀어서 공식통계라고 한다. 이러한 국가통계는 개인, 기업, 정부 등 모든 국가 구성원이 합리적인 판단과 선택을 하는 데에 도움이 될 수 있는 기본 자료이다.

개인측면에서 국가 통계는 개인 생활의 합리성 추구의 판단자료로 이용된다. 예를 들어 보건질병통계[[1]](#footnote-1)를 통해서 내 나이대의 사람들이 주로 어떤 질병에 걸리는지 살펴본 후 보험에 가입할 때 고려할 수도 있고, 조용하고 교통체증을 싫어하는 성향이라면 인구통계조사의 인구밀도[[2]](#footnote-2)자료를 통해 이사 가고 싶은 후보지의 인구밀도를 비교해서 조금 더 적은 곳을 선택하는 등 통계자료를 개인의 의사결정에 활용할 수 있다.

기업의 측면에서 국가통계는 기업 전략 수립의 기본 자료로 활용할 수 있다. 시장 조사, 마케팅, 품질 관리, 고객 관리, 재무 관리, 재고 관리, 성과 평가, 미래 예측, 위기 관리, 비용 관리, 트렌드 분석, 경쟁자 분석 등등 기업에서 행해지는 모든 분석 및 의사 결정에서 통계 데이터가 기본 자료로 활용될 수 있다.

정부 또한 통계 데이터를 이용하여 현재 우리 사회가 가지고 있는 문제점이 무엇인지 분석하고 또한 앞으로 생길 문제점이 무엇인지 예측할 수 있고 이를 토대로 문제점을 해결하고 예방할 수 있는 정책을 기획, 수립하고 알맞은 예산을 배분할 수 있다.

1. 통계학의 세 가지 역할을 설명하라.

통계학의 세 가지 역할로는 자료의 수집, 자료의 요약 및 설명, 그리고 자료로부터 결론 도출이 있다. 먼저 자료의 수집은 통계 자료를 만들기 위하여 자료를 어떤 식으로 수집할 것인지, 전수조사를 할 것인지 표본조사를 할 것인지, 표본조사를 한다면 모집단으로부터 표본은 어떻게 추출할 것인지, 조사를 할 때 질문은 어떤 식으로 할 것인지 등에 관한 방법에 대한 것이다. 우리 생활과 통계 교재 3장에서 다루고 있듯이 이러한 수집 단계에서 적절한 방법을 사용하지 않으면 결과에서 편향이 발생하기 때문에 기껏 힘들게 수집한 자료로 엉뚱한 결과를 예측하여 좋지 못한 결과를 낳을 수 있기 때문에 자료의 수집이 중요하다.

그리고 만약 수집한 자료를 단순히 수집된 상태 그대로 스프레드시트나 데이터베이스의 테이블에 저장된 상태로 보여준다면, 그 숫자가 무엇을 의미하는지 이해할 수 있는 사람은 극소수일 것이다. 이렇게 수집된 자료에서 평균, 중앙값 같은 의미 있는 값을 만들어서 요약하고 막대 그래프, 히스토그램과 같은 형태로 시각화하여 사람들이 알아볼 수 있는 상태로 만들어야 자료가 무엇을 의미하는지 설명할 수 있을 것이고, 이것이 선행되어야 통계자료에 대한 이해를 토대로 의사결정에 활용할 수 있을 것이다.

마지막으로 이러한 자료를 분석하여 가설을 세우고 가설을 근거로 결론을 도출하고 미래에는 값이 어떻게 변할지 예측하는데 사용할 수 있는데, 이 과정에서 오류가 발생할 수 있다. 수집되고 분석된 자료를 통하여 어떠한 결론을 낼 수 있는지, 그 결론은 얼마나 신뢰할 수 있는지 등을 가능하게 하는 것이 통계학의 역할이다.

