확률/통계

2023

[목 차]

Part 1. 통계와 표본조사

01 통계와 표본조사

1 통계학이란?

통계학(statistics)의 어원이 라틴어에서 국가라는 의미를 갖는 'status'에서 유래되었다는 사실이나 'statistics'라는 단어가 나타나기 전 쓰였던 단어가 'political arithmetic(정치산술)'이라는 사실에서 알 수 있듯이 오랜 시간동안 통계라는 것은 한 국가의 지표로서 경제,인구, 정치 상황을 자료나 도표로 나타내는 것과 동일시 되어왔다.
'통계학'을 한자로 쓰면 '統計學'이 되는데 여기서 '統'은 거느릴 통(govern)이고 '計'는 셈할 계(device)이다. 통계학이라는 용어에도 국가를 위한 집계라는 의미가 강하게 있다고 하겠다.

우리의 일상생활에서 얻어지는 다양한 통계자료를 수집, 의미있는 결론을 이끌어내는 작업을 수행하는 학문

2 통계학(statistics)과 통계량(statistics)

- 단수로 쓰이는 statistics '통계학' 학문은 (자료를 수집, 분석, 해석, 설명, 표현하는 수리과학)을 가리킴
(표본과 모집단으로부터 나오는 수치자료를분석하고 파악하는 일을 하는 수학의 한 분야)

- 복수로서 쓰이는 statistics '통계량'은 (자료에 함수를 적용한 결과, 실제로 관찰이 이루어지기 전에 취할 수 있는 모든 값과 그에 대응하는 가능성을 총칭하는 양)이나 통계치(통계량의 실제 관측값)의 복수형이다.
(수치 자료의 총체)
- 우리가 보통 '통계'라고 말하는 단어가 바로 이 통계량이나 통계치를 가리킨다.

단수로 쓰이는 statistics는 '통계학'이라는 학문이고, 복수로서 쓰이는 statistics는 '수치자료의 총체'라고 정의하고 있다.
- **(통계량)** 표본의 특성을 보이는 특성치(特性値)

1. 모집단(population)과 표본(sample)
2. 자료의 축약방법(methods of data reduction)
3. 변동(variation)

3 모집단과 표본

용어	설명	예
모집단 (population)	문제에서 관심의 대상이 되는 모든 자료 즉 연구대상이 되는 가능한 관측값이나 측정값의 집합	전국의 유권자 전체의 자료
표본 (sample)	많은 경우 관심의 대상이 되는 모든 자료를수집하는 대신 이들을 잘 대표한다고 판단되어지는 일부만을 추출하여 조사한 자료 (통계적 처리를 위하여 모집단에서 실제로 추출한 관측 값이나 측정값의 집합)	전국의 유권자를 잘 대표하도록 추출된 일부 유권자의 자료
표본집단을 조사하여 그 특성을 찾아내고 이를 통해 모집단의 특성을 추론함. 이 과정에서 표본집단의 특성은 통계량(statistic)이라고 부르며, 이를 통해 알게된 모집단의 특성을 모수(parameter)라 함		

모집단 전체를 조사하는 것을 전수조사(census)라고 하는 데 시간이나 비용 같은 문제들 때문에 주로 국가기관에서 실시한다.
대중매체가 많이 실시하는 여론조사(opinion survey)나 기업체에서 많이 행하는 시장조사(market research, marketing survey)가 표본조사의 대표적인 예이다.

▶ 모집단 전체를 조사하지 못하고 표본조사를 하게 되는 이유

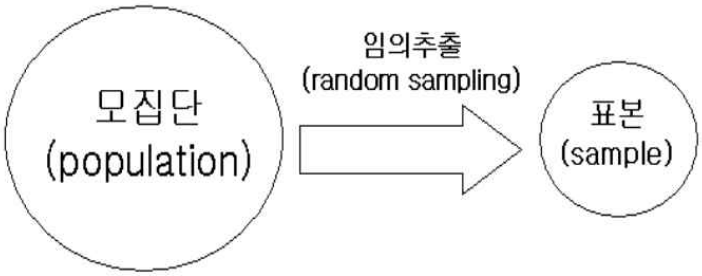
- 전수조사가 시간적 또는 경제적 여건상 불가능한 경우
- 때에 맞추어 조사결과가 제시되어야 조치가 가능한 경우
- 관심 특성치가 파괴를 해야만 얻을 수 있는 자료인 경우
- 전수조사를 함으로써 오차 개입이 커져서 정확도를 오히려 떨어뜨리는 경우

▶ 표본설계의 기본 블록

모집단을 잘 대표할 수 있는 표본을 뽑는 방법은 매우 많다. 그러나 기본적으로 임의 추출법(random sampling)은 많은 방법의 기본 블록을 형성한다. 간단하게 말하면 임의추출법이란 모집단의 구성요소 하나하나가 표본으로 뽑힐 확률이 같게끔 표본을 뽑는방법이다. 따라서 이러한 방법에 의해 만들어진 표본은 모집단을 잘 대표할 수 있는 근거가되는 것이다. 다음은 임의추출법을 적용한 예이다

- 복권추첨 시 대상번호를 모두 통속에 넣어 골고루 섞어서 어떤 번호나 뽑힐 확률이 같은 상황에서 번호를 뽑는 경우
- 선거인명부를 놓고 일련번호를 매긴 뒤 난수표를 이용해서 일부 사람들을 뽑아 전화로 지지후보에 대해서 조사하는 경우
- 100명의 회원 중 5명을 뽑아 오페라 입장권을 준다고 할 때 100명의 이름이나 번호를 카드에 적어서 통속에 놓고 골고루 섞은 후 눈을 감고 5명을 뽑는 경우

임의추출법(random sampling) : 모집단의 구성요소 하나하나가 표본으로 뽑힐 확률이 같은 상황에서 표본을 뽑는 방법



▶ **표본이 모집단을 대표할수 있을까? (표본조사가 타당성이 있을까?)**

Glinko-Cantelli 정리:	모집단의 일부분인 표본을 이용하여 표본의 정보를 얻은 후 이를 확대하여 모집단의 정보로 여기는 것에 대한 타당성 이론 **'모집단에서 iid(independent and identically distributed) 성질을 만족하게 표본을 뽑으면 경험적 누적분포함수(empirical cumulative distribution function)는 표본의 수가 커짐에 따라 점점 이론적 누적분포함수(theoretical cumulative distribution function)에 다가간다.'**
- 요약정리 -	모집단에서 표본을 골고루 섞어 잘 뽑으면 표본의 수가 커짐에 따라 점점 표본은 모집단을 닮아간다.'

***사례1**: 2007년 대통령 선거개표 방송 3사(KBS, MBC, SBS)가 실시한 출구조사결과 실제 개표결과 예측함 MBC와 KBS는 전국 250개 투표구에서 유권자 7만 명에게 출구조사를 했고, SBS는 투표를 마친 유권자 10만여 명을 대상으로 단독 출구조사를 했다.

(모 집 단): 총 투표자수 23,732,854명
(표본집단): KBS와 MBC 공동출구조사는 총투표자수의 0.3%, SBS 출구조사는 총투표자수의 0.4%

***사례2:** 세계 최대의 독립 PR컨설팅사인 에델만의 한국지사, 에델만코리아 (www.edelman.co.kr, 사장 김원규)는 2007년 9월 12일'2007 한국 블로거 성향 조사'라는 이름의 보고서를 발표했다. 에델만 코리아가 한국과학기술원(KAIST) 바이오 및 뇌공학과 정재승 교수팀과 공동으로 진행한 조사는 2006년 12월부터 2007년 2월까지 총 59일 동안 온라인 설문 방식으로 실시됐으며, 블로그와 미니홈피를 운영하고 있는 국내 블로거 총 347명이 참여했다. 조사의 표본 오차는 95% 신뢰수준에서 ±4.3% 포인트다.

(모 집 단) : 우리나라 블로거 전체
(표본집단) : 블로거 347명

- 정리 -

통조림을 판매하는 회사의 대표이며, 통조림에 이물질이 들어가지 않았는지 위생 조사를 실시해야 한다고 가정했을 때, 전수조사를 시행한다면 모든 통조림을 개봉해야 하고, 정작 판매할 통조림을 남아있지 않게 된다. 이런 상황에서 전수 조사를 실시하는 것은 사실상 불가능하다. 모든 상품을 조사하는 전수조사 방법이 아닌 무작위로 일부 상품을 골라 그것들을 대상으로 조사하는 표본조사 방식으로 데이터를 얻어야 한다. 이렇게 얻은 데이터를 이용해 모든 상품의 일반적인 가치를 추청하게 되지만 아무리 정교하게 표본조사를 하더라도 전수조사가 아닌 이상 오차가 발생할 수밖에 없다.
통계학에서는 이렇게 표본 조사로 얻은 데이터에서 허용하는 오차를 보통 5%로 지정하는데, 이를 유의수준 5% 또는 신뢰수준 95%라고 정의한다.

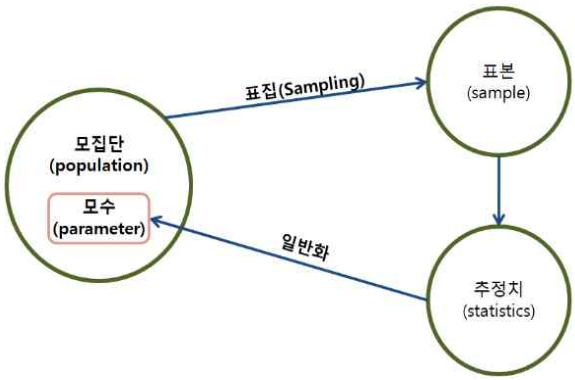
<참고>

▶기술 통계(descriptive statistics)

- 자료의 상태를 있는 그대로 설명하고 기술하는 방식의 통계
- 그래프나 표를 이용해 자료를 요약해 보여주는 것
- 표와 그래프, 자료의 중심과 퍼짐 등

▶추리 통계(inferential statistics)

- 모집단의 모수를 추정하고, 이 추정치를 이용해 모집단의 속성에 대해 추론하는 통계
- 집단 중심: 평균 검정(z 검정, t 검정 등), 분산분석 등
- 변수 중심: 상관관계, 회귀분석 등



모집단, 전집 (population)	연구자가 관심을 두고 있는 전체 집단
모수(parameter)	모집단의 속성을 보여주는 값
표본(sample)	연구를 위하여 선택된 모집단의 일부분
통계치(statistic)	또는 추정치(estimate) 표본의 속성을 보여주는 값
표집(sampling)	모집단으로부터 표본을 추출하는 과정
일반화 (generalization)	표본을 통해 분석한 결과를 모집단으로 확장하여 서술하는 것

02 가설검정

1 표본으로 전체를 설명하는 가설 검정

표본조사로 데이터를 얻었다면 그 데이터를 활용하여 목적에 따른 가설을 세우고, 가설을 입증하기 위한 가설검증을 진행한다. 이러한 가설 검정 방법에는 귀무가설과 대립가설이 있다.
충분한 증거(가설검정의 결과가 통계적으로 유의할 경우)를 제공하여야 대립가설을 채책할수 있다.

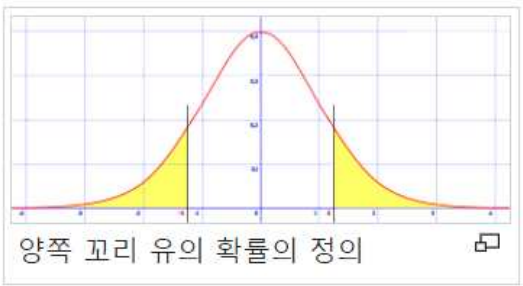
귀무가설(HO)	귀무가설의 사전적 정의는 두 모수값이 서로 차이가 없다고 하는 가설로, 기각될 것을 상정하고 세우는 가설이다. 0가설이라고 한다. 예) “새로운 상품 디자인은 현재 디자인과 별로 다를 게 없다.” 예) “새로운 약품은 시중에 나와 있는 약품과 치료효과 면에서 별로 다른 바 없다.”
대립가설(H1)	조사자가 증명하고자 하는 가설을 대립가설(alternative hypothesis), 혹은 연구가설(research hypothesis)이라 한다. 귀무가설과 반대되는 가설로서, 실제로 주장하고자 하는 가설이다.

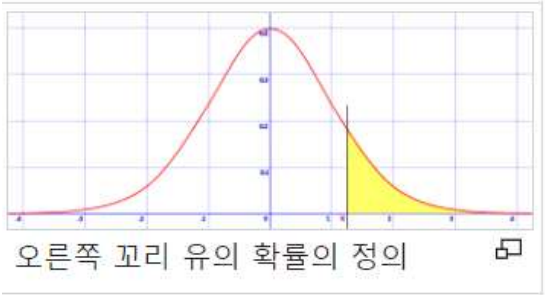
앞서 통조림 위생조사를 실시하여 제품의 붕량률이 3%보다 낮을 때 판매가 가능한 상황이라면 통조림의 분량률이 3%보다 낮다는 것을 증명해야 한다. 이때 가설 검정을 사용하며, 가설 검정 절차는 일반적으로 주장하는 내용과 반대의 주장을 기각하는 방법으로 실제 주장하는 내용을 증명한다.

(1) ‘통조림 불량률이 3%이상이다’ -> 기각될 것을 예상하고 세우는 가설
(2) ‘통조림 불량률이 3%이하이다’ -> 대립가설
대립가설을 선택할수 있는 기준은 P값이 0.05(5%) 미만일 때 가능하다.

2 P값(유의확률: p-value, probability value)

P값은 귀무가설의 채택 여부를 결정하는 기준값으로 귀무가설이 옳다고 가정하고 확률을 구해서 참과 거짓을 판단한다. P값이 0.05 미만이라면 귀무가설을 기각하고, 대립가설을 채택한다. 즉 어떤 사건이 우연히 발생할 확률이 5%보다 낮을 가능성은 거의 없으며, 만약 발생한다면 그것은 우연히 일어난 것이 아니라 유의(통계적으로 의미가 있음) 했기 때문에 일어났다고 해석하는 것이다.


양쪽 꼬리 유의 확률의 정의


오른쪽 꼬리 유의 확률의 정의

통조림 불량률을 예시로 들자면, 생산된 통조림의 불량률이 3% 이상일 확률이 0.05 보다 낮다는 것은 통조림의 불량률이 3% 이상일 가능성이 거의 없으며, 3% 보다 낮을 것이라고 예측하는 것이다. 따라서 통조림의 불량률이 3% 이상일것이라는 귀무가설을 기각하고, 3% 보다 낮다는 대립가설을 채택하게 된다.

3 오류의 종류

가설검증을 위해 귀무가설을 설정했다면 이후에는 정확한 실험을 통해 귀무가설이 옳은지 또는 대립가설이 옳은지를 파악해야 한다. 하지만 그 결과는 100% 완벽하지 않으며 잘못된 결론을 내릴수도 있다. 이러한 오류를 1종오류와 2종오류라 한다.

오류	설명	예
1종오류	귀무가설이 참인데 기각하는 경우	옛날 방식의 제품이 더 성능이 좋았음에도 불구하고 표본 자료에 의거 새로운 제품을 선택하였다면 관리자는 제 1종 오류를 범하는 경우
2종오류	귀무가설이 거짓인데 채택하는 경우	새로운 제품의 성능이 더 좋았음에도 불구하고 예전 방식의 제품을 고집하였을때

4 유의수준과 기각역

귀무가설을 기각하고 대립가설을 채택할 충분한 증거가 자료에 있는 지를 확인하는 작업이 가설검정의 형태가 되는데 충분하다는 의미에는 두가지 방법이 있다.

첫 번째 방법은 분석자가 참을 수 있는 제 1종 오류의 크기를 명시하는 것이다. 이러한 제 1종 오류를 α라 표기하면 통상적으로 0.05, 0.01 혹은 0.10의 값이 많이 쓰인다. 이런 α값을 검정 유의수준(test significance level)이라 한다. 그러면 주어진 α에 대해 통계적으로 기각역을 결정한다. 표본증거가 이 기각역 안에 들어가면 귀무가설을 기각하고 그렇지 않으면 귀무가설을 채택한다. 기각역은 제 1종 오류의 확률이 많아 봐야 α가 되게끔 정확하게 결정되어야 한다. 표본증거가 이 기각역 안에 들어가는 것을

"α 유의수준에서 통계적으로 유의하다."

라고 한다. 예를 들어 α=0.05 라면 "증거는 5% 유의 수준에서 통계적으로 유의하다." 라고 말한다.

두 번째 방법은 α의 값을 명시하기보다 표본증거가 얼마나 유의하냐를 보고하는 것이다. 소위 p-값(p-value)이다. 예제 11.1에서 참의 평균이 $\mu=0$이라고 하자. 물론 이 평균이 0인지 여부는 아무도 모른다. 그리고 표본의 평균이 +2.5가 나왔다고 하자. 이 시점에서 관리자의 옵션은 두 가지일 것이다. 하나는 관측된 표본 값은 매우 비정상적인 값이므로 귀무가설이 참이라고 결론을 내리는 것이고 다른 하나는 표본 값은 아주 정상적인 값이므로 귀무가설을 기각하고 새로운 제품을 선택하는 것이다. 표본의 p-값은 이러한 개념을 계량화시켜 준다.

03 자료의 축약법

통계학에서 하는 큰 작업 중의 하나가 '자료의 축약'이다. 통계학에서의 많은 방법들이 '자료를 어떻게 줄이느냐?'하는 문제와 관련이 있다.

1 대표적인 축약법: 도수분포표와 히스토그램

도수분포표(度數分布表, frequency table):
관측(觀測)값을 몇 개 범주로 나눈 다음 그 범주에 속하는 관측값의 개수(도수, frequency)와 그 도수를 전체 관측값의 개수로 나눈 값(상대(相對) 도수, relative frequency)에 대한 표이다.

<관측값>

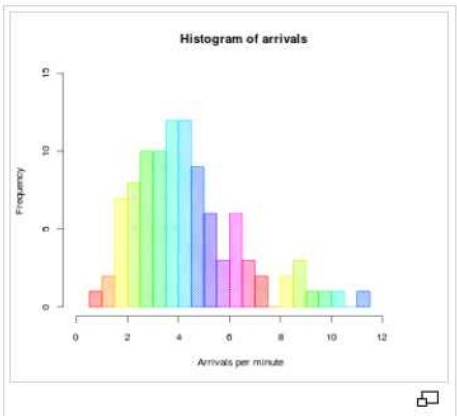
170	178	171	168	173	178	171	174	170	170	175
170	169	166	162	170	171	175	175	171	171	170
172	179	164	170	181	178	180	177	166	169	168
165	163	175	166	178	165	168	167	177	168	177
174	174	176	179	169	173	167	170	173	170	162

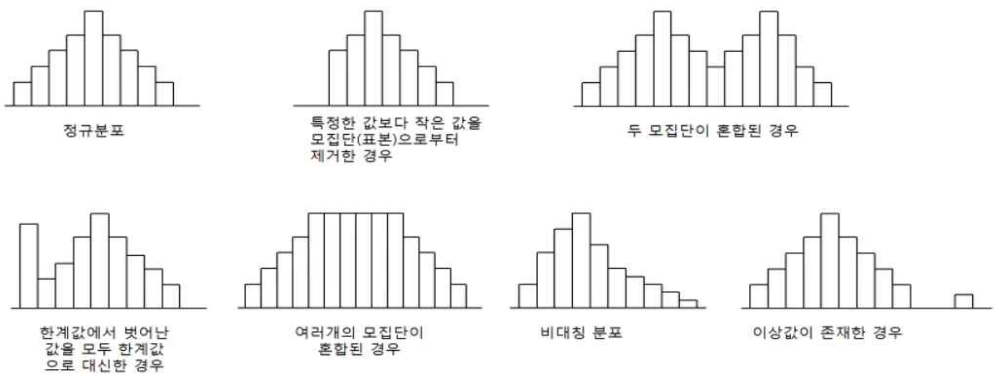
함께하는 통계학의 이해 (최용석, BigBook) 자료 인용

<도수분포표>

계급구간(cm)	도수
161.5 이상 165.5 미만	6
165.5 이상 169.5 미만	12
169.5 이상 173.5 미만	18
173.5 이상 177.5 미만	11
177.5 이상 181.5 미만	8
합계	55

히스토그램(histogram)은 표로 되어 있는 도수 분포를 정보 그림으로 나타낸 것이다. 더 간단하게 말하면, 도수분포표를 그래프로 나타낸 것이다. 보통 히스토그램에서는 가로축이 계급, 세로축이 도수를 뜻하는데, 때때로 반대로 그리기도 한다. 계급은 보통 변수의 구간이고, 서로 겹치지 않는다. 그림에서 계급(막대기)끼리는 서로 붙어 있어야 한다. 히스토그램은 일반 막대그래프와는 다르다. 막대그래프는 계급 즉 가로를 생각하지 않고 세로의 높이로만 나타내지만 히스토그램은 가로와 세로를 함께 생각해야 한다.
공업 분야에서 히스토그램은 품질 관리(QC)를 위한 일곱 가지 도구 중 하나이기도 하다. 여기에는 히스토그램, 파레토도, 체크시트, 관리도, 특성요인도, 순서도, 산포도가 들어간다. 이에 대한 세부 내용은 품질 관리를 참고하라





더 나아가 히스토그램을 이용하여 산술평균과 분산(또는 표준편차)을 구하게 되는 데 이 두 개의 값은 자료의 중심경향과 자료의 흩어진 정도를 나타내는 수치적 측도들이다. 100개의 자료에서 10개의 기둥으로, 다시 2개의 수치 값으로 자료의 축약이 이루어지는 것이다. 여러 집단을 비교할 때 히스토그램을 서로 비교하는 것도 좋지만 각 집단에 대한 산술평균과 분산을 비교하는 것이 더 용이할 때가 많다.