1. 캠핑카 판매지점의 판매량이 다음과 같은 확률분포를 가질 때 기대되는 평균 판매기대수 E(X)를 구하라.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | 0 | 1 | 2 | 3 |
| P(X) | 0.2 | 0.4 | 0.2 | 0.2 |

평균 판매기대수 E(X)는 각각의 판매량X에 대한 확률 P(X)를 곱한 값의 총합이다. 그러므로

E(X) = (0 \* 0.2) + (1 \* 0.4) + (2 \* 0.2) + (3 \* 0.2)

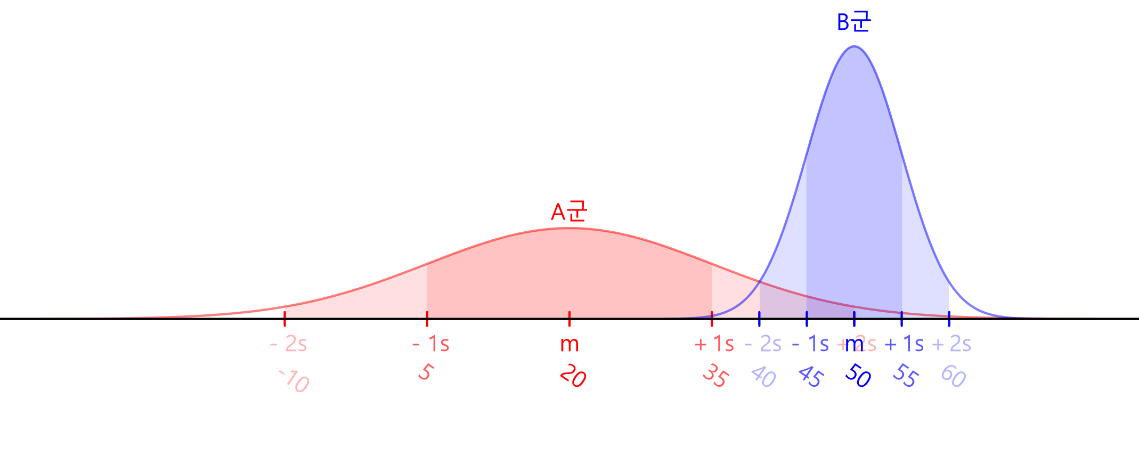
= 0 + 0.4 + 0.4 + 0.6

= 1.4

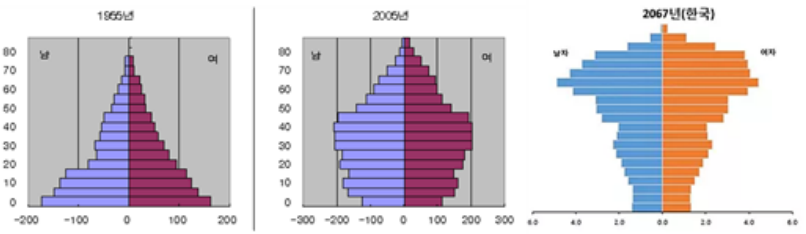
이러한 공식에 따라 평균 판매기대수 E(X)는 1.4이다.

1. 다음 자료로 정규분포의 모수에 의한 변화를 같은 X축상의 분포도 그림과 함께 설명하라.

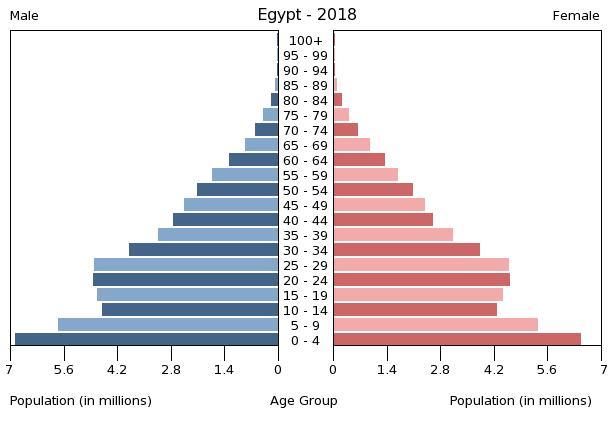
|  |  |  |
| --- | --- | --- |
|  | 평균 | 분산 |
| A군 | 20 | 15 |
| B군 | 50 | 5 |