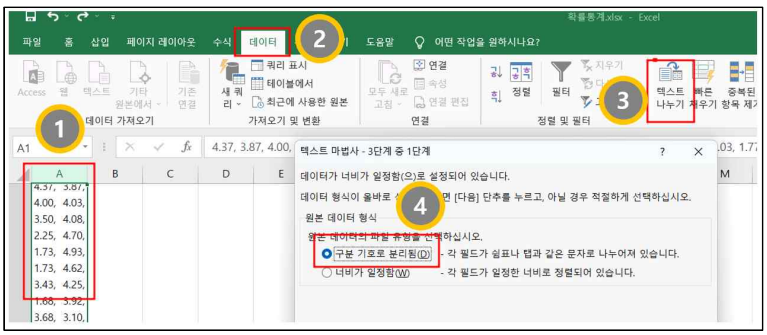
<실습> 엑셀을 이용한 도수분포표 제작하기

다음 사진은 미국 국립공원인 Yellow Stone Park 내에 있는 간헐온천(Old Faithful geyser)을 보여주고 있다. 이러한 간헐천에서 과학자들의 관심을 끈 것은 온천물이 나오는 지속시간(duration)과 온천물이 분출한 후 다음 온천물이 쏟아질 때까지의 간격시간(interval)이다.
다음 자료는 온천물의 지속시간을 분 단위로 잰 자료(107개)이다.

아래의 자료를 이용하여 도수분포표를 작성한 후 히스토그램을 그려 자료의 분포를 확인해본다.

4.37, 3.87, 4.00, 4.03, 3.50, 4.08, 2.25, 4.70, 1.73, 4.93, 1.73, 4.62, 3.43, 4.25, 1.68, 3.92,
3.68, 3.10, 4.03, 1.77, 4.08, 1.75, 3.20, 1.85, 4.62, 1.97, 4.50, 3.92, 4.35, 2.33, 3.83, 1.88,
4.60, 1.80, 4.73, 1.77, 4.57, 1.85, 3.52, 4.00, 3.70, 3.72, 4.25, 3.58, 3.80, 3.77, 3.75, 2.50,
4.50, 4.10, 3.70, 3.80, 3.43, 4.00, 2.27, 4.40, 4.05, 4.25, 3.33, 2.00, 4.33, 2.93, 4.58, 1.90,
3.58, 3.73, 3.73, 1.82, 4.63, 3.50, 4.00, 3.67, 1.67, 4.60, 1.67, 4.00, 1.80, 4.42, 1.90, 4.63,
2.93, 3.50, 1.97, 4.28, 1.83, 4.13, 1.83, 4.65, 4.20, 3.93, 4.33, 1.83, 4.53, 2.03, 4.18, 4.43,
4.07, 4.13, 3.95, 4.10, 2.72, 4.58, 1.90, 4.50, 1.95, 4.83, 4.12

01. '강의자료_확률통계.xlsx' 의 [도수분포표] 워크시트를 선택한뒤 아래작업 실행
   (1) A열 선택
   (2) 데이터
   (3) 텍스트나누기
   (4) 구분기호로 분리됨



02. 구분기호 [쉼표]-[마침] 선택

03. 쉼표단위로 분리된 A행의 자료를 열자료로 변경

(1) A열을 선택한뒤 ctrl+shift+우측 방향키로 범위지정후 복사

(2) 붙여넣기 할 [B3]셀 클릭후

(3) 마우스 우측 클릭하여 [선택하여 붙여넣기]

(4) [행/열 바꿈] -확인

04. 데이터분석을 위한 자료 준비

(1) 1행 자료 제거후
A1에 제목작성

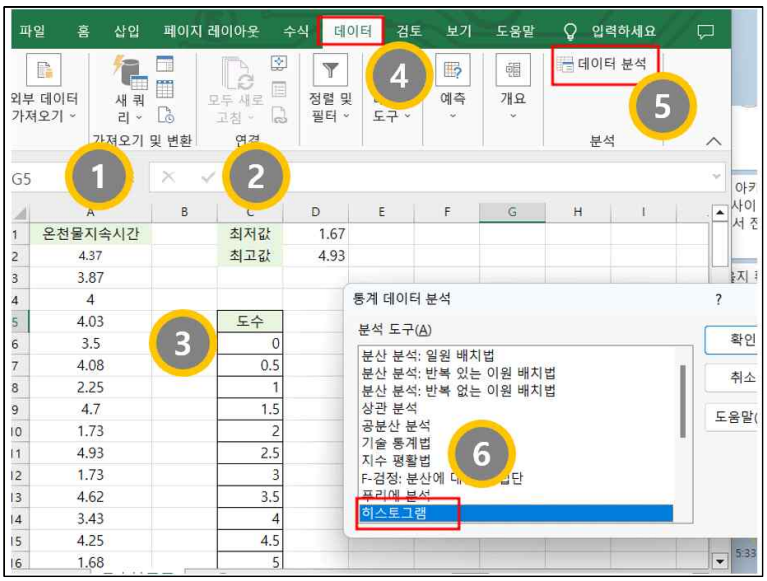
(2) 최저값과, 최고값 확인

(3) 도수 제작
(사용자가 직접입력)
0.5구간, 1구간, 5구간등은
자료값에 따라 설정함

(4) [데이터]

(5) [데이터분석]

(6) [히스토그램] -확인



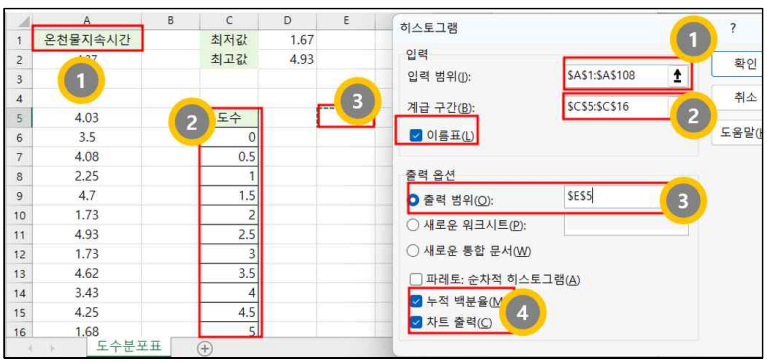
05. 히스토그램 메뉴설정

(1) 입력범위: 온천물지속시간

(2) 계급구간: 도수범위
이름표를 체크하면
범위의 가장상단을 제목으로 인식

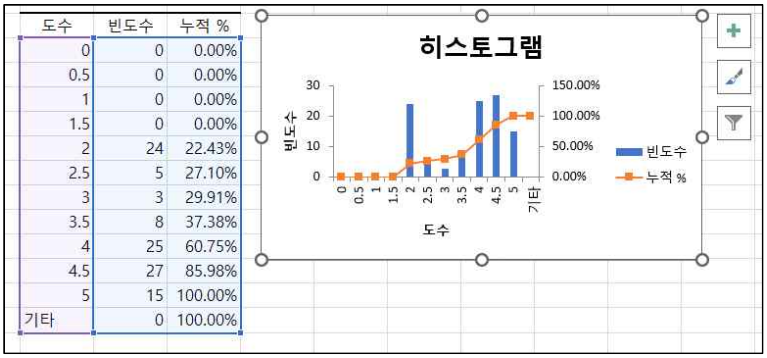
(3) 출력범위: E5셀 클릭

(4) 누적백분율과, 차트 출력 선택



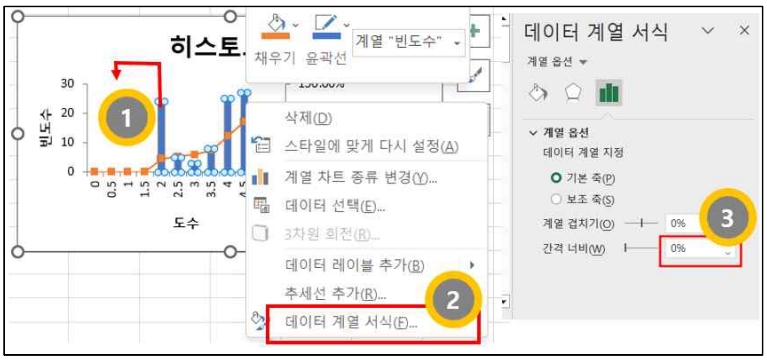
06. 도수표 및 차트 작성됨

도수	빈도수
0	-값에서 0까지
0.5	바로위의 0다음부터~0.5
1	바로위의 0.5다음~1
2	바로위의 1다음~2
2.5	바로위의 2다음~2.5



07. 차트 수정

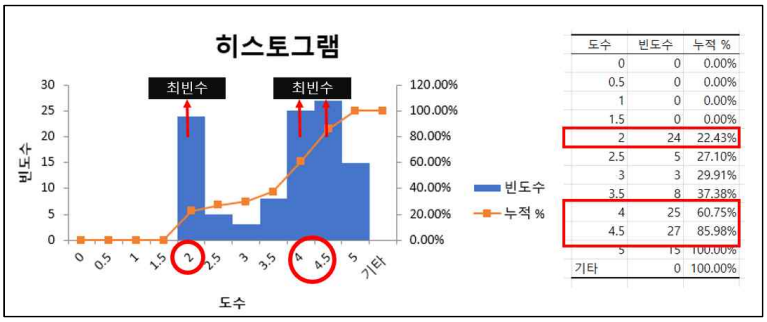
(1) 막대차트 위에서
마우스 우측 클릭
(2) [데이터 계열 서식] 선택
(3) 계열 서식의 [간격 너비]를
0%로 변경함

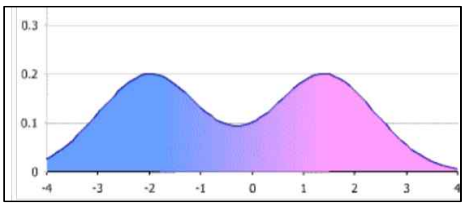


08. 히스토그램 해석

다음 2개구간의 빈도수가 높게 표시되는 쌍봉분포이다.

. 0~2값까지의 온천물의 지속시간 빈도수
. 4~5값까지의 온천물의 지속시간 빈도수





쌍봉분포(雙峰分布, bimodal distribution)는 통계학에서 서로 다른 두 개의 최빈값을 갖는 연속확률분포이다. 왼쪽 그림에서처럼 두 개의 극대값이 있는 확률분포함수 그래프를 나타낸다.

이를 보다 일반화하면 2개 이상의 최빈값을 갖는 확률분포를 다봉분포(多峰分布, multimodal distribution)라고 할 수 있다.

이 히스토그램을 이용하여 산술평균과 분산을 구하니 각각 3.432와 1.151이었으며 이 두 개의 수치적 측도 중 산술 평균은 자료의 중심경향을 알 수 있는 수치적 측도로서 쓰이고 분산은 자료의 흩어진 정도를 알 수 있는 수치적 측도로서 쓰이게 된다. 위의 자료에서는 자료의 중심경향을 알수 있는 수치적 측도로서 최빈값(4.25)도 동시에 제시하는 것이 좋다. 위의 107개의 자료를 축약하여 7개의 기둥을 만들어 자료의 구조를 파악하고 다시 2개의 수치적 측도(산술평균과 분산)로 자료를 축약하여 다른 집단 간의 비교(comparison)에 이 수치들을 이용하게 되는 것이다.

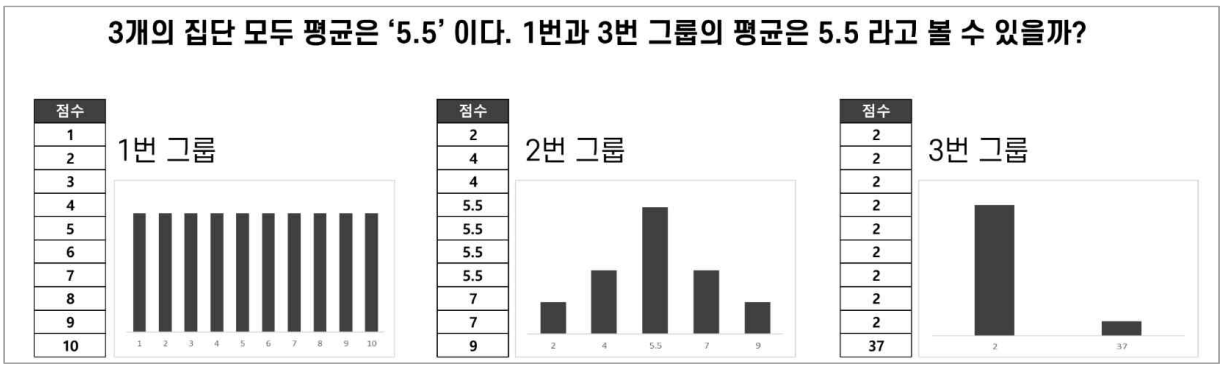
- 도수분포표 용어 -

위의 107개의 자료를 최저값(최저변량)~최고값(최고변량) 1.67~4.93까지 데이터를 0.5구간(구간단위는 데이터에 따라 다름)으로 나누었을 때

(1) 변량: 여러 자료를 수량으로 나타낸 것 (온천물의 지속시간)
(2) 계급: 변량을 일정한 크기로 나눈 구간 (0.5~1, 1~1.5, 1.5~2.5)
(3) 계급의크기: 0.5 (계급의큰쪽끝값-계급의 작은쪽끝값)
(4) 계급의 개수는: ~2, ~2.5, ~3, ~3.5, ~4, ~4.5, ~5 로 총 7개
(5) 계급값: 계급을 대표하는 값 (중앙값) 예) 2~2.5의 가운데인 (2+2.5)/2
(6) 도수: 각계급에 속하는 변량의 개수
(7) 도수분포표: 주어진 전체 자료를 몇 개의 계급으로 나누고 각 계급에 속하는 도수를 조사하여 나타낸 표로서 한자료가 전체에서 어느 위치에 속하는지를 쉽게 알아볼수 있음

- 미리 생각해보기 -

통계의 핵심은 분포이다. 얼마나 퍼져있는가? 치우쳐 있는가? 뾰족한가등



2 대표값: 자료의 중심경향성

- 자료의 특징이나 자료 전체의 경향을 대표할수 있는

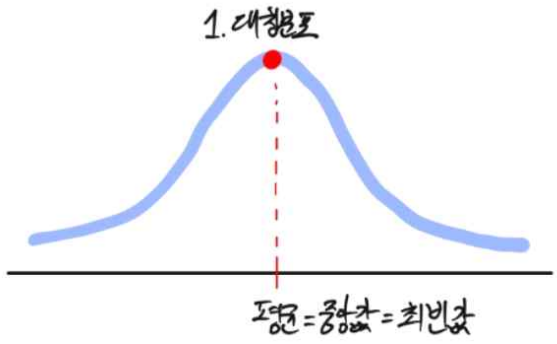
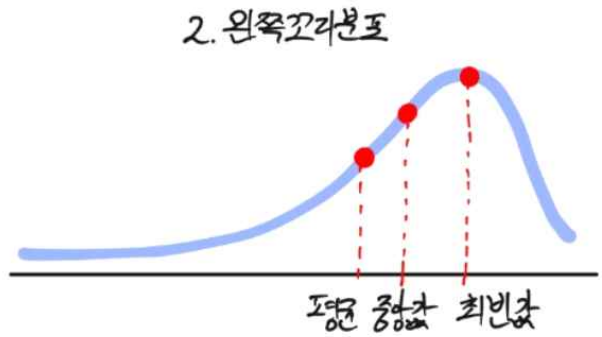
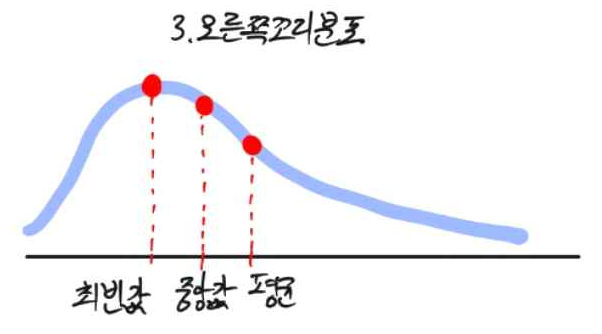
<table>
<tr><th>용어</th><th>설명</th></tr>
<tr><td>중앙값

이상치 (Outlier)에 민감하지 않은 경향을 말함</td><td>
- 도수분포표에서의 중앙값: 계급의 양끝값의 합 / 2
- 전체자료에서 중앙값: 자료를 크기순(오름차순)으로 나열한뒤 가운데값

변량갯수홀수 5개 엑셀함수는 =median(E4:i4) , (1+5)/2=> 3번째위치값
<table>
<tr><td>중앙값</td><td>1위치</td><td>2위치</td><td>3위치</td><td>4위치</td><td>5위치</td></tr>
<tr><td>12</td><td>2</td><td>8</td><td>12</td><td>15</td><td>20</td></tr>
</table>
변량갯수짝수 6개 엑셀함수는 =median(E8:j8)),
(1+6)/2=> 3.5 임으로 (3위치값+4위치값)/2
<table>
<tr><td>중앙값</td><td>1위치</td><td>2위치</td><td>3위치</td><td>4위치</td><td>5위치</td><td>6위치</td></tr>
<tr><td>13.5</td><td>2</td><td>8</td><td>12</td><td>15</td><td>20</td><td>25</td></tr>
</table>
</td></tr>
<tr><td>산술평균 (엑셀 average함수와 도수분포 평균이 꼭 일치하지는 않지만 도수분포표만 있을 때 평균이 필요할 때 사용가능함)</td><td>
자료의총합 / 자료의총갯수 , 변량의(계급값*도수)의 총합 / 도수의총합

[시트명: 대표값]

최저,최고값에서 계급값 정리함
<table>
<tr><th>계급시작 추가</th><th>계급끝</th><th>빈도수 도수</th><th>누적 %</th><th>계급값 (계급시작+계급끝)/2</th><th>계급값*도수</th></tr>
<tr><td>1.5</td><td>2</td><td>24</td><td>22.43%</td><td>1.75</td><td>42</td></tr>
<tr><td>2</td><td>2.5</td><td>5</td><td>27.10%</td><td>2.25</td><td>11.25</td></tr>
<tr><td>2.5</td><td>3</td><td>3</td><td>29.91%</td><td>2.75</td><td>8.25</td></tr>
<tr><td>3</td><td>3.5</td><td>8</td><td>37.38%</td><td>3.25</td><td>26</td></tr>
<tr><td>3.5</td><td>4</td><td>25</td><td>60.75%</td><td>3.75</td><td>93.75</td></tr>
<tr><td>4</td><td>4.5</td><td>27</td><td>85.98%</td><td>4.25</td><td>114.75</td></tr>
<tr><td>4.5</td><td>5</td><td>15</td><td>100.00%</td><td>4.75</td><td>71.25</td></tr>
<tr><td></td><td>합계</td><td>107</td><td></td><td></td><td>367.25</td></tr>
<tr><td></td><td>평균</td><td>(367.25/107)</td><td></td><td></td><td>3.432242991</td></tr>
<tr><td></td><td>엑셀평균함수</td><td></td><td></td><td></td><td>3.459906542</td></tr>
</table>
</td></tr>
<tr><td>최빈값</td><td>
가장빈번하게 나타난수(빈도수 max에 해당하는 값)
엑셀함수 mode, mode.mul, mode.mult
https://xlworks.net/excel-function-mode-mult/
</td></tr>
</table>

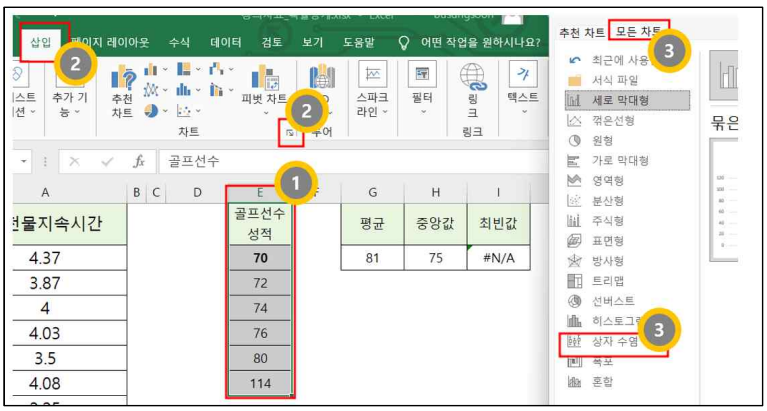
대칭 분포인 경우, **"평균 = 중앙값 = 최빈값"**	
왼쪽 꼬리 분포인 경우, **"평균 < 중앙값 < 최빈값"**	
오른 꼬리 분포인 경우, **"평균 > 중앙값 > 최빈값"**	

<문제> 다음의 골프 선수들의 성적을 가장 잘 묘사하는 값은 무엇입니까?

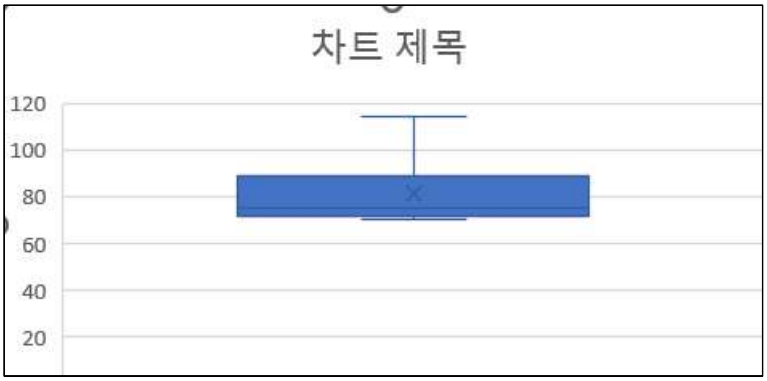
70, 72, 74, 76, 80, 114
(평균)은 은 자료에 나타난 대부분의 성적보다 큰 값을 나타내기 때문에 (중앙값)이 선수들의 성적을 더 잘 묘사합니다.

가장 알맞은 중심경향치는 자료를 가장 잘 묘사하는 값이어야 합니다. 대부분의 골프 선수들은 어떤 점수를 냈나요? 평균에 가까운 점수인가요? 아니면 중앙값에 가까운 점수인가요?
114점을 기록한 골프 선수는 다른 선수들과 너무 다른 점수를 냈기에 이상치라고 할 수 있음. 상한 이상치가 있기 때문에, 자료를 가장 잘 묘사하는 값은 중앙값입니다. 평균인 81점은 대부분의 점수보다 높기 때문에, 자료를 정확히 묘사할 수 없음.