[[3]](#footnote-3)

A군의 평균은 20이고 B군의 평균은 50이다. 그러므로 B군의 분포도는 A군보다 상대적으로 오른쪽에 위치하게 된다. 또한 A군의 분산은 15이고, B군의 분산은 5인데 이는 A군이 B군보다 산포도가 높아서 더 넓게 흩어져서 분포하고 있음을 의미한다. 따라서 A군의 분포는 넓게 퍼진 낮은 형태이고, B군은 평균 근처에 많이 분포된 형태이다. 이와 같이 모수의 변화에 따라 분포도의 형태가 달라지는 것을 그래프로 그리면 더욱 직관적으로 파악할 수 있다.

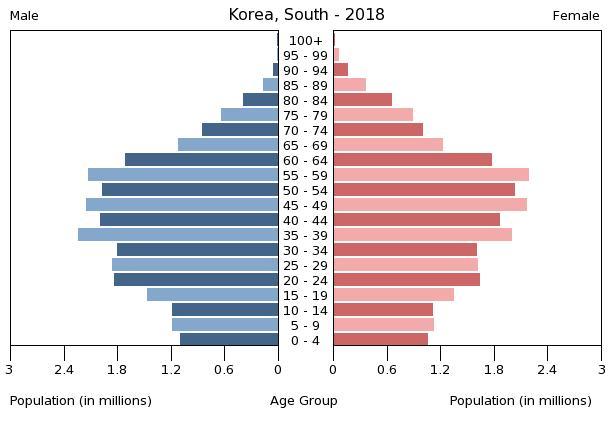
1. 다음의 인구피라미드에 대해 비교 설명하라.

위의 도표는 각각 1955년, 2005년의 인구 구성과 2067년의 예측치를 나타낸 인구피라미드 도표이다. 1955년의 도표는 고연령층의 비중이 매우 적고 아래로 내려올수록 증가하는, 특히 청소년 및 영유아의 비율이 배우 높은 모습을 보이는데 이는 평균수명이 짧고 출산율이 높은 형태로 주로 후진국에서 발견되는 모형이다. 오른쪽의 모형은 2018년 이집트의 인구구조로 1955년 한국과 비슷한 모습을 가지고 있다.



< 2018년 이집트의 인구구조 >

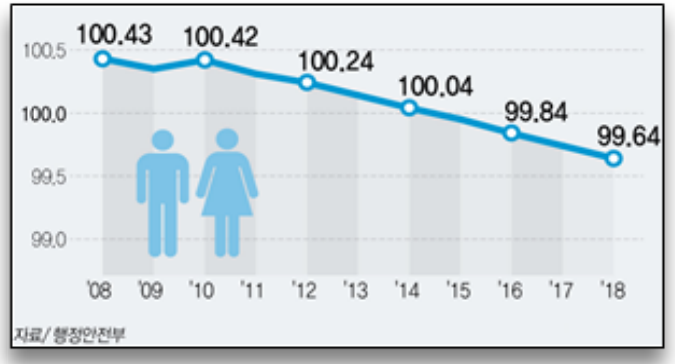
반면에 2005년의 도표는 방추형 형태를 띄고있는데 주로 선진국에서 발견되는 모형이며 평균수명이 줄어들고, 사망률이 줄어들었으나 출생률 또한 줄어들어 인구증가가 정체되고 있다. 왼쪽의 도표는 2005년의 방추형 모형이 13년간 계속 진행된 2018년의 한국인데 출생률이 갈수록 계속 저하하여 아래로 갈수록 얇아지는 모습을 볼 수 있다.



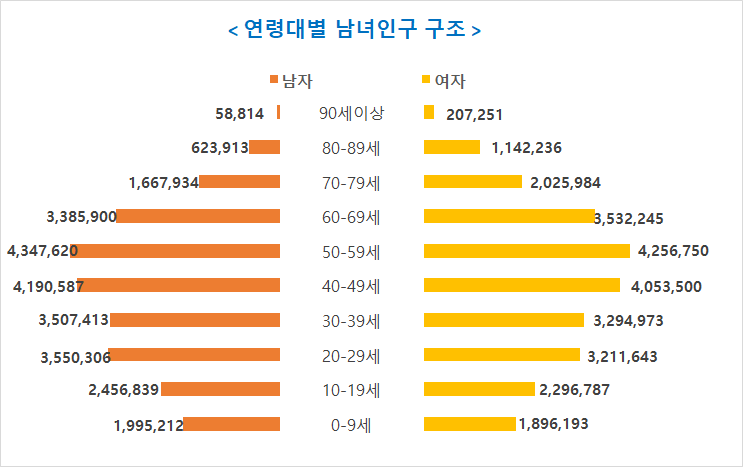
< 2018년 한국의 인구구조 >

마지막으로 2067년의 예상도표는 출산률의 반등이 전혀 없이 고령화가 계속 진행되어 역삼각형에 가까운 형태의 초고령화사회가 된 모습의 인구피라미드이다. 젊은 세대가 적으니 그 젊은 세대가 낳을 수 있는 아이의 수가 한정되어 있고 또 그 자녀 세대가 낳을 수 있는 아이도 적을 수밖에 없는 악순환이 계속 진행된 결과로 그래프가 아래로 갈수록 계속 얇아짐을 알 수 있다.

1. 출생성비의 정의와 최근 성비 변화에 대한 다음의 통계자료에 대해 설명하라.



출생성비란 여아 100명당 남아의 수를 의미한다. 왼쪽의 도표[[4]](#footnote-4)에 따르면 2000년대 이전에는 전통적인 사고방식인 남아선호사상 등의 요인으로 1990년대 최대 117.2명까지 높았다가 점점 낮아져서 2006년 이후로는 자연출생성비 105에 가까운 수치를 유지하고 있다.

반면에 전체인구에 대한 남녀성비는 2000년대까지는 약간 남초였다가 2015년, 2016년 즈음에 여성 인구가 더 많아지는 역전을 겪은 후 쭉 여초가 유지되고 있는데, 이는 한국 여성의 평균 기대수명 84.6세가 남성 평균 기대 수명 77.9세보다 약 7세가량 높음으로 인하여 60대 이상의 세대부터는 연령대가 높을수록 여초현상이 더욱 뚜렷해지는 것이 왼쪽의 도표[[5]](#footnote-5)에서 관측된다. 이렇듯 노년인구 증가 및 노년층의 여초현상이 한국 전체 인구 성비가 남초에서 여초로 전환되는 흐름에 크게 기여했음을 유추할 수 있다.

1. 통계분석 초기에 자료의 이상치를 파악하기 위한 검토 방법을 설명하라.

이상치(outlier)란 자료의 집합에서 다른 자료들이 가진 일반적인 패턴에서 벗어난 극단치를 지닌 값이거나 오류로 인해 발생한 값인 경우가 많다. 통계 자료를 정리하고 분석하기 전에 이런 이상치가 존재함을 인지하고 이러한 값을 전처리를 하여야 통계 자료 분석의 결과로 잘못된 결과를 도출하는 경우를 예방할 수 있을 것이다.

A diagram of a graph

Description automatically generated with medium confidence이상치를 파악하기 위한 대표적인 방법으로는 상자그림[[6]](#footnote-6)이 있다. 자료를 크기 순으로 나열한 뒤 하사분위값인 25%지점, 상사분위값인 75%지점을 끝으로 하는 상자를 만든 후, 특이항의 경계를 표시하고 특이항의 경계를 벗어나는 값을 이상치로 간주하는 방법이다. 이러한 방법을 사용하면 특이항의 범위를 벗어나는 값을 직관적으로 파악할 수 있어서 이상치의 존재여부를 시각화하는데에 유용하다.