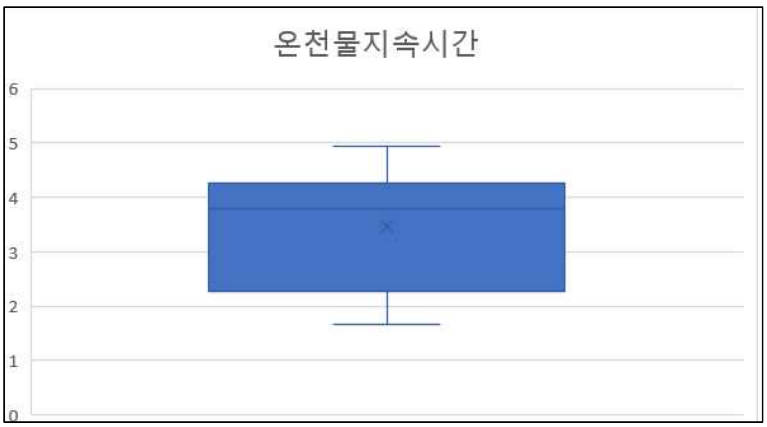
01. '강의자료_확률통계.xlsx' 의 [이상치] 워크시트를 선택한 뒤 아래작업 실행
(1) E1:E7 셀 선택
(2) 삽입-차트
(3) 모든차트-상자수염
* 최빈값은 중복된값이 없음으로 N/A로 출력됨



02. 대부분의 데이터(네모상자)가 아래쪽으로 내려 있음, 즉 ~80까지의 데이터셋이 많음을 알수 있음으로 평균값인 81 보다.
중앙값인 75가
대푯값으로 적당함.
구글검색 '박스플랏'



03. 온천물지속시간 변량을 가장 잘 묘사하는 값으로 사용할수 있는 대푯값은?
(1) 중앙값
(2) 평균
(3) 최빈값

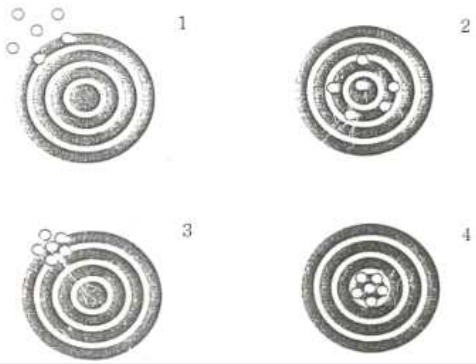


04 자료의 변동

'변동'이라는 문제는 현대생활에서 아주 중요한 통계학의 주제가 되고 있다. 기업체들이 자주 언급하는 식스시그마(6) 운동도 변동에 초점을 맞춘 개념이다. 예로 키가 모두 큰 부모 하에서 주로 키 큰 자녀들이 나오지만 키가 작은 자녀도 나타나고, 키가 모두 작은 부모 밑에서 주로 키가 작은 자녀들이 나오지만 키가 큰 자녀도 나타나는 현상도 변동의 문제인 것이다.

우측그림은 사격장에서 6발 사격을 한 후 총알의 흔적을 나타내는 그림이다. 2번과 4번을 보면 총알의 흔적의 중심과 동심원들의 중심이 일치한다. 이러한 경우를 품질공학에서는 정확하다(accurate)고 하고 통계학에서는 비편향 하다(unbiased)고 표현한다. 반면 3번과 4번을 보면 총알의 흔적이 조밀하게 모여 있음알 수 있다.

- 정확성과 정밀성 -



이러한 경우를 품질공학에서는 정밀하다(precise)고 하고 통계학에서는 유효하다 (efficient)고, 또는 변동이 적다고 표현한다. 그러면 2번과 3번 중 4번으로 개선하기가 수월한 것은 어느 것인가? 2번에서 4번으로 바꾸는 것보다 3번에서 4번으로 바꾸는 것이 훨씬 수월함을 경험상 알 수 있다. 이러한 현상을 제품의 품질에 적용하여 보면 제품 품질의 변동이 크면 설혹 제품품질의 중심이 목표값과 같더라도 제품 품질의 변동을 줄이 기가 어려우나 제품 품질의 변동이 적으면 제품품질의 중심을 목표값으로 이동하기가 훨씬 수월하다는 것이다. 제조업체 중심으로 많이 거론되는 식스시그마 운동의 핵심도 제품의 '변동'에 있다고 하겠다.

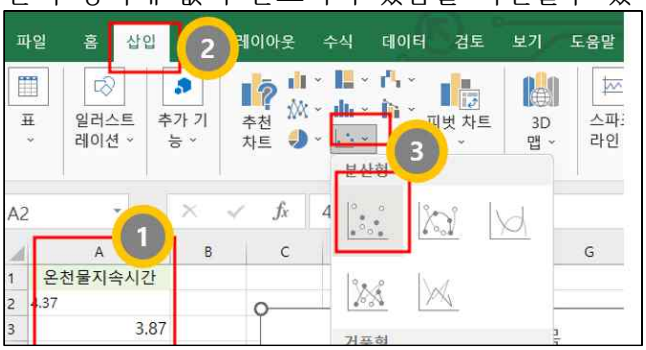
▶ [산포도] 자료가 흩어져있는 정도를 나타내는 값

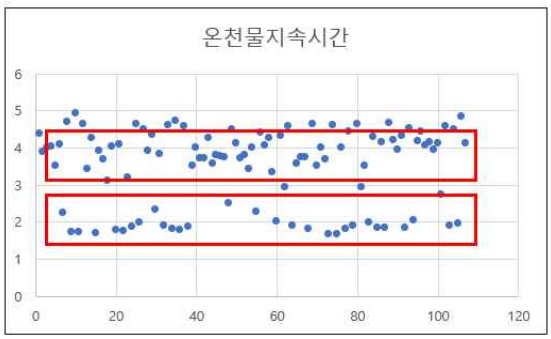
산포도(散布度) 또는 변산도(變散度)는 변량이 흩어져 있는 정도를 하나의 수로 나타낸 값이다.

[산포도차트] '강의자료_확률통계.xlsx' 의 [산점도] 워크시트를 선택한뒤 아래작업 실행

(1) A1셀클릭 (2) [삽입] (3) [산점도]차트 선택

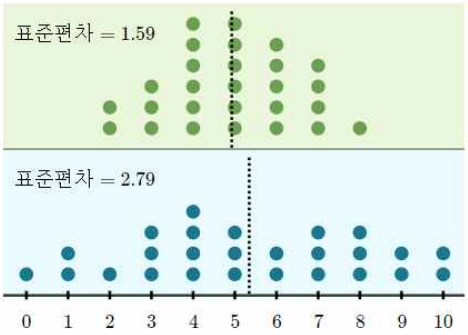
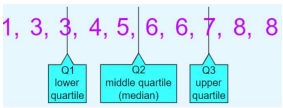
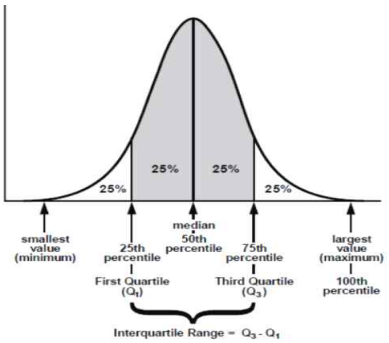
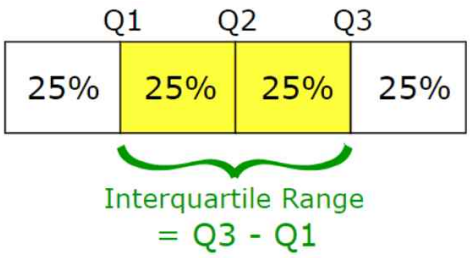
산점도 차트를 보면 변량값은 1~5값으로 분포되어 있으며 대부분 아래 산점도 차트의 빨간색 영역에 값이 분포되어 있음을 확인할수 있다.





[산포도] 데이터의 퍼짐정도를 알수 있는 척도

- 도수분포에서 분포정도가 크면, 분산과 표준편차도 커짐

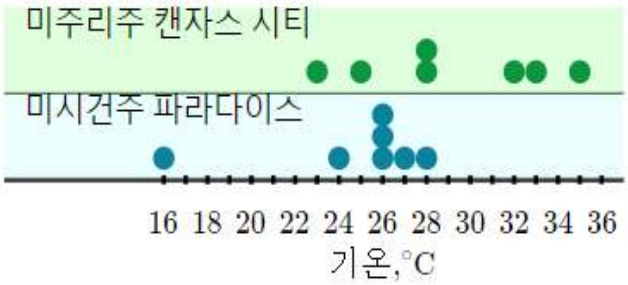
용어	설명	
편차	변량에서 평균 뺌 $\lvert x - \bar{x} \rvert$	편차는 각 관측값이 자료의 중심(평균)으로부터 떨어져 있는 정도로 (관측값-표본평균)로 구할 수 있다. 평균을 중심으로 데이터들이 모여있기 때문에 편차들의 합은 0이고, 편차의 평균도 0이다. 때문에 편차로 퍼진 정도를 측정하는 것은 적합하지 않다.
분산 모분산(/n) ->var.P 표준분산(/n-1) ->var.S	편차 제곱의 평균 $\sigma^2 = \frac{\sum_{i}^{n}(x_i - \bar{x})^2}{n}$ $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	퍼진 정도를 측정하는 것은 관측값이 중심위치에서 얼마나 떨어져 있는가를 알기위한 것으로 편차의 부호는 중요하지 않고, 편차의 크기만 중요하다. 분산은 편차 제곱의 합을 구한 후 관측값의 개수에서 1을 빼준다. 1을 빼주는 이유는 편차의 합이 0이라는 제약이 있기 때문이다.
표준편차	분산의 양의 제곱근 $\sqrt{\frac{\sum \lvert x - \bar{x} \rvert^2}{n}}$	표준편차는 분산에 제곱근을 씌워 관측값의 단위와 맞춰준다. 표준편차가 0에 가까울수록 측정값이 평균(점선)에 가까움을 의미합니다. 측정값이 평균으로부터 멀리 떨어져 있을수록 표준편차는 커진다. 
사분위수	→ Q1 = 1사분위수 = 25 percentile → Q2 = 2사분위수 = 50 percentile → Q3 = 3사분위수 = 75 percentile → Q4 = 4사분위수 = 100 percentil 	데이터 표본을 4개의 동일한 부분으로 나눈 값 
사분위수범위 IQR (Interquartile Range)	Q3-Q1 전체자료의 50%를 포함하는 범위	

(참고) 모집단과 표본집단의 표기법

Parameter 모수		Statistic 통계량
μ	Mean 평균	$\bar{X}$
σ^2	Variance 분산	S^2
σ	Standard Deviation 표준편차	S

▶ 다음 이미지의 값 그래프를 이용하여 질문에 답

[질문1] 아래의 산점도표를 살펴보면 어떤 도시의 기온이 더 흩어져 있다고 할수 있나?



(1) 미주리주 캔자스 시티 기온의 분포가 더 간격이 있는 것으로 보아, 캔자스 시티의 기온이 더 많이 흩어져있는 것으로 보임.

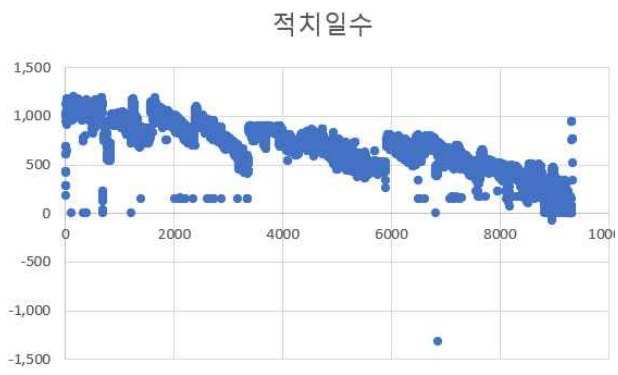
(2) 미시간주 파라다이스 기온의 분포가 집중되어 있는 것으로 보아, 파라다이스의 기온이 더 많이 흩어져있는 것으로 보입니다.

[질문2] 이 경우, 사분위수 범위가 산포도보다 측정하기에 더 적합하다. 그 이유는?

(1) 파라다이스 기온의 하위 이상치는 자료의 범위에 영향을 미치지만 사분위수 범위에는 영향을 미치지 않는다.
(2) 캔자스 시티의 사분위수가 파라다이스의 사분위수보다 크기 때문에, 캔자스 시티의 기온이 더 흩어져 분포함을 알 수 있다.
(3) 범위를 보면, 파라다이스의 자료들이 더 뭉쳐서 분포한다.

[실습1] 시트명 [실습-잔채총일람] 자료를 이용하여 아래의 차트를 완성하고 해석하여 봅니다..

산포도

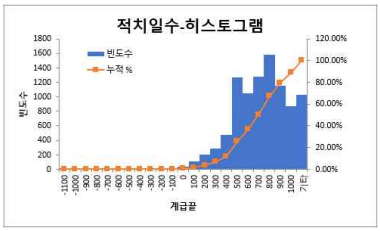


중심경향 대푯값 선정

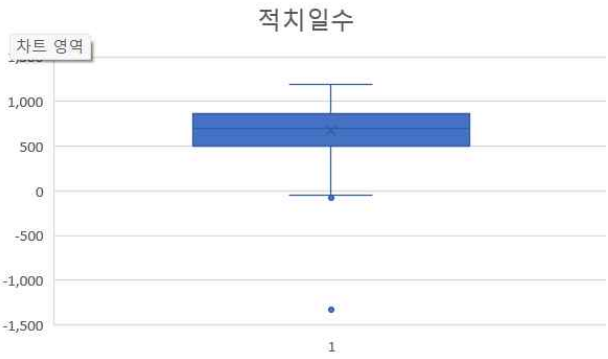
E	F	G	H
최저	최고	중앙값	최빈값(한개출력)
-1323	1196	699	1005

히스토그램

계급끝	빈도수	누적 %
-1100	1	0.01%
-1000	0	0.01%
-900	0	0.01%
-800	0	0.01%
-700	0	0.01%
-600	0	0.01%
-500	0	0.01%
-400	0	0.01%
-300	0	0.01%
-200	0	0.01%
-100	0	0.01%
0	36	0.40%
100	105	1.52%
200	207	3.74%
300	283	6.77%
400	471	11.81%
500	1271	25.42%
600	1050	36.66%
700	1276	50.32%
800	1577	67.21%
900	1157	79.59%
1000	876	88.97%
기타	1030	100.00%

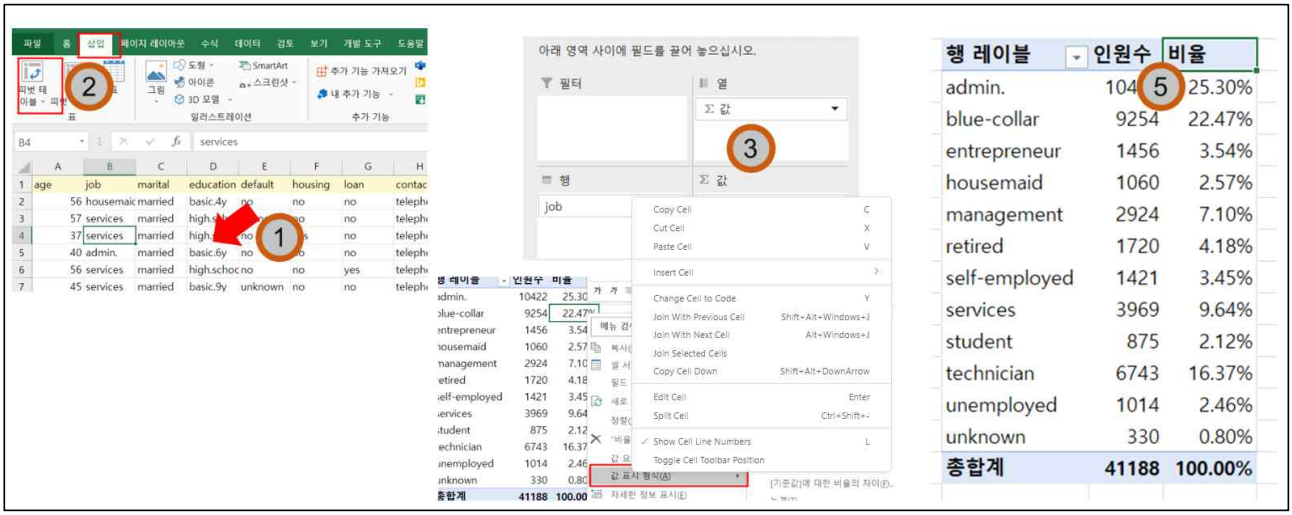


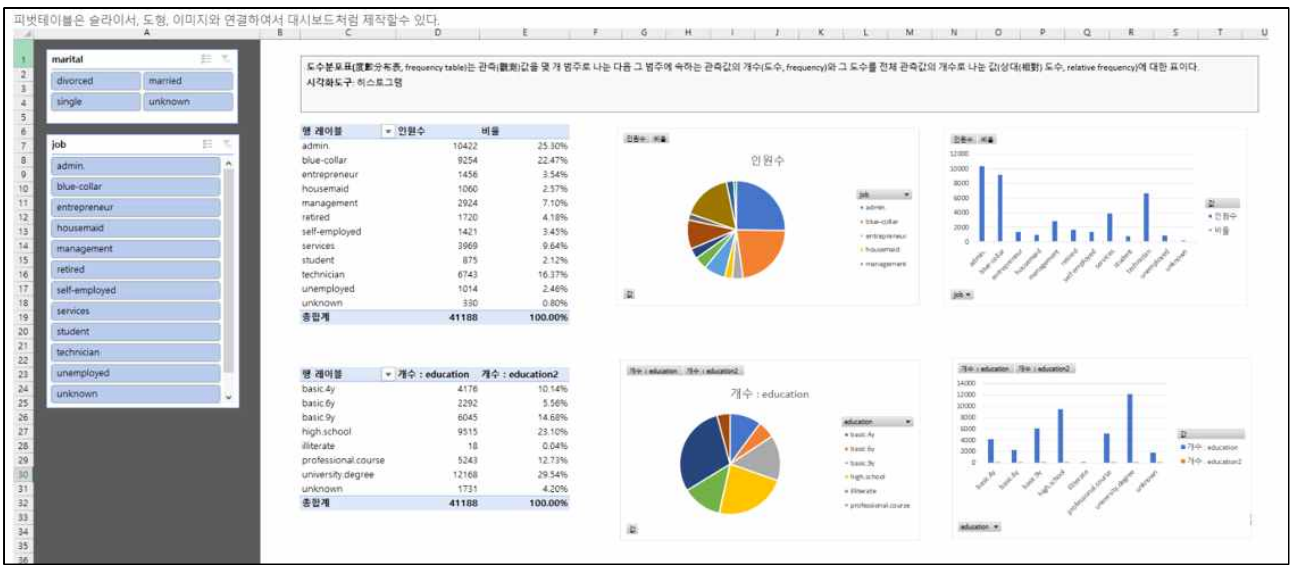
이상치확인



[실습2] 시트명 [실습-잔채총일람] 자료를 이용하여 아래의 도수분포표를 참고하고 자료를 완성함.

*질적자료(문자형)에서의 도수분포표





[실습3] 시트명 [실습-국내철강 원자재 가격데이터] 년도별 가격 추이 정리 자료를 이용하여 데이터의 특성을 이해하고 요약

	A	B	C	D	E	F	G	H	I
1	연도-월	철광석($/톤)	유연탄($/톤)	철스크랩($/톤)	철스크랩(엔/톤	철근(천원/톤)	열연(천원/톤)	후판(천원/톤)	냉연(천원/톤)
2	20-5	95	91	195	17750	648	660	650	718
3	20-4	84	111	202	17980	645	675	665	736
4	20-3	88	145	229	18525	596	700	695	740
5	20-2	87	142	236	19575	590	710	710	730
6	20-1	92	130	243	23060	591	702	702	716
7	19-12	91	121	222	22675	539	698	698	710
8	19-11	85	122	200	21093	568	690	690	713
9	19-10	89	125	189	21160	610	706	714	734
10	19-9	93	128	223	22850	648	728	728	748
11	19-8	90	140	242	24400	666	726	726	750
12	19-7	117	165	226	24420	687	720	720	752
13	19-6	111	179	238	27275	696	735	738	754
14	19-5	99	184	269	28920	695	734	744	750
15	19-4	93	178	296	31700	696	735	745	750
16	19-3	86.2	180.1	309.7	31500	695	726	736	712
17	19-2	88.9	180.4	293.6	28700	694	714	724	690
18	19-1	77.3	179.1	303	28260	688	706	721	730
19	18-12	70.5	185.2	324.9	28600	711	720	745	783
20	18-11	72.2	193.6	323.4	32360	721	736	789	810
21	18-10	72.9	185.1	305.1	34725	699	740	795	810
22	18-9	69.3	176.8	301.3	34075	668	740	795	810
23	18-8	67.3	159	321.1	33440	683	730	792	810

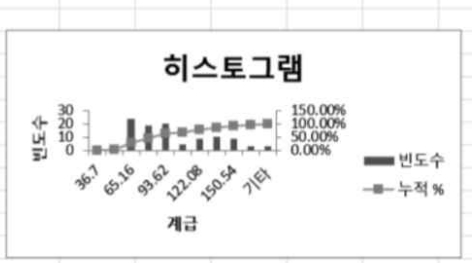
데이터분석 - 기술통계

K	L
철광석($/톤)	
평균	91.83271028
표준 오차	3.309674606
중앙값	85
최빈값	93
표준 편차	34.23554033
분산	1172.072222
첨도	-0.514163196
왜도	0.648866387
범위	142.3
최소값	36.7
최대값	179
합	9826.1
관측수	107
신뢰 수준(95.0%)	6.561751811

데이터분석 - 히스토그램 및

K	L
철광석($/톤)	
평균	91.83271028
표준 오차	3.309674606
중앙값	85
최빈값	93
표준 편차	34.23554033
분산	1172.072222
첨도	-0.514163196
왜도	0.648866387
범위	142.3
최소값	36.7
최대값	179
합	9826.1
관측수	107
신뢰 수준(95.0%)	6.561751811

계급	빈도수	누적%
36.7	1	0.93%
50.93	4	4.67%
65.16	24	27.10%
79.39	19	44.86%
93.62	20	63.55%
107.85	5	68.22%
122.08	9	76.64%
136.31	10	85.98%
150.54	9	94.39%
164.77	3	97.20%
기타	3	100.00%



데이터분석 - 기술통계					
철광석 이외에 철스크랩, 열연, 냉연 자료를 이용하여 기술 통계표를 작성하여 봅니다.					