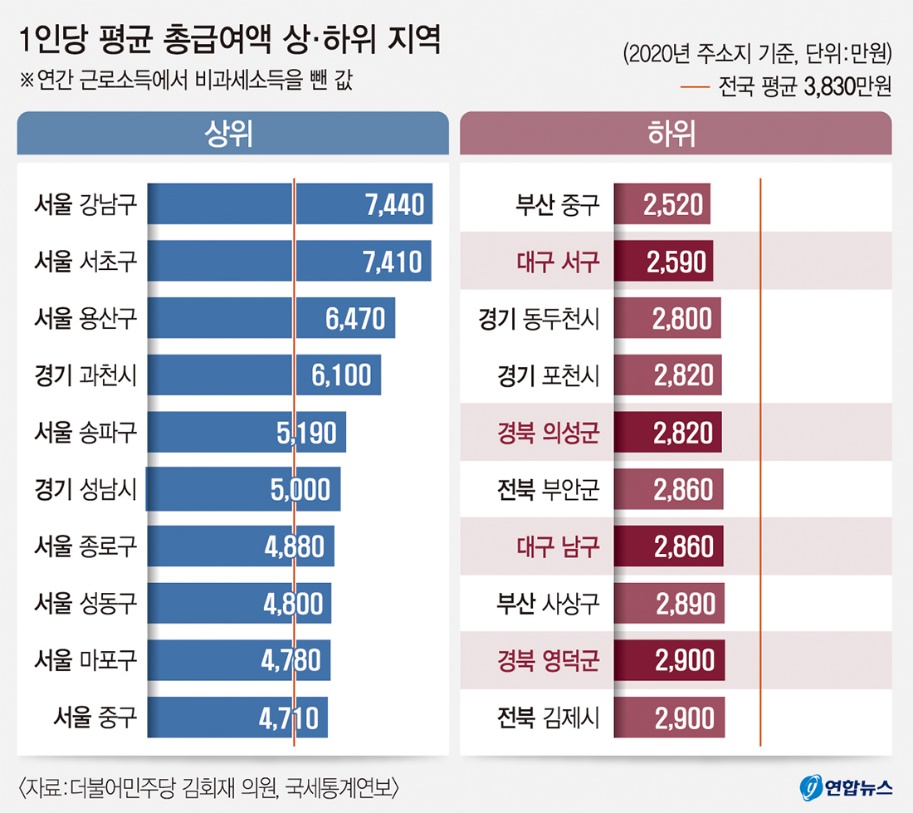
1. 모집단과 표본의 관계와 임의추출(random sampling)을 도표로 설명하라.

자료의 정리

표본 추출

분석 및

통계적 추론

모집단과 표본의 관계는 위와 같은데, 우리가 알고자 하는 정보인 모수를 가진 조사대상의 전체 집단을 모집단이라 한다. 모집단을 조사하기 위에 뽑힌 모집단의 일부를 표본이라고 한다. 이러한 표본이 모집단을 제대로 대표하려면 표본이 잘 섞여 있고 무작위로 추출되어야 한다. 예를 들어 우리나라 국민의 평균 소득을 알고 싶어 하는 상황을 가정해보자. 우리나라 국민 전체의 개인 소득을 모두 전수조사 하기는 시간과 비용이 매우 많이 필요하기 때문에 국민 중 일부 집단을 추출해내서 조사하는 것이 더욱 합리적일 것이다. 평균 소득을 조사하기 위해 추출된 일부 국민이 표본이다. 하지만 만약에 표본이 골고루 섞여서 임의로 추출되지 않고 오른쪽의 도표[[7]](#footnote-7)에서 볼 수 있듯이 소득이 높은 지역의 국민이나 소득이 낮은 직업군, 소득이 높은 연령대 등 특정한 집단이 편향되어 표본에 많이 포함되거나 적게 들어가는 일이 발생할 경우 치우침이 발생하여 왜곡된 결과가 도출될 가능성이 높다. 이러한 왜곡을 방지하고 표본이 모집단에 대한 대표성을 띈 좋은 표본이 되기 위해서는 모집단의 정의, 표본크기, 조사 방법, 조사 시기, 표촐 추출 방법, 등이 모두 고려되어야 한다.

1. 정규분포의 특징을 설명하라.

정규분포[[8]](#footnote-8)는 평균과 표준편차에 의해 특성, 즉 어떻게 분포하고 있는지 그 형태가 정해A diagram of a normal distribution

Description automatically generated지고 평균을 중심으로 좌우대칭인 종 모양인 특징을 가지고 있다. 평균(μ)과 표준편차(σ) 두 가지 값이 정규분포의 특성을 결정하기 때문에 정규분포 N은 N(μ, σ2)의 형태로 표현한다. 또한 평균에서 1표준편차 내에 약 68.27%의 자료가 존재하고, 2표준편차 내에 95.45%, 3표준편차내에 99.73%가 존재하고 3표준편차를 벗어나는 값은 극히 드문 편이다.

1. 자료의 그래프에 의한 표현의 장점과 그 종류를 설명하라.

A diagram of a spiral staircase

Description automatically generated [[9]](#footnote-9)

자료를 숫자나 표의 형태가 아닌 그래프로 시각화해서 제시하는 것에는 여러 장점이 있다. 우리가 통계 자료를 만드는 이유는 도출해낸 결과를 의사결정에 사용하기 위해서인데, 숫자나 표의 형태로 제시한다면 통계나 해당 분야에 대한 기초적인 지식이 없는 사람은 그 숫자들이 무엇을 의미하는지, 어떤 추세를 보이는 것인지 직관적으로 이해하기 힘들다. 위의 그래프는 나이팅게일이 19세기 크림 전쟁 도중에 작성한 것이다. 나이팅게일은 전쟁 도중 병원에서 일어난 모든 일을 빠짐없이 기록하고 수치화 함으로써 그 결과를 토대로 대부분의 병사는 부상이 아닌 위생문제로 사망함을 깨닫게 되고 이를 장미 다이어그램을 만들어서 많은 사람에게 위생의 중요성을 직관적으로 설파할 수 있었다. 시각화가 되지 않았다면 많은 사람들에게 이해하기 힘든 어려운 숫자들의 나열이었을 자료가 시각화를 통해 직관적으로 누구나 이해하기 쉬운 자료로 탈바꿈하는 것이다.

1. 보건의료빅데이터개방시스템, 다빈도질병 통계, <https://opendata.hira.or.kr/op/opc/olapHifrqSickInfoTab1.do> [↑](#footnote-ref-1)
2. e-나라지표, 지역별 인구 및 인구밀도, <https://www.index.go.kr/unity/potal/main/EachDtlPageDetail.do?idx_cd=1007> [↑](#footnote-ref-2)
3. 정규분포 그래프 그리기, <https://javalab.org/normal_distribution_graph/> [↑](#footnote-ref-3)
4. 갈수록 여아선호 … 출생성비 42년만에 최저, <https://stock.mk.co.kr/news/view/52509> [↑](#footnote-ref-4)
5. 2021년 4월 주민등록 기준 연령별 및 성별 인구 현황, <https://m.blog.naver.com/ljwphsjh/222344140825> [↑](#footnote-ref-5)
6. boxplot() 함수: 상자 그림 그리기, <https://thebook.io/080217/0262/> [↑](#footnote-ref-6)
7. 대구 서구 근로자 평균연봉 2590만원…농촌지역 의성보다 적어, <https://www.yeongnam.com/web/view.php?key=20220824010003002> [↑](#footnote-ref-7)
8. 확률분포(8): 정규분포, http://infoso.kr/?p=3748 [↑](#footnote-ref-8)
9. Diagram of the causes of mortality in the army in the east, https://en.m.wikipedia.org/wiki/File:Nightingale-mortality.jpg [↑](#footnote-ref-9)