철스크랩($/톤)		열연(천원/톤)		냉연(천원/톤)	
평균	291.6222643	평균	711.5046729	평균	845.7336449
표준 오차	6.849988166	표준 오차	8.821256187	표준 오차	12.67000304
중앙값	305.1	중앙값	726	중앙값	810
최빈값	351.1811024	최빈값	700	최빈값	950
표준 편차	70.85682856	표준 편차	91.24778352	표준 편차	131.0595306
분산	5020.690153	분산	8326.157997	분산	17176.60056
첨도	-0.905673881	첨도	0.01936967	첨도	-0.64912293
왜도	-0.224167976	왜도	-0.297426053	왜도	0.363123192
범위	277.8208661	범위	418	범위	605
최소값	138.6811024	최소값	490	최소값	575
최대값	416.5019685	최대값	908	최대값	1180
합	31203.58228	합	76131	합	90493.5
관측수	107	관측수	107	관측수	107
신뢰 수준(95.0%)	13.58076778	신뢰 수준(95.0%)	17.48899836	신뢰 수준(95.0%)	25.11951334

첨도 비교	
표준편차비교	
평균비교	
표준편차와 평균비교	

05 데이터의 안정성 확인

표준편차로 대부분의 데이터가 포함되는 범위인 '평균±2x표준편차'를 설명하려면, 해당 데이터에서 집단은 반드시 정규분포를 가져야 한다.
하지만 실무에서 다루는 데이터는 일부 정규분포를 갖는 경우도 있지만, 완벽한 정규분포를 갖는 데이터는 없으므로 보통 '왜도' 와 '첨도'를 바탕으로 데이터의 정규분포 여부를 대략적으로 판단한다. 실무에서 왜도와 첨도로 정규분포를 판단하는 기준은 다음과 같다.

-2 < 왜도 < 2
-3 < 첨도 < 8

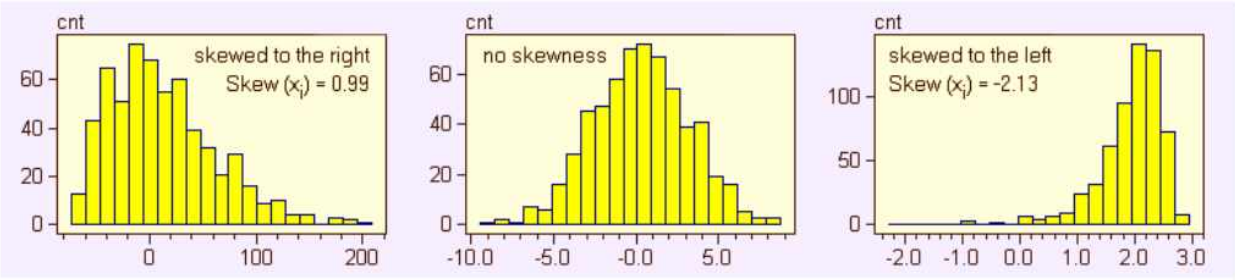
*** 데이터가 분포하고 있는 모양을 비교하여 주는 통계량**

- **왜도(skewness) 비대칭도: 좌우대칭의 분포를 기준으로 분포가 왼쪽이나 오른쪽으로 치우쳐 있을 때 그 정도를 나타내는 척도**
- **첨도(kurtoisis) : 분포의 모양이 얼마나 뾰족한가를 나타내는 척도**
- 데이터의 분포가 한 쪽으로 쏠려있을 경우, 평균보다 중앙값이 더 중심경향을 잘 보여준다
- right-skewed(positively skewed: 오른쪽으로 꼬리가 길게 있는 왜도) 분포에서는 중앙값보다 평균이 더 크다(Median<Mean)

1 왜도(skewness)

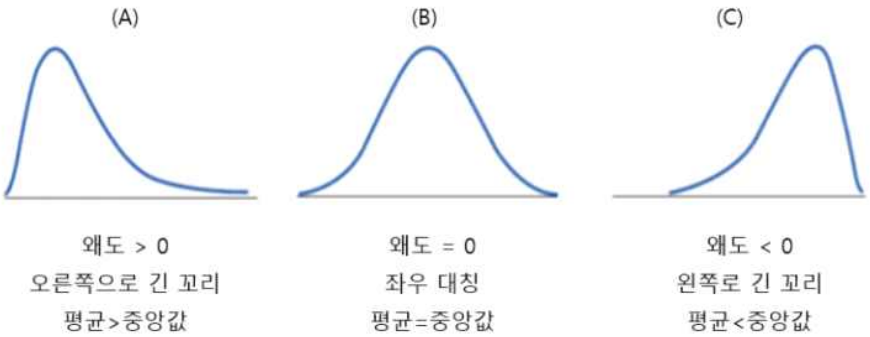
중심경향을 논할 때 중앙값과 평균의 위치를 파악하는 방법으로 왜도가 있다.
왜도(skewness)는 기본적으로 편차의 3제곱의 개념으로 설명이 가능하다. 식 (5.6)은 왜도에 대한 정의인데 편차를 표준편차로 나눈 후 세 제곱을 한 다음 평균을 내는 개념이다.

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3 \quad =\text{skew(salary)}$$



왜도의 크기

따라서 분자와 분모의 단위가 같아지므로 왜도의 단위는 없다. 그리고 왜도의 부호는 편차의 세 제곱에 따른 부호에 따른다. 자료가 왼쪽으로 왜도가 발생한 경우는 부호가 음이 될 것이고 반대로 오른쪽으로 왜도가 발생되면 양의 부호가 된다. 물론 좌우 대칭인 경우는 왜도는 0이 된다. 왜도의 값은 그 자체보다는 두 자료의 왜도를 상대적으로 비교하는데 많이 쓰인다. 기준 점은 물론 대칭인 자료인 경우의 왜도는 0이다.



- 왜도가 0보다 크면 오른쪽으로 긴 꼬리, 0보다 작으면 왼쪽으로 긴 꼬리를 갖음-

- 엑셀-

› SKEW 함수 사용법 간단예제

1 데이터 분포의 왜도 계산하기

```
=SKEW(범위)
'범위 안 데이터 분포의 왜도를 계산합니다.
```

2 왜도가 절대값 2 보다 작을 경우 "정규분포" 출력하기

```
=IF(ABS(SKEW(범위))<2,"정규분포","검증필요")
'왜도의 절대값이 2보다 작을 경우 "정규분포", 그렇지 않을 경우 "검증필요"를 출력합니다.
```

*엑셀 SKEW 함수로 계산 될 데이터 수가 3개 미만이거나 표준편차가 0이면 SKEW 함수는 #DIV/0! 오류를 반환

- 엑셀-

엑셀 '왜도' 시트의 [B7:B18]셀을 피벗테이블로 10~50을 계급크기 '10'으로 구간을 나누어 집계하여 막대차트를 작성하여 보면 왼쪽으로 막대가 몰려있고 오른쪽 꼬리가 있는 차트 모양이 출력된다. (B6셀의 SKEW 왜도 함수의 출력값은 >0 값임)

행 레이블	개수 : 왜도 > 0
10	4
20	3
30	2
40	1
50	1
총합계	11

행 레이블	개수 : 왜도 <0
10	1
20	1
30	2
40	3
50	4
총합계	11

(출처) https://www.oppadu.com/%EC%97%91%EC%85%80-skew-%ED%95%A8%EC%88%98/

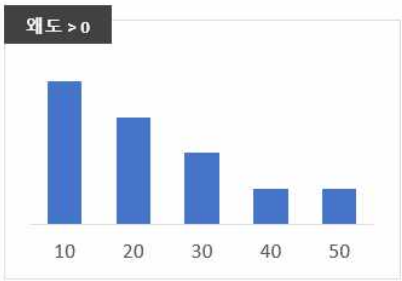
엑셀 SKEW 함수 :: 통계함수

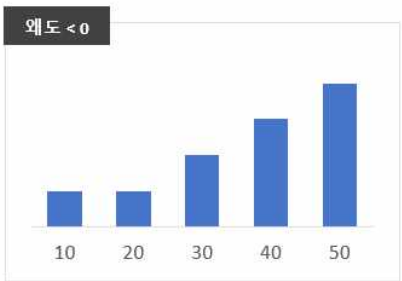
= SKEW (값1, [값2], ...)

데이터 분포의 왜도를 계산합니다.

0.892
왜도 > 0
10
10
10
10
20
20
20
30
30
40
50

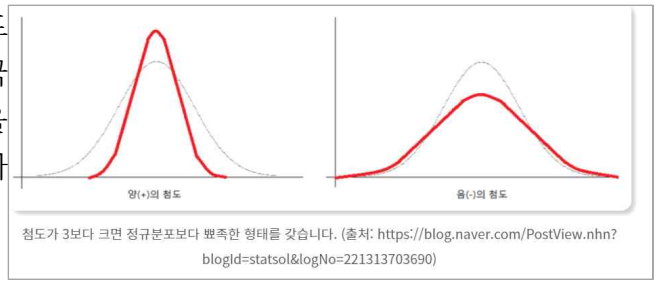
-0.892
왜도 <0
10
20
30
30
40
40
40
50
50
50
50





2 첨도(kurtosis)

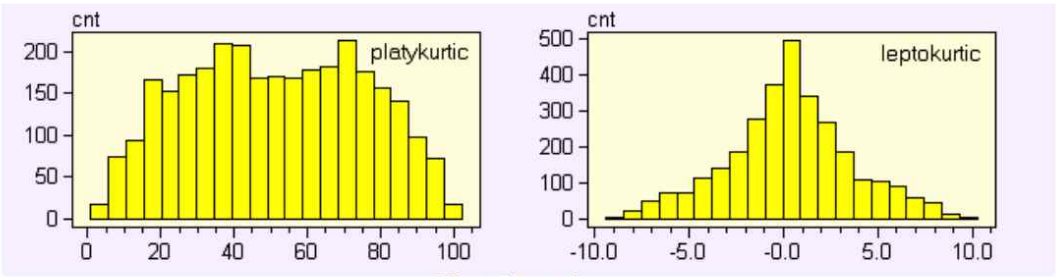
첨도(尖度, 영어: kurtosis 커토시스)는 확률분포의 꼬리가 두꺼운 정도를 나타내는 척도이다. 극단적인 편차 또는 이상치가 많을수록 큰 값을 나타낸다. 첨도값(K)이 3에 가까우면 산포도가 정규분포에 가깝다.



첨도가 3보다 크면 정규분포보다 뾰족한 형태를 갖습니다. (출처: https://blog.naver.com/PostView.nhn?blogId=statsol&logNo=221313703690)

첨도(kurtosis)는 편차의 4제곱의 개념으로 설명이 가능하다. 식 (8.7)은 왜도에 대한 정의인데 편차를 표준편차로 나눈 후 4제곱을 한 다음 3을 빼는 개념이다. 여기서 3을 빼는 이유는 기준이 되는 정규분포인 경우 값이 3이기 때문이다.

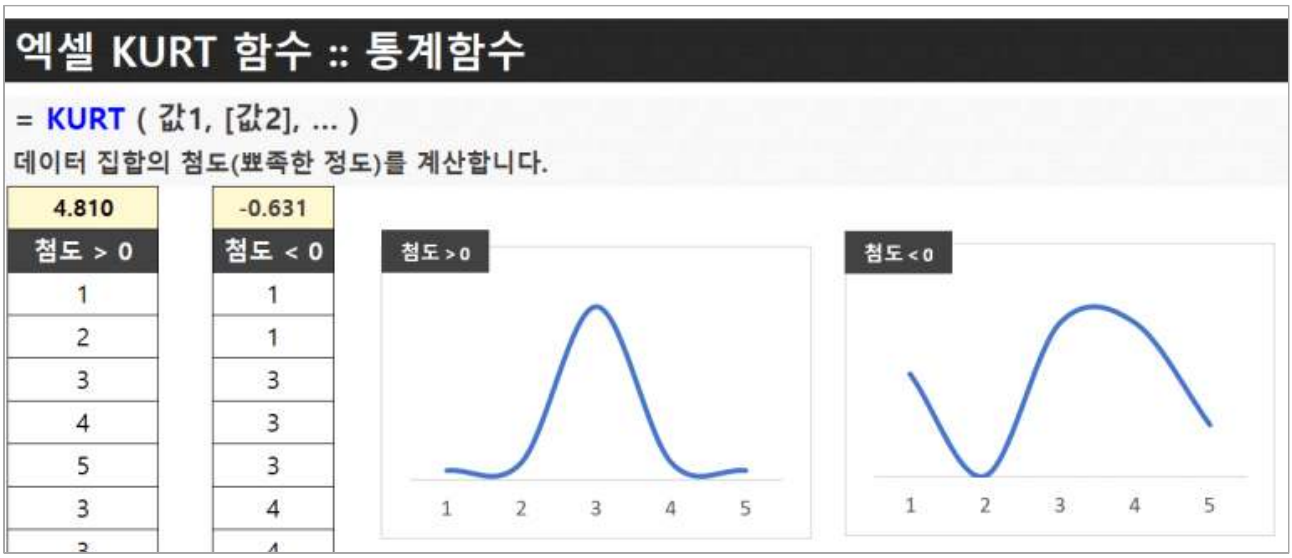
$$\left\{\frac{1}{n}\sum\left(\frac{x_i-\bar{x}}{s}\right)^4\right\}-3$$



첨도의 크기

자료의 중심에 있는 부분이 정규분포에 비해 얼마만큼 뾰족하냐(peakedness)를 측정 하는 방법이다. 역시 분자와 분모의 단위와 같기 때문에 첨도의 단위는 없다. 판정은 다음과 같이 하면 된다.

첨도가 0보다 작으면 중심부분이 짧고 뚱뚱하다(short and fat).
첨도가 0이면 정규분포의 정상부근의 모양새이다(normal).



엑셀 KURT 함수는 데이터 집합의 첨도(뾰족한 정도)를 계산하는 엑셀 통계함수이다. 실무에서는 일반적으로 첨도의 절대값이 8 이내일 경우 정규분포 형태를 갖는다고 판단할 수 있으며, 첨도가 3에 가까울 때 산포도가 가장 정규분포에 가깝다고 판단한다.

› KURT 함수 사용법 간단예제

1 데이터 집합의 첨도 계산하기

```
=KURT(범위)
'데이터 집합의 첨도를 계산합니다.
```

2 첨도를 계산하고 첨도의 절대값이 8보다 크면 검토필요 문구 출력하기

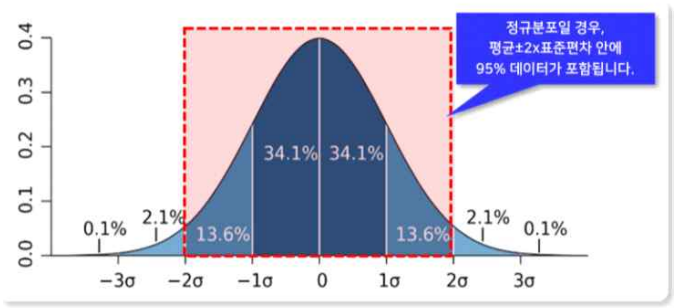
```
=IF(ABS(KURT(범위))>8,"검토필요","")
'데이터 집합의 첨도가 8보다 클 경우 검토필요라는 문구를 출력합니다.
```

3 정리: 엑셀 함수

표준편차는 데이터가 정규분포를 가진다는 전제 하에서 여러 통계분석에 사용되는 핵심 지표임
STDEV.P 함수와 STDEV.S 함수가 나오는데 P는 Population(모집단)의 약자, S는 Sample(표본집단)의 약자이며, 실무에서는 대부분 샘플링 된 데이터(표본집단)을 다루므로, 특별한 상황을 제외하면 STDEV.S 함수를 사용하는 것이 일반적이다.
현업에서 다루는 대부분의 데이터는 집단의 일부분만 샘플링하게 되고 이렇게 분석된 데이터를 바탕으로 모든 집단을 설명할 수 있는 결과를 제시하게 된다.

통계지표	설명
☰ 표준편차(STDEV)	데이터 분포가 **평균으로부터 얼마나 떨어져있는지를 나타내는 지표**입니다. 데이터 집단이 정규분포를 가질 경우, 평균으로부터 **2x표준편차 떨어진 범위 안에 95%의 데이터가 포함**됩니다.
☰ 왜도(SKEW)	데이터 분포가 **얼마나 좌/우로 기울어져있는지 나타내는 지표**입니다. 양수일 경우 왼쪽, 음수일 경우 오른쪽으로 기울어집니다. 일반적으로 **-2~2 사이일 경우 정규분포를 갖는다**고 이야기합니다.
☰ 첨도(KURT)	데이터 분포가 **얼마나 뾰족한 형태로 이루어졌는지 나타내는 지표**입니다. 값이 클수록 더욱 뾰족한 형태로 분포하게 됩니다. 일반적으로 **8보다 작을 경우 정규분포를 갖는다**고 이야기합니다.

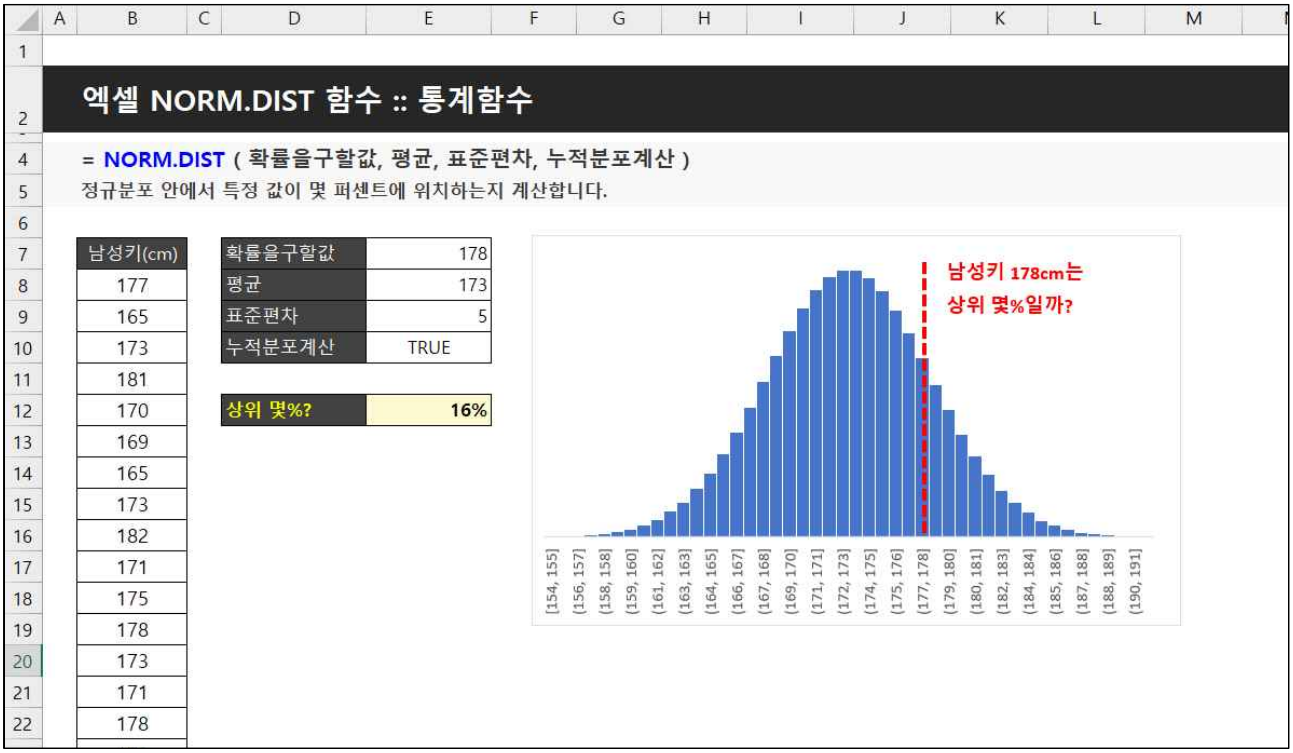
이럴 경우, 선택된 표본이 집단의 모든 값을 대표할 수 있다는 가설을 세우기 위해, 데이터 집단이 '정규분포를 따른다' 라는 가정하에 데이터 분석을 진행하게 된다. 그리고 데이터 집단이 정규분포를 가질경우, 집단의 95% 데이터는 평균으로부터 2x표준편차가 떨어진 범위안에 포함된다.



데이터가 정규분포를 가질 경우, 평균±2x표준편차 범위에 95%의 데이터가 포함됩니다.

따라서, 실무에서 표준편차는 '데이터가 정규분포를 가질 경우, 95%에 해당하는 대부분의 데이터는 평균±2x표준편차 안에 들어간다' 라고 해석할 때 주로 사용된다.

엑셀 NORM.DIST 함수는 정규분포 안에서 특정 값이 몇 퍼센트에 위치하는지 계산하는 함수임



인수	설명
☰ 확률을구할값	**정규분포에서 몇%에 위치하는지 확률을 구할 값**입니다.
☰ 평균값	**분포의 평균**입니다. (AVERAGE 함수로 계산)
☰ 표준편차	**분포의 표준편차**입니다. (STEDEV.S 함수로 계산)
☰ 누적분포함수계산	**TRUE일 경우 누적분포함수로 계산**합니다. FALSE이면 확률밀도함수로 계산합니다. **대부분 실무에서 몇%에 위치하는지 확률을 구할 경우 TRUE(누적분포함수)로 계산**합니다. FALSE로 입력하면 몇%확률로 발생하는지가 계산됩니다.

엑셀 NORM.INV 함수 :: 통계함수

= NORM.INV (확률, 평균, 표준편차)

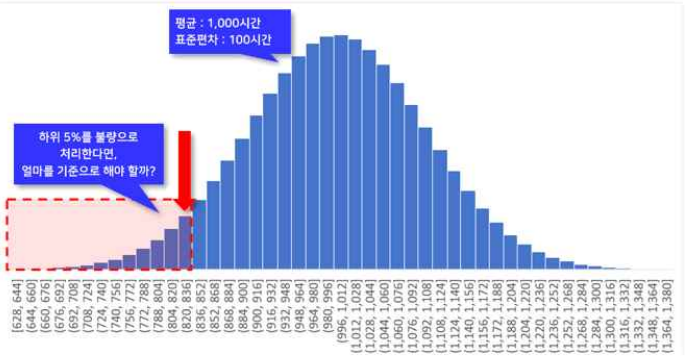
평균과 표준편차의 정규분포에서 특정 확률의 값을 계산합니다. (정규누적분포의 역함수 계산)

기대수명
1,037
1,015
887
937
968
972
979
1,031
1,058
1,000
1,082
1,043
979

평균	1,000
표준편차	100
확률	5%

값	835.64061

94368.96869



NORM.INV 함수는 아래와 같은 상황에 아주 유용하게 사용할 수 있습니다.

- 배터리의 수명을 분석하니 평균이 1,000시간, 표준편차가 100시간으로 계산되었습니다. 그 중 **하위 5% 제품을 Fail로 처리한다면, 몇시간을 기준으로 Fail 제품을 분류**해야 할까요?
- 고객별 매출 현황을 분석해보니 평균이 12만원, 표준편차가 2만원으로 계산되었고, 매출증대를 위해 최소 구매금액을 설정하려고 합니다. **하위 10% 고객을 대상으로 최수주문금액을 설정한다면, 최소 주문금액은 얼마**가 되어야 할까요?

> NORM.INV 함수 사용법 간단예제

1 평균 1,000시간 / 표준편차 100시간일 때, 하위 5%에 해당하는 시간

```
=NORM.INV(0.05, 1000, 100)
=835.5 시간
```

2 평균 12만원 / 표준편차 2만원 일 때, 하위 10%에 해당하는 금액

```
=NORM.INV(0.1, 120000, 20000)
=94,368 원
```

Part 2. 통계적 추론

01 통계적 추론이란?

1 통계적 추론 개념

전체 모집단을 조사하는 것이 불가능하기 때문에 작은 집단의 정보를 활용하여 더 큰 모집단의 속성을 추론한다.
즉 모평균이나 모분산과 같은 모수에 대한 어떤 결정을 내리기 위하여, 모집단에서 표본을 추출하여 데이터를 얻고 이 데이터를 기초로 하여 통계이론에 의한 결론을 내리게 된다.
따라서 가능한 한 정확하게 모수를 측정하는 방법을 개발하는 것이 통계적 분석에서 중요한 관건이다. 이 분야를 다루는 통계학을 추리통계학 또는 통계적 추론(statistical inference)이라 한다. 통계적 추론이란 표본이 갖고 있는 정보를 분석하여 모수에 관한 결론을 유도하고, 모수에 대한 가설의 옳고 그름을 판단하는 것을 말한다.

연구자들의 통계적 추론에 대한 개념을 살펴보면 다음과 같다.

> 통계적 추론은 통계적 정보로부터 또 다른 정보를 이끌어 내거나 정보를 판단하는 사고 수단입니다. 통계에서는 자료 그 자체와 함께 통계적인 방법을 동원하여 자료 너머를 탐색하고 추측하는 사고 활동인 통계적 추론이 통계적 사고의 중심이 됩니다. 또한 통계는 확률을 이용해 실제적인 자료를 처리하는 방법을 다루는 분야이므로, 통계 영역에서의 추론은 상황에 근거하며 자료에 의존적인 귀납적 추론을 주로 사용합니다. 귀납적 추론을 근간으로 하는 통계적 추론은 방대한 자료를 대신할 일부 자료인 표본을 토대로 하여 전체 자료인 모집단의 여러 가지 특징들에 대해서 추론하는 것입니다. 예를 들어 지구가 태양 주위를 돈다는 것을 연역적으로 추론하는 경우는 지구과학의 일반화된 법칙에 의존하여 내일도 지구가 태양의 주위를 돈다는 것을 증명하는 것이지만 귀납적 추론을 바탕으로 하는 통계적 추론은 수집한 자료에 근거하여 내일에 대한 이 확신을 믿는 것이 타당한지를 확률적으로 답합니다.

출처: 이종학(2011), "예비 교사의 통계적 추론 능력에 대한 연구", 한국학교수학회논문집 제 14권 제3호, 299-327

▶ 통계적 추론은 조사자의 관심에 따라

(1) 모수의 추정과 (2) 모수에 대한 가설검정이라는 두 가지 문제로 나눌 수 있다.

(1) 모수의 추정(estimation)이란 미지수인 모수에 대한 추측 혹은
추측치를 그 수치화된 정확도와 함께 제시하는 것이다.
(2) 모수에 대한 가설검증이란 모수에 대한 여러 가지 가설들이 적합한지 혹은
적합하지 않은 것인지를 추출된 표본으로부터 판단하는 것이다.

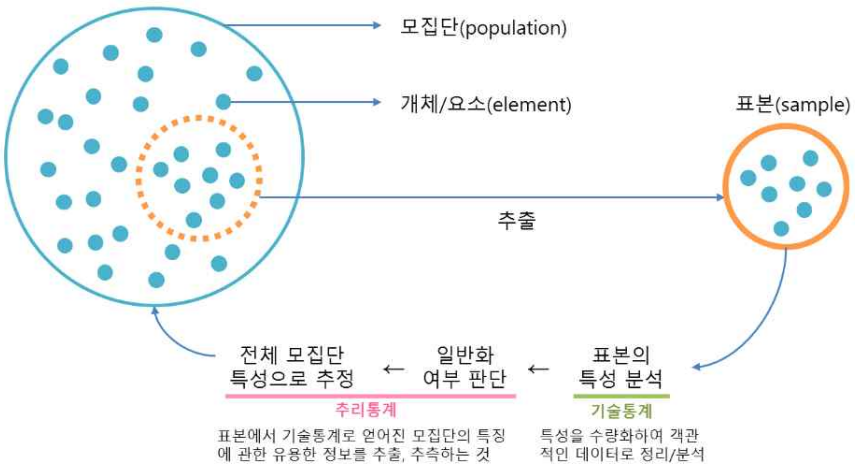
통계적 추론 (statistical inference) = 추정 (estimation) + 가설검정 (testing hypothesis)

정규분포를 이용하여 모평균과 모비율에 대한 추정과 검정으로 설명하면 다음과 같다.

예) 어느 도시에서 청소년기의 성장에 관한 연구를 하기 위해
중학교 1학년 남학생 30명을 임의추출하여 키를 측정하였다.

표본으로부터 얻은 30명 키의 평균으로 그 도시 전체 중학교 1학년 남학생의 평균키를 추론할 수 있다.

(1) μ를 하나의 값으로 추정한다. -> 점추정
(2) μ를 포함할만한 적당한 구간을 정한다. -> 구간추정
(3) μ 값이 5년 전의 평균값과 얼마나 다른지 판단한다. -> 가설검정



2 통계적 추론 사례

가계 로트폴리오의 위험 수준 특정	평균분산모형을 이용한 가계 포트폴리오의 위험 수준 측정 연구에서는 2001년 1월부터 2008년 8월까지의 종합주가지수, 회사채 총수익률지수, 정기예금금리, 주택매매가격 종합지수를 바탕으로 주식과 채권, 무위험자산 및 부동산의 표준 편차를 산출하여 자산별 위험을 측정하였고, 이를 2007년 펀드투자자 조사 자료에 대입하여 가계 포트폴리오의 위험수준을 측정하였다. 이처럼 자료를 통해 투자 활동이나 재무 설계시 가계 포트폴리오의 위험 수준을 파악하여 투자 의사 결정에 도움을 얻을수 있다.
소셜 미디어 및 웹사이트의 상관관계 측정	어도비에서는 디지털 게시자의 소셜미디어 활동과 웹사이트 방문간 잠재적인 관계를 이해하고자 어도비 디지털 게시자의 2주간 시간별 트위터 언급과 웹사이트 방문간의 긍정적인 관계를 도출하였다.
당뇨병 진행도 예측	442명의 당뇨병 환자를 대상으로 한 조사에서 회귀 분석을 활용하여 당뇨병 진행도를 예측할수 있었다. 독립변수로는 나이, 성별, bmi, 평균, 혈압, 혈액검사 수치 등을 사용하했다. 조사한 데이터를 바탕으로 bmi 지수와 평균 혈압이 당뇨병 진행도와 양의 상관관계를 가지고 있다는 사실을 알수 있었다.

02 자료의 종류

1 명목척도(명명척도, nominal scale)

문자열 자료로서 자료의 정보가 가장 적은 자료이다. 서열을 알 수 없는 각 범주(category)에 대응되는 도수(frequency:빈도수)만 주어지는 자료이다.

예) 성별, 주소, 선호하는색상, 국적

측정내용	가능측도	불가능측도
자료의 중 심경향 측정	최빈값(mode)	중앙값(median) 산술평균(arithmetic mean)
흩어진 정도 측정	다양도(성)지수 (diversity index)	분산(variance), 표준편차(standard deviation), 범위(range)

***(참고) 다양성 지수(Diversity Index)란 특정 생태계의 종의 수나 사회에서 인종 또는 민족의 다양성을 나타내는 지수임/ 지수가 높아질수록 해당 지역에 인종과 민족이 다양하게 분포하고 있음을 의미한다. (0~1 또는 0~100으로 표시)**
예를 들어 미국 사회의 다양성 지수가 55라면, 미국 사회에서 임의로 선택한 두명의 인종이 다를 확률이 55%라는 의미이다

다음과 같은 서양인 머리색에서 경우1의 최빈값은 갈색 이고 다양성지수는 0.899이다.
경우2는 검정색 범주에 모든 592 도수가 몰려있음으로 다양성지수가 0이며,
경우3의 머리색의 분포가 모든 범주마다 같으므로 다양성지수가 1또는 100이다.

경우	색깔	검정	갈색	빨강	금색	합계	다양성지수
1	도수	108	286	71	127	592	0.899
2	도수	592	0	0	0	592	0
3	도수	148	148	148	148	592	1

두 번째 자료는 다양성지수가 0이므로 산포도가 제일 작다. 즉, 592명이 모두 머리색이 검정색으로 산포도가 0이 된다. 세 번째 자료는 다양성지수가 1이므로 산포도가 제일 크다. 즉, 592명이 4개의 머리색에 골고루 나뉘어 148명으로 동일하다. ■

2 순서척도(순위, 서열척도 ordinal scale, rank scale)

명목척도에 순위라는 정보를 더 갖는 자료이다. 범주 사이의 서열이 존재하는 자료이다.
예) 시험점수가 아닌 시험을 먼저 끝내서 제출한 순서로 서열
시합에서 빨리 도착한 순서대로 나열
설문결과의 1(매우불만족) / 2(불만족) / 3(보통) /4(만족) /5(매우만족)

서열척도는 +,- 등의 산술연산이 불가능함. 1등과 2등사이의 값이 1차이이지만 1등이 100점, 2등이 70점일수도 있어서 간격차이를 서열척도의 2-1로 계산할수 없음.
만족의 정도또한 사람마다 다를수 있어, 불만족+보통=매우만족과 같은 계산이 불가능함.

측정내용	가능측도	불가능측도
자료의 중심경향 측정	- 최빈값(mode) - 중앙값(median)	- 산술평균(arithmetic mean)
흩어진 정도 측정	- 다양도(성)지수(diversity index) - 사분위수간 범위 (IQR, inter-quartile range)가	- 분산(variance), - 표준편차(standard deviation) - 범위(range)

대학생 2명중 1명은 스스로 유행에 관한 한 캔비족'이라고 생각하는 것으로 나타났다. 아르바이트 구인·구직 포털 아르바이트천국(www.alba.co.kr)이 대학생 380명을 대상으로 유행 민감 정도'에 대한 설문조사에서 대학생 2명 중 1명은 자신이 유행에 민감하다고 생각하는 것으로 나타났다. 민감하다는 의견이 36%로 가장 많았고 보통(29%), 매우 민감(19%) 순이었다.

민감하지 않다는 의견은 전체의 16%에 불과했다. 유행에 민감한 이유로는 예쁘고 멋져보여서'가 49%로 가장 많았고 남에게 뒤지지 싫어서' 20%, 연예인처럼 되고 싶어서'라는 의견이 17%였다. 이런 최신 유행 스타일들은 외국 패션 잡지나 연예인을 통해 먼저 접할 수 있기 때문에 자신도 유행스타일을 따라가면 연예인들처럼 예쁘고 멋져 보일 것이라는 환상을 갖고 있는 것으로 나타났다. 이런 이들을'캔비족(can be·될 수 있다)'이라고 부르는데 유명 연예인의 옷과 액세서리 등 패션을 모방하며 스스로의 가치를 연예인과 동격화하려는 사람들을 말한다. 패션·헤어스타일이 대표적이며 심지어는 다이어트 방법이나 성형까지 그 영역이 확대되고 있다.
'유행에 민감한가?'에 대한 답변을 표로 정리하면 다음과 같다.

민감하지 않다	보통	민감하다	매우 민감하다	합계
16%	29%	36%	19%	100%

위의 답변은
"민감하지 않다"< "보통"< "민감하다"< "매우 민감하다"라는 서열이 매겨진다.
"민감하다"라는 범주가 최빈값이면서 중앙값이 된다

3 구간척도(등간척도, interval scale)

순위척도에 '차이(difference)'라는 정보가 부여된 자료이다.

다음 기사는 0000.00.00년 매일경제 TV 기사 내용이다.

' 소설 인 오늘 전국 곳곳 흐리고 눈비'

오늘 아침 어제보다 기온이 크게 올랐습니다. 포근하다고 느껴질 정도로 며칠동안 이어진 영하권 추위를 무색하게 하는 날씨인데요. 또 오늘은 절기상 소설"입니다. 보통 겨울의 첫 추위가 나타나는 때라고 하는데요. 그런데 오늘 추위는 없고 대신 소설을 알리는 비가 내리겠습니다. 지금도 중부지역에 약하게 비가 내리고 있는데요. 오늘 흐리고 곳곳에 오락가락 약하게 비가 이어지겠습니다. 내리는 비의 양은 많지 않겠지만 우산 챙기셔야 겠습니다. 구름모습입니다. 중부지역에 비구름이 들어와 있습니다. 밤에는 이 구름이 남부지역까지 내려오겠습니다. 지역별 오늘 날씨입니다. 오늘 중부지역은 흐리고 비가 내리겠습니다. 호남과 경북지역은 흐려져서 밤 한 때 비가 내리겠습니다. 현재기온 서울이 5도, 전주와 광주가 6도로 어제보다 기온이 5도 이상 크게 올랐습니다. 낮기온은 서울이 8도, 강릉 11도, 부산 16도로 어제와 비슷하거나 조금 더 높겠습니다. 내일 차차 날씨가 맑아지겠습니다. 이번 주말에도 포근한 날씨는 계속 되겠고요. 월요일 에는 중북부지역에 비가 예상됩니다. 날씨였습니다.

==> 기온에 해당되는 자료가 구간척도가 된다

.

(예) 앞에서 언급한 날씨 기사에서 "부산의 낮기온 16^0C은 서울의 낮기온 8^0C보다 온도가 8^0C 높다"고 표현할 수 있으나 "부산의 낮기온 16^0C은 서울의 낮기온 8^0C보다 온도가 2배 온도가 높다"고 표현할 수는 없다.

(특징) 대다수의 대표값과 산포도가 가능하다. 그러나 절대영점(absolute zero point)이 없기 때문에 비율(ratio)이 지켜지지 않는다. 절대영점은 0의 값이 없는 것을 나타내느냐의 개념이다. 온도가 0^0C라면 없다는 이야기가 아니라 단지 얼음이 어는 기준을 나타낼 뿐이다

4 비율척도(비척도, ratio scale)

구간척도에 절대영점이 부여된 자료이다. 자료 사이에 차이뿐만이 아니라 비율도 비교할 수 있게 된다.
몸무게가 80kg인 사람은 몸무게가 40kg인 사람보다 몸무게가 40kg 더 나간다(차이)고 할 수도 있고 2배 더 나간다(비율)고 할 수도 있다.

(특징) 자료 사이에 차이 뿐만이 아니라 비율도 비교할 수 있게 된다.

명목척도와 순서척도를 묶어 범주형자료(categorical data) 또는 질적 자료 qualitative data)라고 부르고 구간척도와 비율척도를 묶어 수치자료(numerical data) 또는 양적 자료(quantitative data)라고 부른다. 수치자료는 관측 가능한 값이 연속적인 연속형 자료(continuous data)와 관측 가능한 값이 이산적인(관측 가능한 값을 셀 수있는) 이산형 자료(discrete data)로 구분하기도 한다. 자료가 어떤 종류이냐에 따라 자 료에 대한 통계분석 방법이 달라지므로 자료의 성격을 잘 파악하여야 한다

- 메모 -

03 신뢰성과 타당성

1 개념

조사결과는 조사대상 전체가 아니라 일부 즉 표본에 의거해 나온 결과이므로 어쩔 수 없이 표본오차, 즉 우리가 알고자 했던 성질과 표본에서 얻어진 값과의 차이인 표본오차가 항상 발생되기 마련이다. 통계적 신뢰성은 이러한 표본오차와 관련된 문제이다. 표본오차가 작으면 작을수록 조사결과는 신뢰성이 있다고 이야기 할 수 있는데 이러한 오차를 과학적으로 관리를 하기 위해서는 표본추출은 확률적으로 이루어져야 한다.

그러나 그 표본오차는 조사방법과 표본크기에 의하여 영향 받는다. 어떤 조사방법에 따라 표본이 추출 되었느냐에 따라 같은 비용과 시간으로도 훨씬 효율적인 결과, 즉 표본오차의 폭을 크게 줄일 수 있는 여지가 나온다.

그리고 표본의 크기는 크면 클수록 이런 표본오차의 크기는 줄어든다. 그러나 비용, 그리고 시간 관계상 표본의 크기를 무작정 크게 할 순 없다. 또한 표본의 크기를 늘린다 하더라도 오차의 폭이 그렇게 줄어들지 않는 현상이 벌어지는 시점이 온다. 즉, 비용과 신뢰성의 타협(트레이드오프, trade-off) 이 발생되는데 대략적으로 1,000명 내지 1,500명의 표본의 크기를 가지고 여론조사는 실시된다.

통계적 타당성이란 통계조사가 과연 의도하는 것을 측정하고 있느냐 하는 것임
예) 선거여론에서의 가장 큰 문제는 투표 참여의향과 관련된 추계를 하기 위해서는 투표할 의사가 전혀 없는 응답자가 보인반응을 각 후보 지지율에 산입해서는 안됨 (1) 우선 개별 응답자가 투표할 사람인지 그렇지 않은지를 알아낸후 (2) 각 후보의 지지율을 추정한다.

***참고**

어떤 경우는 조사대상수가 많아질수록 오차가 클 가능성이 높아진다. 미국 갤럽조사의 일반 여론조사에 있어서 대상자는 2천명 전후이다. 이 정도의 숫자만을 조사하여도 최근 4회에 걸친 대통령선거예측에서의 평균오차는 0.8%였다. 어떻게 이렇게 적은 수로 전 미국인의 의견을 정확하게 파악할 수 있는가? 이 원리는 조지-갤럽박사는 다음과 같이 쉽게 설명한다.

- "주부들은 냄비에 국을 끓여 간을 볼 때 잘 저은 뒤 한 두 숟갈 떠서 맛을 본다. 백배나 더 큰 가마솥에 국을 끓였더라도 잘 저어졌다면 백 숟갈이나 떠서 맛을 보지 않는다."
- "여기 흰 공, 검은 공 10만개를 7만개, 3만개의 비율로 잘 섞어놓고 또 다른 상자엔 천개의 공을 700개, 300개의 비율로 섞어놓았다. 각 상자에서 백 개씩 집어내도록 한다. 한쪽에 백배나 더 많은 공이 들어있는 것이 확실하나 집어내는 확률은 어느 쪽도 같은 것이다."

표본의 크기와 오차와의 관계를 보면 6백 명에서 1천2백 명으로 늘려 조사하면 오차가 4%에서 2.9% 로 감소하며 2천4백 명으로 늘릴 경우 2% 아주 적게 감소할 뿐이다. 이 경우 인구가 천만이든 억 만이든 관계없이 표본의 크기는 언제나 거의 일정하다. 이런 원리를 잘 이해하지 못하고 가능한 많은 수를 조사하는 것이 정확도를 재는 척도인 것처럼 오인되어 막대한 시간과 돈이 낭비되는 경우가 흔한 것이다.

2 표본크기의 결정

표본의 크기란 통계적으로 믿을만한 추정치를 얻기 위해 조사해야 하는 조사단위의 수를 의미하는데 통계조사에서 표본의 크기는 조사목적, 부분집단별 통계치의 필요성 여부, 전체적인 조사비용과 계획 등 여러 요인을 고려해서 결정해야 한다.

앞서 설명한 것처럼 표본크기가 늘면 표본오차는 줄지만 데이터 수집, 데이터처리, 분석 등 조사의 전 과정에서의 비용이 증가하게 된다. 거기에 표본크기가 늘어나면 조사원의 업무량과 조사과정에 대한 관리 감독도 어려워져서 표본조사에 따른 총 오차가 증가하게 되는 경우도 있다. 따라서 표본크기를 결정할 때는 전체적인 조사비용과 계획을 고려해서 결성하여야 한다. 또한 표본크기는 표본오차에 영향을 미치는 가장 중요한 요소 이지만 좋은 통계조사가 되기 위한 여러 조건 중의 하나라는 사실이다.

표본오차의 크기는 모집단 크기에 따라 좌우되는 것이 아니라 표본크기에 좌우된다는 점을 염두에 두어야 한다. 예를 들어 국회의원 선거구 한 지역에서 뽑은 200명의 표본을 조사한 경우와 우리나라 전체에서 뽑은 200명을 조사한 경우의 표본오차는 비슷하다.

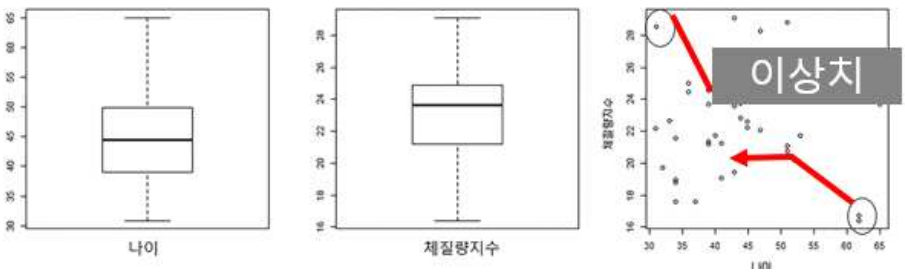
3 DataCleaning

- 결측치(missing values)

결측치란 일부 요인에서 관측값이 얻어지지 않은 것을 말하며, 이를 무시하고 관측된 자료만 분석하게 되면 편향(bias)이 발생할 수 있다. 따라서 결측치가 발생하면 적절한 방법에 의해 대체된 값으로 채워넣고분석을 하기도 한다.

- 이상치(outlier)

잘못 평가된 값으로 잘못된 분석결과를 초래할 수 있는 값을 의미한다. 즉, 데이터의 전체적인 패턴에서 벗어난 관측값을 말한다. 일반적으로 표본크기가 80 이하일 때는 표준화 점수가 2.5 이상이면 이상치이고, 표본크기가 80 이상일 때는 3~4면 이상치이다.

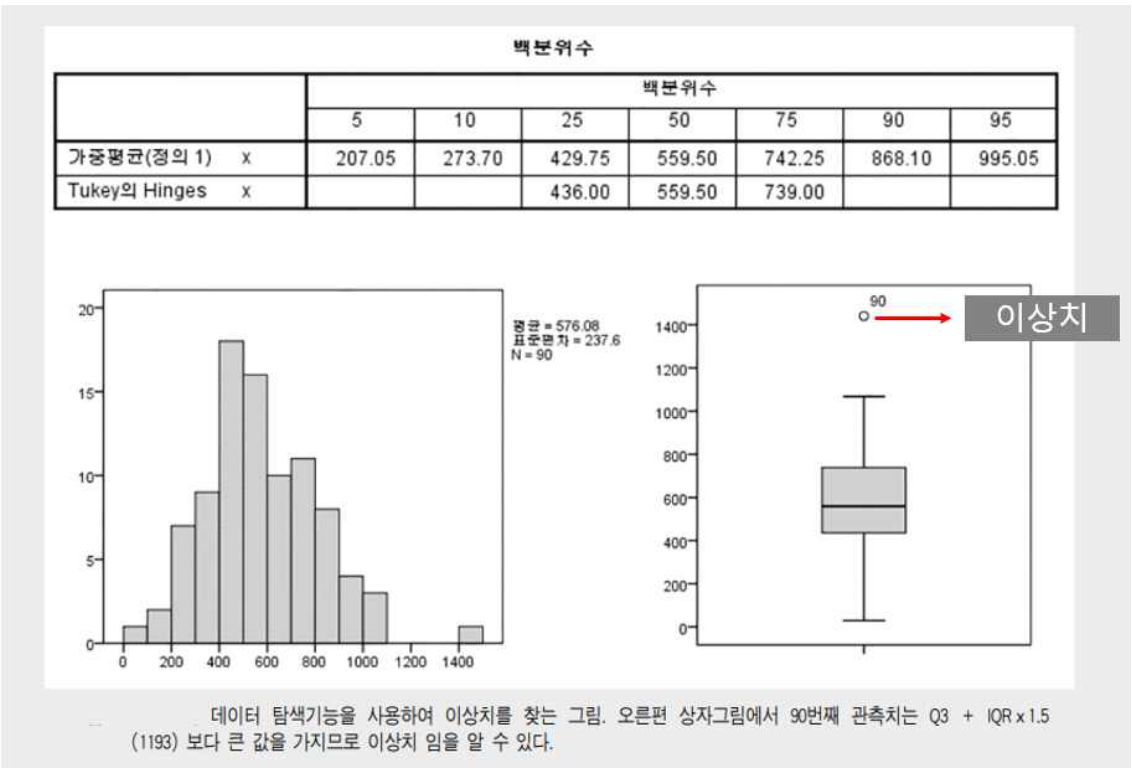


▶이상치 종류

이상치	설명
데이터 입력 오류	데이터를 수집하는 과정에서 발생할 수 있는 에러를 말한다. 예를들어 100을 입력해야 하는데, 1000 을 입력하면 10배의 값으로 입력이 된다. 이렇게 입력된 값은 전체 데이터의 분포를 보면 쉽게 발견을 할 수 있다.
측정 오류	데이터를 측정하는 과정에서 발생하는 에러를 말한다. 예를 들어 몸무게를 측정하는데, 9개의 체중기는 정상 작동, 1개는 비정상 작동을 한다고 가정 할 때, 한 사용자가 비정상적으 로 작동하는 체중계를 이용할 경우에 에러가 발생하게 된다
실험 오류	실험을 할 때 생기는 에러를 말한다. 100미터 달리기를 하는데, 한 선수가'출발'신호를 못듣고 늦게 출발했다고 가정하자. 이때 그 선수의 기록은 다른 선수들보다 늦을 것이고, 그의 경기시간은 이상치가 될 수 있다. 즉, 실험조건이 동일하지 않은 경우에 발생할 수 있다
고의적인 이상치	self-reported measures 에서 나타나는 에러를 말한다. 예를들어 음주량을 묻는 조사가 있다고 가정하자. 대부분의 10대 들은 자신들의 음주량을 적게 기입할 것이고, 오직 일부만 정확한 값을 적을 것이다. 이런 경우, 정확하게 기입한 값이 이상치로 보일 수도 있다
표본추출 에러	데이터를 샘플링하는 과정에서 나타나는 에러를 말한다. 대학 신입생들의 키를 조사하기 위해 샘플링을 하는데, 농구선수가 포함 되었다면 농구선수의 키는 이상치가 될 수 있다. 이것은 샘플링을 잘못한 경우이다

▶이상치 검출

이상치는 그림을 이용한 탐색을 통해서 발견할 수있다. 주로 상자그림(Box-plot), 히스토그램, 산점 도(Scatter Plot)를 사용하며, 수치적으로는 다음의 기준에 의해 이상치를 찾을 수 있다.



백분위수

	백분위수						
	5	10	25	50	75	90	95
가중평균(정의 1) x	207.05	273.70	429.75	559.50	742.25	868.10	995.05
Tukey의 Hinges x			436.00	559.50	739.00		

데이터 탐색기능을 사용하여 이상치를 찾는 그림. 오른편 상자그림에서 90번째 관측치는 Q3 + IQR x 1.5 (1193) 보다 큰 값을 가지므로 이상치 임을 알 수 있다.

▶이상치 제거

- 데이터가 정규분포를 갖고 있지 않는 경우

● 사분위수를 사용

(모든 데이터 표본 값을 갖고 있을 경우에 사용, 보다 정확함)

QUARTILE 함수 사용

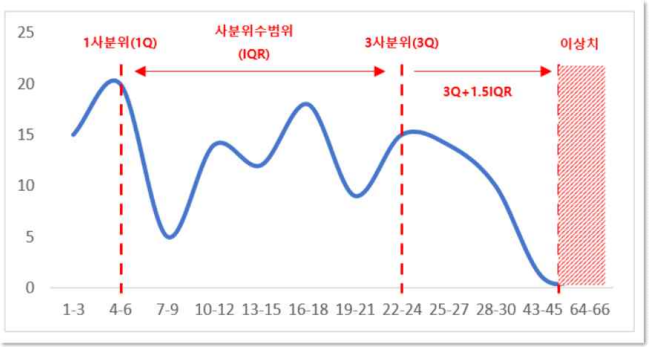
1Q = QUARTILE.INC(범위, 1) '1사분위 값을 계산

3Q = QUARTILE.INC(범위, 3) '3사분위 값을 계산

IQR = Q3 – Q1

최소유효값 = 1Q – 1.5*IQR

최대유효값 = 3Q + 1.5*IQR



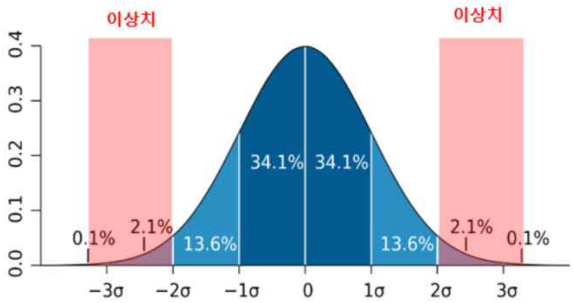
● 표준편차를 사용

(정규분포를 갖는 경우, 모든 가설검정은 정규분포 가정 하에 진행)

STDEV.S 함수 사용

표준편차 = STDEV.S (범위) '표준편차 계산

'평균 ±2*표준편차를 벗어나는 데이터는 이상치로 판단 후 제거



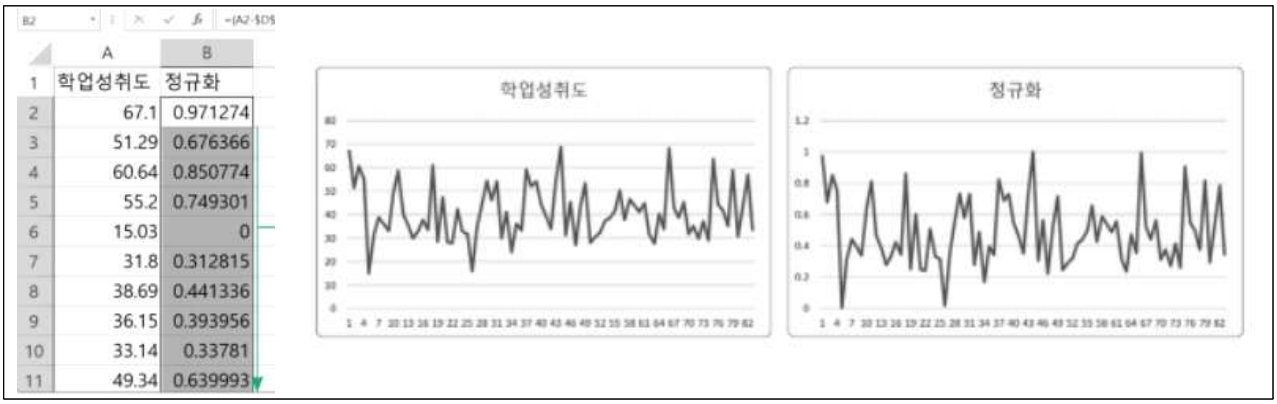
04 특성이 다른 데이터를 비교하는 정규화와 표준화

데이터는 특성에 따라 단위도 다르고 값의 범위도 큰 차이가 있을수 있다. 이러한 다름을 고려하지 않고 단순히 데이터의 수치만 비교한다면 제대로 된 비교를 할수 없다. 사람의 키가 180cm인 사람이 몸무게가 70kg인 사람보다 더 건장한 사람이라고 할수 없으며, 단위가 같더라도 범위가 크게 다르면 데이터간의 비교가 힘들다. (예: 100명중 99등과 1000명중 99등은 다르다.)
따라서 데이터를 제대로 비교하기 위해서는 이러한 특성이나 단위를 고려하지 않아도 되게끔 데이터의 상황, 즉 범위를 비슷하게 만드는 정규화와 표준화하 필요하다.

▶ 엑셀함수로 제공되지 않음

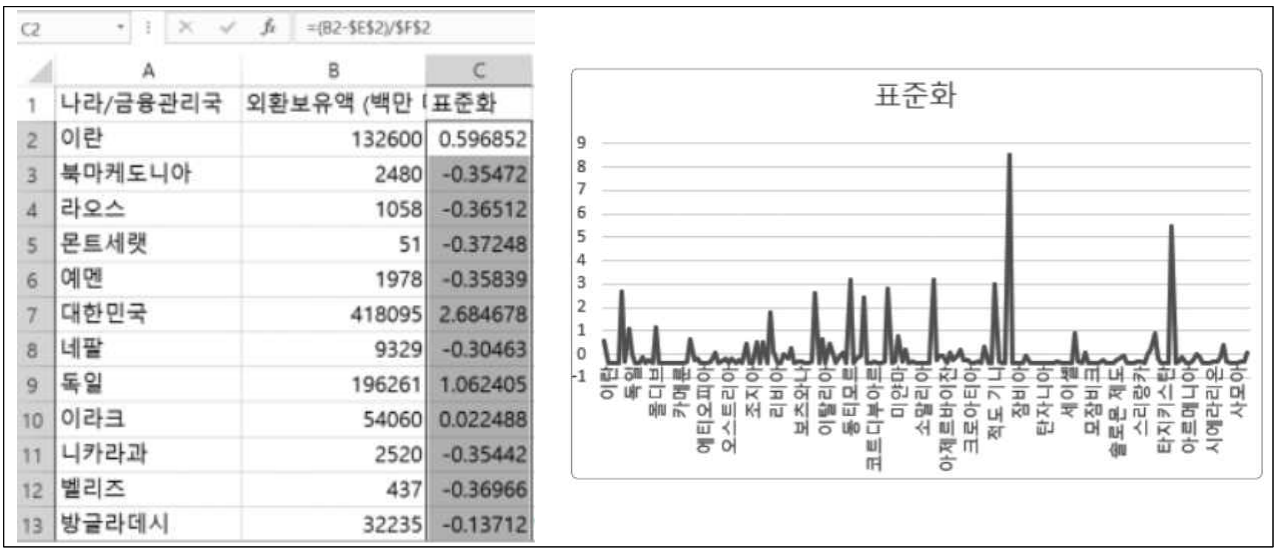
정의	개념	공식
정규화	데이터를 특정구간으로 변경(0~1값) - 보통 데이터 군 내에서의 특정 데이터의 위치를 확인하고자 할 때 사용 - 과거 대비 현재 데이터의 위치를 파악하기 용이 - 예) 과거의 하루 코로나 19 확진자 수에 비해 오늘 코로나 확진자 수가 어느정도 위치에 있는지를 확인할 때	$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$
표준화	데이터가 평균을 기준으로 얼마나 떨어져 있는가를 나타냄 어떤 특성의 값들이 종 모양의 정규분포를 따른다고 가정하고 값들을 평균0, 표준편차1이 되도록 함 - 평 균 0: 데이터비교쉬워짐 - 표준편차 1: 데이터간 간극이 줄어듬 - 표준화를 적용할 대상의 단위가 다를 때 데이터를 같은 기준으로 볼수 있게 함. - 예) 몸무게 데이터를 표준화하여서 평균0을 기준으로 몸무게가 음수값이 나오면 저체중, 양수값이 나오면 과제충으로 판단할수 있음	$Z = \frac{x - m}{\sigma}$ ▲ 표준화 = (요솟값 – 평균) / 표준 편차

01. ‘강의자료_확률통계.xlsx’ 의 [정규화] 워크시트를 선택한뒤 B열에 정규화 수식 작성 및 꺾은선 차트 작성



	A	B
1	학업성취도	정규화
2	67.1	0.971274
3	51.29	0.676366
4	60.64	0.850774
5	55.2	0.749301
6	15.03	0
7	31.8	0.312815
8	38.69	0.441336
9	36.15	0.393956
10	33.14	0.33781
11	49.34	0.639993

02. ‘강의자료_확률통계.xlsx’ 의 [표준화] 워크시트를 선택한뒤 B열에 표준화 수식 작성 및 꺾은선 차트 작성



=(B2-E2)/F2

	A	B	C
1	나라/금융관리국	외환보유액 (백만 (	표준화
2	이란	132600	0.596852
3	북마케도니아	2480	-0.35472
4	라오스	1058	-0.36512
5	몬트세랫	51	-0.37248
6	예멘	1978	-0.35839
7	대한민국	418095	2.684678
8	네팔	9329	-0.30463
9	독일	196261	1.062405
10	이라크	54060	0.022488
11	니카라과	2520	-0.35442
12	벨리즈	437	-0.36966
13	방글라데시	32235	-0.13712

05 표본데이터의 신뢰성 평가지표 – T분포, F분포

분석한 결과(가정)이 과연 타당한 값인가? 데이터 분석 도구를 사용하면 유의확률(p-value)값이 자동 계산된다.

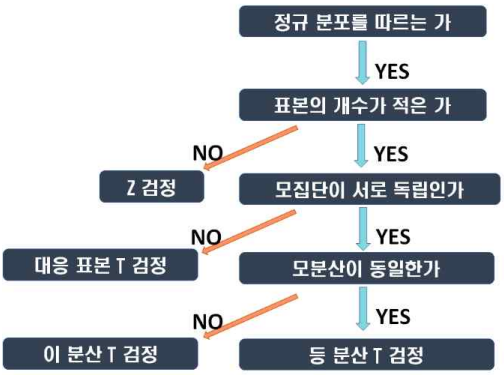
독립 표본 검정이란, 두 모집단의 관계가 서로 독립일 때 사용되는 검정이다. 독립 표본은 대립 표본에 비해서 조금 복잡한데, 그 이유는 표본의 갯수와 모분산의 동일성에 따라 검증 방법이 조금씩 다르기 때문이다.

▶표본개수충분->z검정

1. 표본 개수가 충분하고 , 모분산이 동일할 때
2. 표본 개수가 충분하고,
   모분산의 동일 여부를 알 수 없을 때

▶표본 개수가 불 충분할 때 T 검정

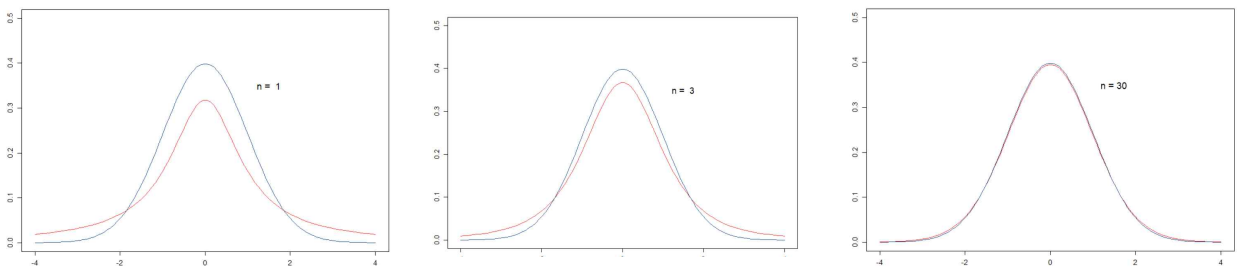
3. 표본 개수가 불충분 하고, 모분산이 동일할 때
4. 표본 개수가 불충분 하고,
   모분산의 동일 여부를 알 수 없을 때



1 T분포

독립 표본 T검정이란, 양적 자료 이고, 두 모집단의 관계가 서로 독립이며, 표본 갯수가 불충분할 때 사용하는 통계 검증 방식이다.

t 분포는 모집단 표준편차를 알 수 없을 때 표본 평균과 모집단 평균 사이 표준화된 거리를 설명하며,표본평균을 이용해 정규분포의 평균을 해석할 때 많이 사용함.
(구글검색, 't분포 기네스')

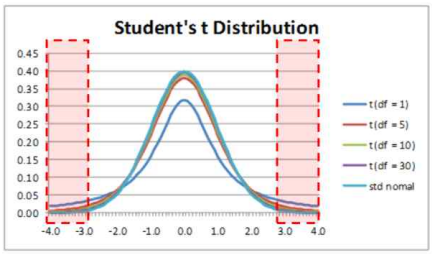


평균들의 표집 분포에서

t = (표본 평균 - 모집단 평균) / 표본 평균들의 표준 편차 = (m-M) / sm = (m-M) / (s/√n)

T분포가 클수록 ➔ 정규분포에 가까워진다!

2보다 크면 OK! **2보다 작으면 검토가 필요**할 수도 있다.



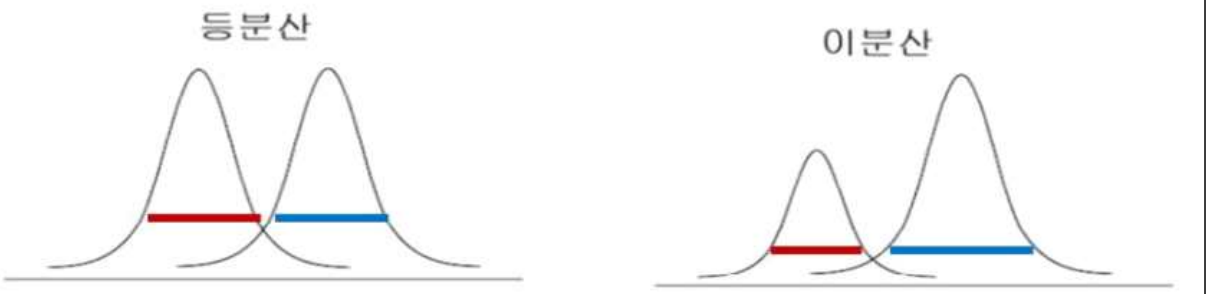
하지만, 데이터 분석도구를 사용하면 유의확률(p-value)가 자동으로 계산되기 때문에 크게 중요치는 않다.

T분포 = 계수 ÷ 표준오차로 계산되는데 계수(중요도)가 낮거나, 표준오차(오차량)이 크면 T분포는 자동으로 낮게 계산!

*먼저 F검증을 통해 등분산 여부를 확인한후 등분산 또는 이분산을 실행한다.

(F검정에서 P단측 검정값이 0.05보다 크면 등분산임)

https://free-chicken-forever.tistory.com/105



*두개의 변수값의 분산이 일정하면 등분산, 서로 다르면 이분산이라함.

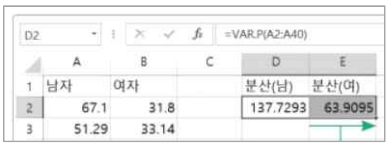
2 Z검정

- z검정은 모집단의 분산값을 알고 있을때 사용함

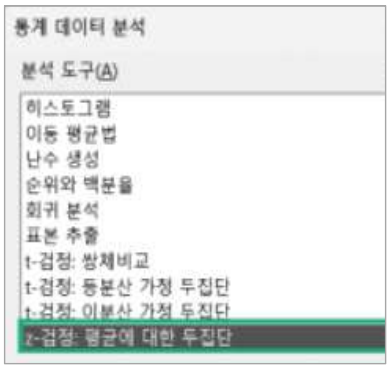
◐ 엑셀 [Z검정시트] 실행

- 'D1,E1 셀에 VAR.S 함수를 이용하여 남,여의 분산값을 계산
- [데이터-데이터분석]의 [z-검정-평균에 대한 두집단] 선택

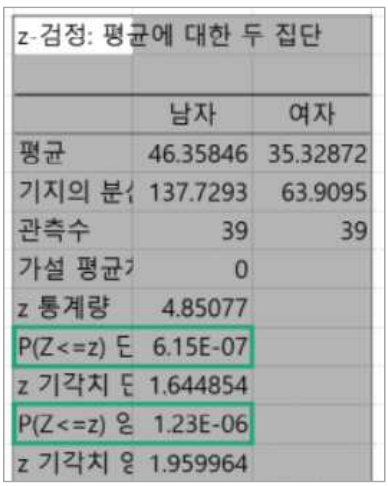
	A	B	C	D	E
1	남자	여자		분산(남)	분산(여)
2	67.1	31.8			
3	51.29	33.14			
4	60.64	29.93			
5	55.2	32.91			
6	15.03	37.72			
7	38.69	33.44			
8	36.15	28.48			
9	49.34	28.22			
10	58.55	27.89			
11	40.01	42.18			



	A	B	C	D	E
1	남자	여자		분산(남)	분산(여)
2	67.1	31.8		137.7293	63.9095
3	51.29	33.14			



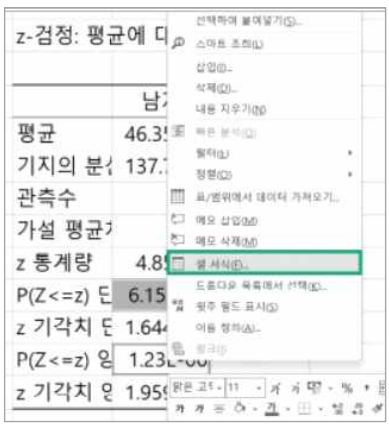
- 변수1은 남자 / 변수2는 여자
- 변수1의 분산은 남자 / 변수2의 분산은 여자
- 이름표선택 / 출력범위 설정
- 출력된 pvalue값에 대한 소숫점 자릿수를 더 자세히 표시하기 위하여 [셀서식]에서 소수자릿수 10



z-검정: 평균에 대한 두 집단

	남자	여자
평균	46.35846	35.32872
기지의 분	137.7293	63.9095
관측수	39	39
가설 평균	0	
z 통계량	4.85077	
P(Z<=z) 단	6.15E-07	
z 기각치 단	1.644854	
P(Z<=z) 양	1.23E-06	
z 기각치 양	1.959964	

-소수점 자리숫 출력 -



- 검정결과 확인

z-검정: 평균에 대한 두 집단

	남자	여자
평균	46.35846154	35.32872
기지의 분산	137.7293	63.9095
관측수	39	39
가설 평균차	0	
z 통계량	4.850769831	
P(Z<=z) 단측 검정	0.0000006149	
z 기각치 단측 검정	1.644853627	
P(Z<=z) 양측 검정	0.0000012298	
z 기각치 양측 검정	1.959963985	

위 실습을 통해 증명하고자 하는 바는 다음과 같습니다.

- **단측 검정 대립가설**: 여자 그룹보다 남자 그룹의 평균이 더 크다.
- **양측 검정 대립가설**: 어느 쪽일지는 모르지만 두 그룹의 평균에는 차이가 있다.

Z 검정 실습 결과 단측 검정과 양측 검정의 P값이 0.05보다 작으므로 '두 집단 간에는 유의미한 평균 차이가 없다.'(가설 평균차: 0)는 귀무가설이 기각되었습니다. 그러므로 위의 대립가설이 참이라는 결론을 얻을 수 있습니다.

3 T검정

- 모집단의 분산값을 모를 때 사용

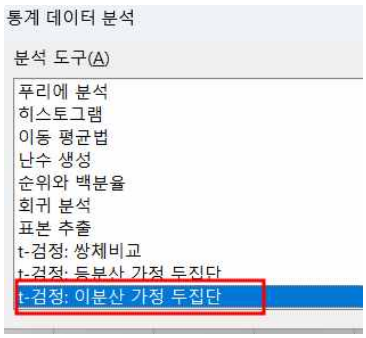
(1) 등분산가능한지 먼저 F검정을 수행한결과 0.01로서 0.05보다 작음으로 등분산이 아니다.

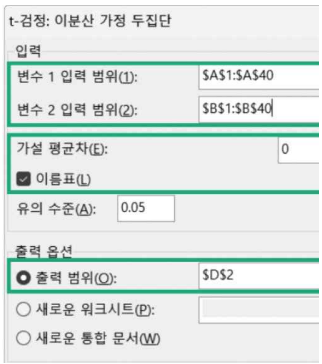
F-검정: 분산에 대한 두 집단

	남자	여자
평균	46.35846154	35.3287179
분산	141.3538028	65.5912799
관측수	39	39
자유도	38	38
F 비	2.15507005	
P(F<=f) 단측 검	0.010081454	
F 기각치: 단측	1.716687144	

(2) 엑셀시트: T검정(등분산 검정을 시행할수 있다)

	A	B
1	남자	여자
2	67.1	31.8
3	51.29	33.14
4	60.64	29.93
5	55.2	32.91
6	15.03	37.72





t-검정: 이분산 가정 두 집단

	남자	여자
평균	46.35846154	35.32872
분산	141.3538028	65.59128
관측수	39	39
가설 평균차	0	
자유도	67	
t 통계량	4.788176648	
P(T<=t) 단측 검정	0.0000048358	
t 기각치 단측 검정	1.6679161141	
P(T<=t) 양측 검정	0.0000096716	
t 기각치 양측 검정	1.9960083540	

T 검정(이분산) 결과 단측 검정과 양측 검정의 P값이 0.05보다 작으므로 '두 집단 간에는 유의미한 평균 차이가 없다.'(가설 평균차: 0)는 귀무가설이 기각되었습니다. 그러므로 다음과 같은 대립가설이 참이라는 결론을 얻을 수 있습니다.

- **단측 검정 대립가설**: 여자 그룹의 평균보다 남자 그룹의 평균이 더 크다.
- **양측 검정 대립가설**: 어느 쪽인지 모르지만 두 집단의 평균에는 차이가 있다.

4 분산분석

'분산분석(ANOVA분석)'을 할 때 사용되는 통계지표 * 분산분석 = 두개 이상 다수의 집단을 비교하는 분석 방법	$F = \frac{\text{집단간분산}}{\text{집단내분산}}$ * 분산 = 데이터가 얼마나 퍼져있는가?

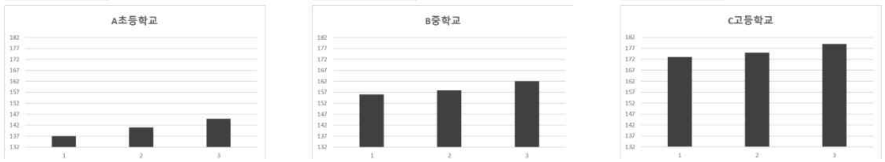
초등학교, 중학교, 고등학교에서
무작위 3명 학생의 키를 조사한 결과

➔ 집단간차이 > 집단내차이

A초등학교
137
141
145

B중학교
156
158
162

C고등학교
173
175
179



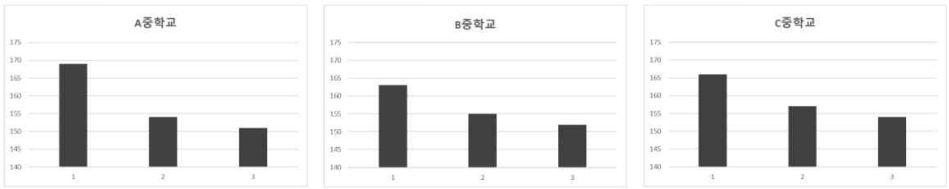
3개 중학교에서
무작위 3명 학생의 키를 조사한 결과

➔ 집단간차이 < 집단내차이

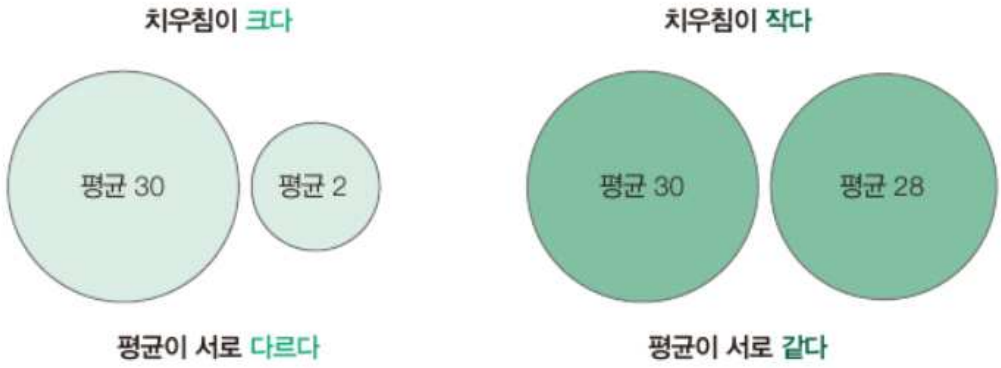
A중학교
169
154
151

B중학교
163
155
152

C중학교
166
157
154



▲ T 검정 vs. 분산 분석



▲ 분산 분석의 개념

분산 분석은 독립 변수의 개수에 따라 일원배치와 이원배치로 나뉘게 됩니다. 개념은 간단합니다. 이름으로도 어느 정도 유추할 수 있듯이 일원배치는 하나의 독립 변수를, 이원배치는 2개의 독립 변수를 이용한 분산 분석입니다.

	A	B	C	D	E
1	문화시설	여자 10대	여자 30대	여자 60대	
2	문화시설	26	53	75	
3	문화시설	26	54	73	
4	문화시설	25	58	78	
5	문화시설	24	57	84	
6	문화시설	28	64	81	
7	문화시설	29	66	91	
8	문화시설	33	70	70	
9	문화시설	33	65	64	
10	문화시설	24	53	56	
11					

	A	B	C	D	E
1		여자 10대	여자 60대		
2	KBS홀 [부산]	29	91		
3	광천문예회관	12	84		
4	국립극장	1	35		
5					
6					
7					
8					
9					
10					
11					

▲ 일원배치 분산 분석용 데이터와 이원배치 분산 분석용 데이터

06 자료의 연관성

1 관계성 측정

변수가 2개 이상일 때 변수 간의 관계를 알아보고자 한다면 제일 대표적인 방법은 산점도이다. 산점도는 특히 자료의 개수가 클 때 유용하며 x-축에는 y-축의 변수를 설명하는 변수를 지정하는 것이 좋다. 그리고 산점도를 통해 다음과 같은 사항을 시각적으로 확인하여 본다.

- 두 변수 간의 관계 형태 (선형, 비선형)
- 관계 강도의 크기
- 관계의 방향(음, 양의 방향)
- 이상점 유무

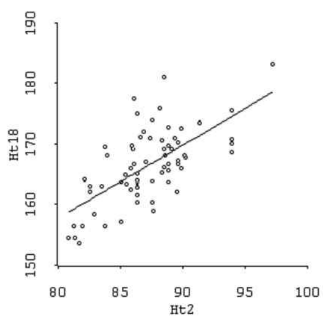
산점도에 두 변수의 관계를 잘 보여줄 수 있는 선을 덧붙여 놓기도 한다. 이와 같은 방법으로 세 가지가 제안되는데 모두 기본적으로 선과 자료점 간의 거리가 최소화하도록 그려야 한다.

- 직선
- 곡선을 나타내는 이차나 다항함수 곡선
- 평활선

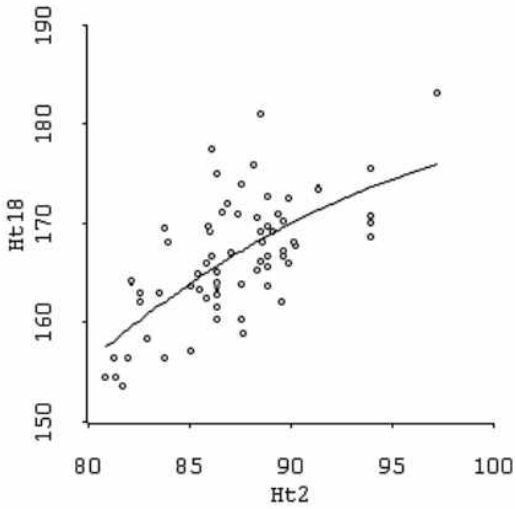
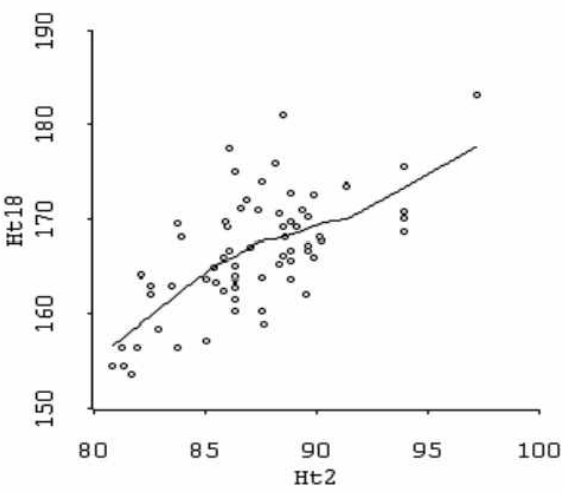
다음은 미국 버클리 대학에서 발표한 2세, 9세, 18세 때의 여성의 몸무게와 키의 자료이다. 몸무게는 wt2, wt9, wt18, 키는 ht2, ht9, ht18로 구분되어 있다. 예를 들어 wt2는 2세 때의 몸무게이고 ht2는 2세 때의 키이다. 몸무게는 kg으로, 키는 cm로 측정하였다.

다음 차트는 산점도에 두 변수 간의 관계가 직선이라 짐작하고 산점도에 직선을 적합시킨 차트이다. 자료의 x-축은 2세 때 키이며 y-축은 18세 때 키를 나타낸 것이다. 선은 직선을 적합 시켰을 때 생기는 거리의 제곱의 합을 최소화하는 선을 그린 결과이다. 이런 추정 방법을 최소제곱법이라 한다. 산점도의 결과에 의하면 대부분의 자료는 선을 따라 양의 방향으로 움직임을 알 수 있다.

	A	B	C	D	E	F
1	wt2	wt9	wt18	ht2	ht9	ht18
2	13.6	32.5	56.9	87.7	133.4	158.9
3	11.3	27.8	49.9	90	134.8	166
4	17	44.4	55.3	89.6	141.5	162.2
5	13.2	40.5	65.9	90.3	137.1	167.8
6	13.3	29.9	62.3	89.4	136.1	170.9
7	11.3	22.8	47.4	85.5	130.6	164.9
8	11.6	30	59.3	90.2	136	168.1
9	11.6	24.3	50	82.2	128	164
10	12.4	29.9	58.8	85.6	132.4	163.3
11	17	44.5	80.2	97.3	152.5	183.2
12	12.2	31.8	59.9	87.1	138.4	167
13	15	32.1	56.3	88.9	135.2	163.8



그러나 2세때 키가 비교적 크다 하더라도 18세에 도달 했을 때에는 키의 증가 속도가 다른 경우에 비해 둔화된다고 가정한다면 이차 함수 식을 이용하여 선을 적합하는 것이 좋을 것이다.
그리고 만약 무슨 선을 그어야 할지 확신이 서지 못하는 경우는 평활선을 그려 추세를 알아 볼 수 있다. 많은 통계소프트웨어는 사용자가 전문적인 지식 없이도 선을 그릴 수 있는 인터페이스를 제공한다. 평활선 차트에서 추세선이 잠시 중간부분에서는 잠시 멈칫 하다 다시 증가하는 형상을 보인다

이차함수식 적용	평활선
	

주의할 점은 자료가 통제된 실험계획에 의해 수집이 되지 않는 한 변수 간의 인과관계를 산점도가 알려주지는 않는다는 것이다. 다음 절에서는 직선에 의한 관계식을 가정하였을 때 두 변수 간의 관계를 요약하는 방법에 알아보도록 하자.

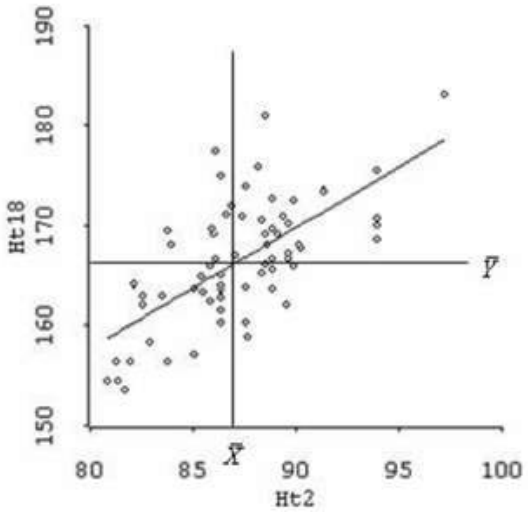
2 공분산 및 상관계수

산점도에 표시된 변수 관계의 크기와 부호를 정량화하기 위해서 만들어진 값들이다.

2-1. 공분산(Covariance)

미국 버클리 대학에서 발표한 키와 몸무게 자료를 이용한 공분산 자료에서 2세 때의 키와 18세의 키의 산점도를 다시 그렸는데 x,y 축과 더불어 $\bar{x}$ 와 $\bar{y}$ 를 새로운 축으로 가운데 설정하였다.

$$Cov(X,Y)=\frac{\sum_{i}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$



- 직선을 가정한 평면 4분할 -

$(x_i-\bar{x})(y_i-\bar{y})$ 의 쌍의 값은 (x_i, y_i) 쌍의 값이 어느 분면에 위치하고 있는 점이냐에 따라 그 부호가 바뀐다. 제 1사분면에 있다면 + 곱하기 + 의 부호인 양의 부호가 부여될 것이고 역시 제 4사분면에 위치하고 있다면 - 곱하기 - 가 되어 양의 부호일 것이다. 유사한 이유로 (x_i, y_i) 이 2사분면이나 3사분면에 위치하고 있었다면 음의 부호를 가질 것이다. 모든 n개의 $(x_i-\bar{x})(y_i-\bar{y})$ 의 값을 다 더하면 크기와 부호에 따라 양 혹은 음의 부호를 가지는 하나의 값을 가질 것이다. 이를 n-1로 나는 것이 공분산이다. 따라서 공분산 역시 평균, 분산과 마찬가지로 평균의 개념이 포함되어 있다. 엑셀에서는 변수 X와 Y의 공분산은 다음과 같은 명령문을

$$= cov(X,\ Y)$$

이용하면 된다. 그러나 공분산은 X의 원래 단위와 Y의 원래 단위의 곱형태이며 크기는 단위에 따라 값이 달라지기 때문에 직접 사용하기에는 문제가 있다. 따라서 이 단위에 상관없이 두 변수의 관계를 알아보는 방법은 없을까? 이것이 상관계수이다.

2-2. 상관계수(correlation coefficient)

- **분석될 데이터는 반드시 연속된 '숫자'여야 한다.**
- **상관관계 분석 전 적절한 가설이 우선 세워져야 한다.**

예) 김대리(35)의 업무성과가 이대리(32)보다 높은걸 보니, 나이와 업무지표가 상관이 있을까?
예) 나이가 아니라면, 연봉과 업무성과가 상관이 있을까?
잘못된 예) 마케팅부서와 영업부서가 상관이 있을까?

상관계수는 공분산을 각각의 변수의 표준편차로 나눈값으로 정의된다. 따라서 분자와 분모의 단위가 같아지므로 상관계수는 단위가 없는 상태가 된다. 그리고 이렇게 정의된 상관계수는 -1과 1 사이에서(-1과 1을 포함해서) 값을 가진다. 변수가 어떤 단위가 주어진다 하더라도 모든 상관계수의 값은 -1과 1 사이에서 값을 가진다는 의미이다. 상관계수는 $Corr(X, Y)$ 로 표기한다.

$$Corr(X, Y) = \frac{Cov(X, Y)}{\mathrm{sd(X)sd(Y)}}$$

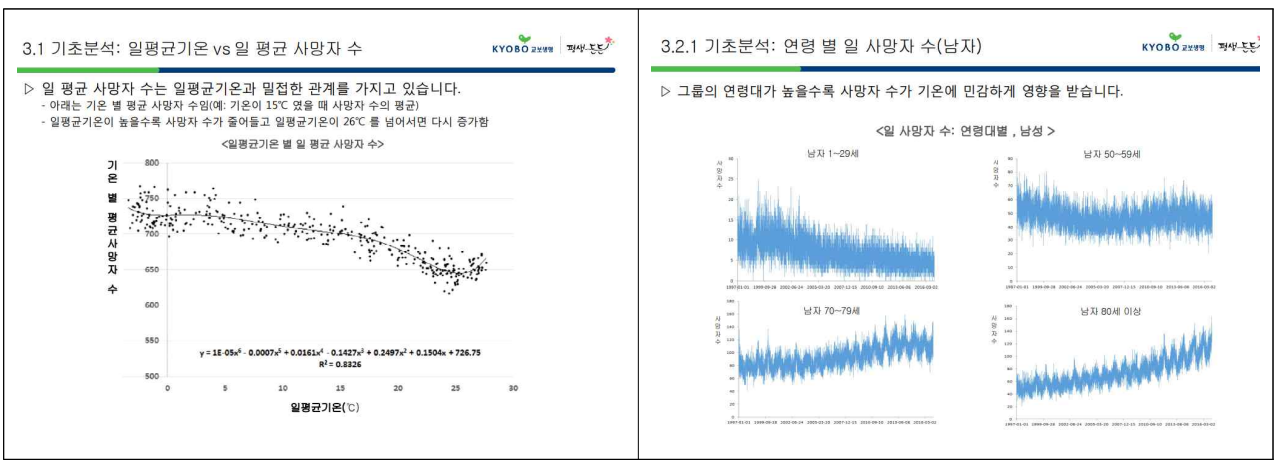
여기서 sd는 표준편차를 의미한다. 엑셀에서는 상관계수의 명령문은 다음과 같이

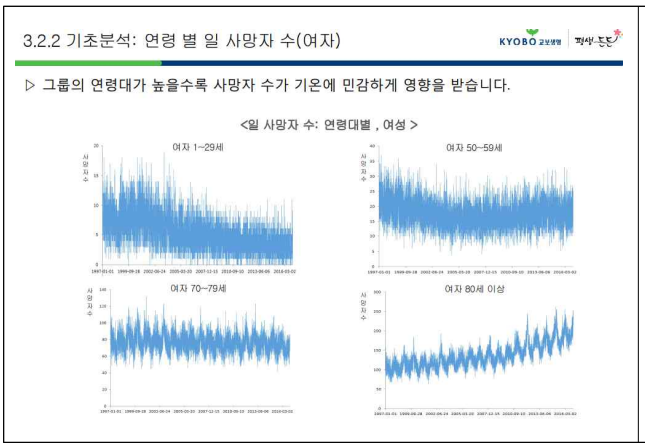
$$= correl(X,\ Y)$$

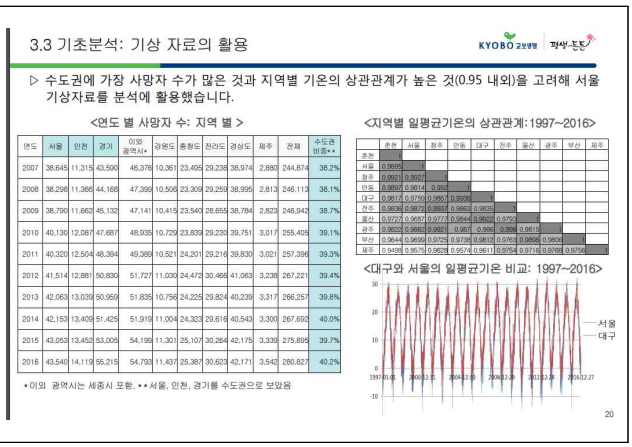
구현된다. 앞의 산점도에서 보았던 두 변수의 관계의 크기를 상관계수로 나타낸다면

$$Corr(X, Y) = \frac{Cov(X, Y)}{\mathrm{sd(X)sd(Y)}} = \frac{13.22923}{3.330523 \cdot 6.074886} = 0.663335$$

*(예1) 상관계수를 이용한 데이터 분석 사례

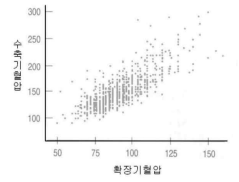


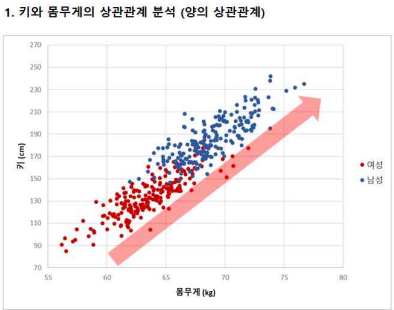


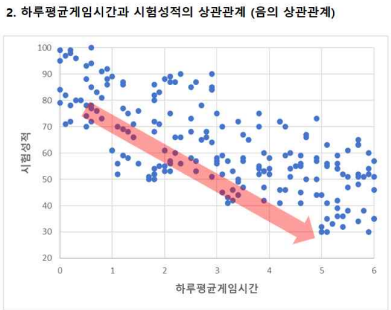


*(예2)

사람들의 확장기혈압(Diastolic BP (BP: Blood Pressure)) 과 수축기혈압(Systolic BP)에 대한 산점도이다. 표본상관계수를 구하면 0.792로 두 변수 사이의 직선관계가 강함을 알 수 있다.







▶ 상관계수의 두 변수 값의 관계 정도에 따른 상관계수의 값 (산업군에 따라 0.4가 강한상관일수도 있음)

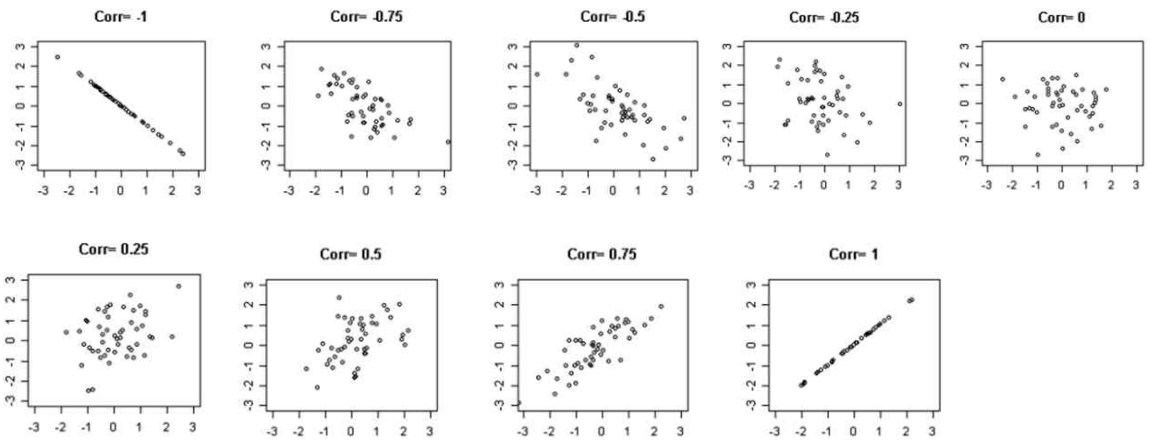
±0.81 ~ ±1.00 : 매우 강한 관계

±0.61 ~ ±0.80 : 강한 관계

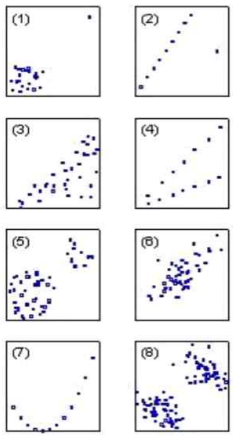
±0.41 ~ ±0.60 : 보통

±0.21 ~ ±0.40 : 약한 관계

±0.00 ~ ±0.20 : 관계 없음 (또는 매우 약한 관계)



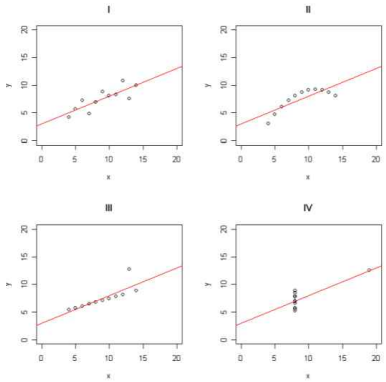
또한 같은 상관계수 값이라 하더라도 자료의 관계의 형태에 따라 그 의미를 동일시 해서는 안 된다. [그림 6.8]은 Cleveland, Kleiner, 그리고 Tukey의 책에서 인용한 그림인데 관계의 형태에 상관없이 상관계수는 0.7로 나온다. 이렇듯 상관계수는 산점도의 형태를 보지 않는 이상 두 변수와의 관계를 정확하게 반영하였다고 이야기 못한다.



[그림 6.8] 동일한 상관계수 = 0.7

4종류의 데이터셋에 대하여 산점도를 그렸을 때 표본상관계수는 모두 0.82 이고 추정된 회귀직선식은 모두 같았으나 점의 패턴이 모두 다르다.

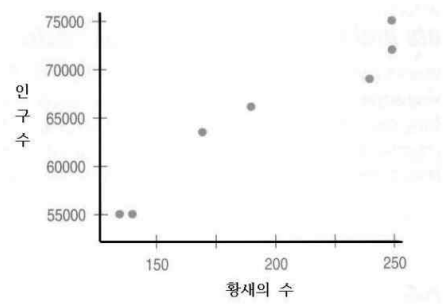
$$\hat{y} = 3 + 0.5x$$



[그림] 4 종류의 데이터셋에 대한 산점도과 추정된 회귀직선식

2-3. 허위상관

1930년대 독일의 Oldenburg시의 7년 동안의 황새수와 시 인구수에 대한 산점도이다. 표본상관계수는 0.97이어서 황새 수와 시 인구수에 상관관계가 매우 강하다고 할 수 있다. 그렇다고 황새 수가 많아지면 인구수가 늘게 된다고 인과관계로 이야기하면 안 된다. 이러한 상관을 허위상관(spurious correlation)이라고 한다. '황새가 아기를 물어온다'는 신화 때문에 황새가 잘 깃드는 나무를 심었던 역사가 서양에서 있음을 상기하면 통계의 오용이 우리의 삶에 영향을 주고 있음을 다시 한 번 느끼게 된다. 그러면 황새 수와 시 인구수에 상관관계가 매우 강하게 나타난 이유는 무엇일까? 황새 수와 시 인구수에 다 같이 관계를 갖는 숨겨져 있는 변수(예: 시의 발전성)가 있을 것이다.

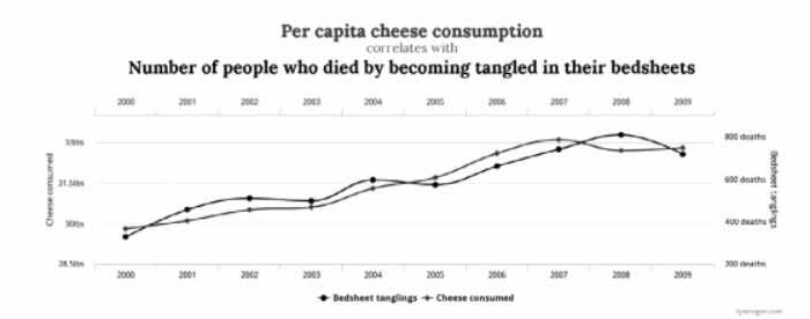


WT2	1					
WT9	0.692539	1				
WT18	0.392152	0.692089	1			
HT2	0.64455	0.522928	0.363717	1		
HT9	0.607125	0.727612	0.609332	0.738356	1	
HT18	0.445099	0.42605	0.497935	0.663335	0.807808	1

- 상관계수 행렬 -

미국의 한 조사 결과 대학 졸업 여부와 소득 사이에는 상관관계가 존재한다고 합니다. 즉, 대학 졸업자는 대체로 소득이 높은 경향을 보인다는 것입니다. 하지만 이러한 상관관계를 인과관계와 혼동해서 대학을 나와야 소득이 높아진다고 해석하는 것은 오류입니다. 소득이 높아서 대학에 진학하였을 수도 있고, 대학을 나오지 않더라도 높은 소득을 얻을 수 있기 때문입니다.

예) 담뱃값이 오르면 흡연률이 낮아진다. → 음의 상관관계 (인과관계는 아니다)
예) 정크푸드를 많이 먹는 청소년이 범죄를 저지를 확률이 높다. → 양의 상관관계 (인과관계는 아니다)



▲ 1인당 치즈 소비량과 침대 시트에 얽혀 죽은 사람들의 수(출처: www.tylervigen.com)

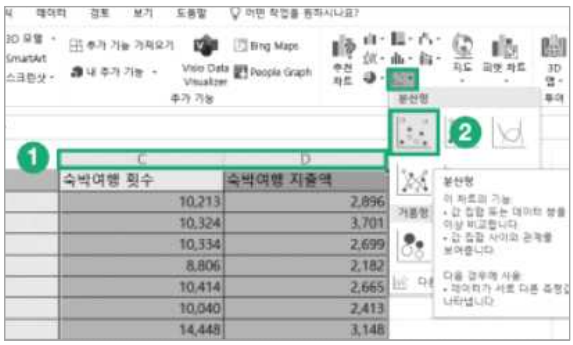
▲ 미국 메인주의 이혼율과 1인당 마가린 소비량(출처: www.tylervigen.com)

- 엑셀실습 -

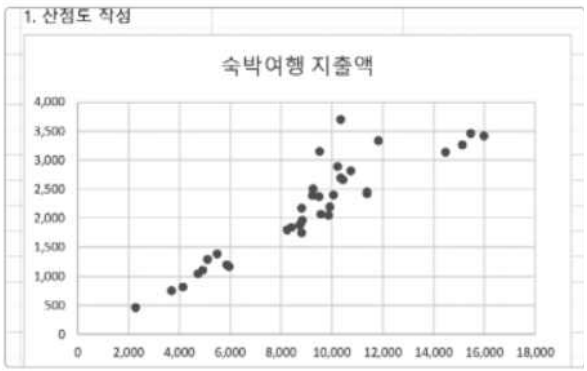
01. 공공데이터에서 제공되는 '월별_관광_숙박여행_횟수.xlsx' 자료를 작업에 필요한 형태로 전처리 함.

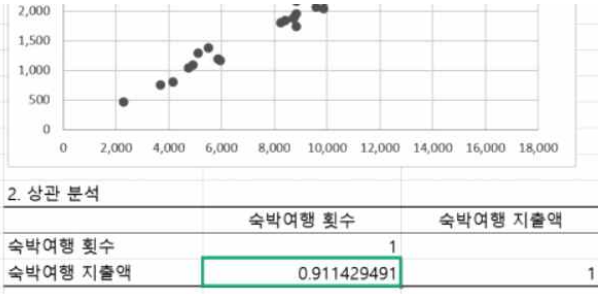
02. 숙박여행 횟수, 숙박여행 지출액을 범위지정후

03. [삽입]-[산점도 차트 클릭]



05. 데이터분석 - 상관관계를 이용하여 correl함수를 실행함.





2. 상관 분석

	숙박여행 횟수	숙박여행 지출액
숙박여행 횟수	1	
숙박여행 지출액	0.911429491	1

- 엑셀실습 -

01. [상관행렬_철강원자재가격동향] 시트를 이용하여 아래와 같이 상관행열을 구한뒤 조건부서식을 이용하여 시트를 완성

<원본>

	B	C	D	E	F	G	H	I
1	철광석($/톤	유연탄($/톤	철스크랩($	철스크랩(엔	철근(천원/	열연(천원/	후판(천원/	냉연(천원/톤
2	95	91	195	17750	648	660	650	718
3	84	111	202	17980	645	675	665	736
4	88	145	229	18525	596	700	695	740
5	87	142	236	19575	590	710	710	730
6	92	130	243	23060	591	702	702	716
7	91	121	222	22675	539	698	698	710
8	85	122	200	21093	568	690	690	713
9	89	125	189	21160	610	706	714	734
10	93	128	223	22850	648	728	728	748
11	90	140	242	24400	666	726	726	750
12	117	165	226	24420	687	720	720	752
13	111	179	238	27275	696	735	738	754

<완성본>

L	M	N	O	P	Q	R	S	T
	철광석($/톤)	유연탄($/톤	스크랩($/톤	스크랩(엔/톤	철근(천원/톤	열연(천원/톤	후판(천원/톤	냉연(천원/톤)
철광석($/톤)	1							
유연탄($/톤)	0.3941784	1						
철스크랩($/톤)	0.7106321	0.423016	1					
철스크랩(엔/톤)	0.5129033	0.459697	0.818101	1				
철근(천원/톤)	0.7209834	0.580767	0.778996	0.589344	1			
열연(천원/톤)	0.7024151	0.729263	0.756609	0.682333	0.866575	1		
후판(천원/톤)	0.8242514	0.390519	0.908496	0.682868	0.793082	0.749276	1	
냉연(천원/톤)	0.7868453	0.337229	0.876827	0.581709	0.771744	0.735897	0.901218	1

2-4. 카이제곱 검정

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

카이제곱검정 (Chi square test)

: χ^2 검정은 카이제곱 분포에 기초한 통계적 방법

: 관찰된 빈도가 기대되는 빈도와 유의하게 다른지를 검증

: 범주형 자료로 구성된 데이터 분석에 이용

: 카이제곱 값 $\chi^2 = \Sigma(\text{관측값} - \text{기댓값})^2 / \text{기댓값}$

: 카이제곱검정에는 두 가지 형태가 있으며, 같은 카이제곱 통계량과 분포를 사용하지만 다른 목적을 가짐

ㄱ) **Goodness of fit test (적합도 검정)**

-관찰된 비율 값이 기대값과 같은지 조사하는 검정 (어떤 모집단의 표본이 그 모집단을 대표는지 검정)

ㄴ) **Test of homogeneity (동질성 검정)**

-두 집단의 분포가 동일한지 검정

ㄷ) **Test for independence (독립성 검정)**

-Contingency table에서 있는 두 개 이상의 변수가 서로 독립인지 검정

-기대빈도는 두 변수가 서로 상관 없고 독립적이라고 기대하는 것을 의미하며, 관찰빈도와의 차이를 통해 기대빈도의 진위여부를 밝힘

-귀무가설 : 두 변수는 연관성이 없음 (독립)

-대립가설 : 두 변수는 연관성이 있음 (독립X)

: 이 때 일어날 법한 일인지, 희귀한 경우인지의 판단 기준은 confidence level 혹은 p value

예) 남녀성별에 따라 선호하는 공부장소에 차이가 있는가?

전제조건1) 연구가설의 종속변인은 범주형 자료여야 한다.

전제조건 2)기대빈도가 5이하인 셀이 전체의 20% 가 넘지 않도록함

해결방안1) 표본의 크기를 늘려서 빈도수 5이하인 셀을 제거함

해결방안2) 표본수를 늘림

해결방안3) 항목수를 줄임

<빈도표예시>

	독서실	집
남자	9	6
여자	6	7

예)

한 PC방 주인이 매주 똑같은 수에 손님이 온다고 했다. 이 가정을 검정해보자. 일주일간 온 손님 수를 관찰했다.

월	화	수	목	금
50	60	40	47	53

	A	B	C	D	E	F
1	요일	관측치	예상치	(O-E)^2/E		
2	월	50	50	0		
3	화	60	50	2		
4	수	40	50	2		
5	목	47	50	0.18		
6	금	53	50	0.18		
7			X^2	4.36	=SUM(D2:D6)	
8			p-값	0.3595	=CHISQ.DIST.RT(D7,4)	

4단계: 결과 해석하기

X^2은 **4.36**이다. 그리고 그에 상응하는 p-값(p-value)은 **0.3595**다. 이 p-값(p-value)은 0.05보다 크지 않기 때문에 귀무가설(null hypothesis)을 기각할 수 없다. 그러므로 관측치는 예상치와 다르다고 할만한 충분한 근거가 없다. 결과적으로 PC방 사장이 말한 매주 같은 수에 손님이 온다는 말은 틀린 말이라고 할 수 없다.

07 결정계수

결정계수는 회귀 모델에서 독립변수가 종속변수를 얼마만큼 설명해 주는지를 가리키는 지표이며 설명력이라고 부르기도 한다. 결정계수가 높을수록 독립변수가 종속변수를 많이 설명한다는 뜻인데 이 계수는 독립변수의 수가 증가하면 상승한다.

결정계수 : 회귀분석의 결과가 얼마나 정확한지 나타내는 지표

(상관계수의 제곱) = [0 ~ 1 사이의 값]

예를 들어, 결정계수가 0.703 이라면,

회귀분석으로 계산된 '**종속변수(y)에 대해 70%를 설명할 수 있다**'

라고 해석 가능하다.

결정계수 >= 0.65 : 유의미하다

결정계수 >= 0.5 : 쓸만하다

결정계수 < 0.5 : 검토 필요

엑셀의 회귀분석 결과로

1) 독립변수가 2개 이상이거나

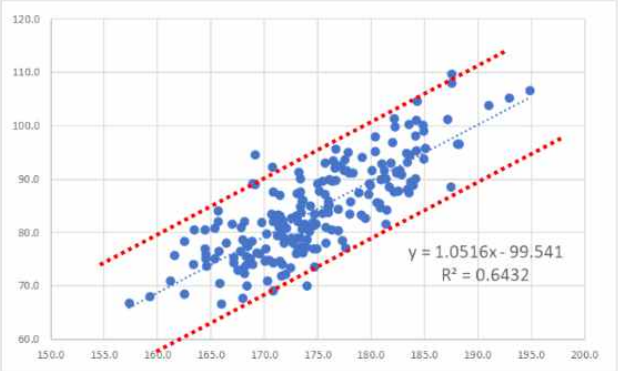
2) 표본의 개수가 200개 미만일 경우

조정된 결정계수를 보는 것이 바람직

(보수적 예측)

1	요약 출력	
2		
3	회귀분석 통계량	
4	다중 상관계수	0.860410573
5	결정계수	0.740306355
6	조정된 결정계수	0.709754161
7	표준 오차	29.12650434
8	관측수	20

15						
16		계수	표준 오차	t 통계량	P-값	하
17	Y 절편	123.2490003	2.743254964	44.92802	0.00000	1
18	몸무게	0.611626012	0.032373437	18.89284	0.00010	0
19						



몸무게를 독립변수(x)로 사용할 경우,

계산된 y값 대비 실제 값이 이상치 있을 확률이 0.0001 (0.01%) 이다

e.g. 표준오차가 5cm라고 가정하면..

빨간색 라인 밖에 있을 확률 = 0.01%