들어가기 전, 통계학이란?

통계학의 사전적 의미는 '통계치에 의미를 부여할 수 있는 합리적인 배경을 연구하는 학문'

통계학은 단순히 자료의 나열에 그치는 것이 아니라, 자료를 효과적으로 모으는 방법, 수집한 자료를 합리적으로 해석하는 방법 그리고 자료로부터 올바른 결론을 이끌어내어 적용하는 방법 등을 다루는 학문

통계학은 크게 두 개로 분류할 수 있다.
1. 기술통계학(descriptive statistics) - 어떤 자료를 단순하게 표현하는 방법
2. 추측통계학(inferential statistics) - 자료를 통하여 모집단에 대한 어떤 특성을 이반화하여 통계적 의사결정을 하는 방법 (= 서술통계학)

통계학을 배워야 하는 이유
1. 각종 수치 정보가 일상생활 곳곳에 있기 때문이다.
2. 통계적 방법은 작게는 개인이나 가정, 크게는 기업이나 국가의 운영에 필요한 의사결정에 중요하게 활용되며, 실제로 필요한 정보를 획득할 수 있다.
3. 각종 정보를 얼마나 믿을 수 있는지를 평가하는 정보의 질적 가치에 대한 척도를 제공

ⅰ <자료와 정보>

통계분석에서 가장 중요한 것은 자료이다.
자료는 수치로 표현되어 분석을 위해 가치 있게 사용될 수 있다.

1. 질적자료(qualitative data) 또는 범주형자료(categorical data)
- 자료 그 자체를 수치로 표현할 수 없는 자료
(ex) 산업, 채권, 지역분류, 성별, 생활수준, 종교성향, 교육수준

질적자료
1. 명목자료(nominal data) : 자료의 순서 또는 크기를 정의할 수 없는 형태
2. 순서자료(ordinal data) : 자료의 순서는 정의되나 크기가 정의되지 않는 형태

2. 양적자료(quantitative data) 또는 수치형자료(numerical data)
- 수치로 표현할 수 있는 자료
(ex) GNP, 경제성장률, 수출입현황, 임금, 키, 몸무게

양적자료
1. 이산자료(discrete data) : 관측값 사이에 공백이 있고 산발적으로 측정되는 자료
2. 연속자료(continuous data) : 지정된 구간 안에서 관측값 사이에 공백 없이 측정되는 자료

그러나 대부분의 자료는 그 자체로서 의미를 지니는 경우가 드물다.

자료는 통계처리과정을 거쳐 정보라는 의미 있는 형태로 전환되므로 조사자는 자료 수집 단계부터 통계적으로 어떻게 의미 있는 것으로 만들지 잘 설계해야 올바른 결과를 얻을 수 있음

<모집단과 표본>
통계조사의 목적은 자료에서 얻은 정보를 통해 심의 대상인 집단의 특성 또는 집단과 집단 간 관계의 특성을 파악하는 것

모집단(population)
통계분석에서 관심 대상이 되는 모든 사람, 응답 결과, 실험 결과, 측정값 등에 대한 전체 집합

모집단을 구성하는 자료 수에 따라
유한 모집단(finite population) | 무한 모집단(infinite population) 으로 구분
모집단을 이루는 자료의 개수를 모집단 크기(population size)라고 한다.

표본(sample)
통계분석을 위하여 모집단에서 실제로 추출한 관측값이나 측정값의 집합

표본추출(sampling) - 모집단에서 표본을 추출하는 일
표본 크기(sample size) - 표본에 포함된 자료의 개수
확률표본(random sample) or 임의표본 - 표본에 속한 각 관측값이 무작위로 선택된 경우
편향표본(biased sample) - 표본에 속한 각 관측값이 고의로 선택된 경우

전수조사(complete enumeration) - 통계조사에서 조사 대상이 되는 집단 전체를 조사하는 것
표본조사(sample survey) - 조사 대상이 되는 집단 전체에서 일부만을 뽑아 조사하는 것

모수(parameter) - 모집단의 특성을 나타내는 수치로, 모평균, 모분산, 모표준편차, 모비율 모집단의 자료 수를 나타내는 N 등이 있다.

통계량(statistic) - 표본의 특성을 나타내는 수치로, 모집단을 대상으로 하는 경우와 구별하기 위해 모집단과 다른 기호를 사용한다.
-표본평균, 표본분산, 표본표준편차, 표본비율, 표본 자료의 수 n

<기술통계학과 추측통계학>
기술통계학은 자료 수집, 정리, 표현과 관련된 것으로, 관심 집단의 특성을 수치적으로 기술하고 계산하는 데 필요한 통계적 기법과 관련된 분야
(이때 관심 집단은 모집단이든 표본이든 상관없다)
=>자료 집단의 여러 가지 특성을 기술하는 통계학이 기술통계학

추측통계학은 표본으로부터 얻은 정보를 이용하여 알려지지 않은 모집단의 특성을 추론하는 것
즉, 모집단에서 표본을 추출할 때 확률을 이용한 확률표본으로 만들어 통계적 모형으로 설정한다.

모집단 or 표본 -> 정리 요약 -> 특성 분석 (기술통계학)
표본 -> 통계적 모형 설정 -> 모형의 타당성 조사 -> 의사결정 (추측통계학)

통계 패키지(statistical package) - 컴퓨터로 자료처리와 통계분석을 할 수 있도록 다양한 통계분석 기법을 하나의 프로그램으로 정리한 것
(ex) SAS, SPSS, MINITAB, R 등 엑셀 활용
엑셀의 장점 – 대중성, 편의성, 호환성, 탁월한 통계처리 기능, 다양한 제반 기능

	질적자료		양적자료	
정의	자료의 크기를 정의할 수 없는 형태의 범주형 자료		자료의 크기와 순서를 정의할 수 있는 형태의 수치 자료	
종류	명목자료	순서자료	이산자료	연속자료
예	혈액형, 성별, 인종	학점, 부서 평가	자녀 수, 산업재해 발생 수, 대학별 취업자 수	키, 몸무게, 온도, 투자 수익률, 시장점유율
도수분포표의 종류	범주형 도수분포표		범주형 도수분포표 계급형 도수분포표	계급형 도수분포표

자료 배열이란
무질서하게 수집한 자료를 작은 값에서 큰 값 순으로 또는 큰 값에서 작은 값 순으로 정돈하는 것을 의미한다.
-한정된 양의 자료를 조직하고 파악하는 데 유용
-자료의 양이 많으면 배열 방법을 통해 얻을 수 있는 정보의 질과 양은 한정적

도수분포표(frequency distribution table)
자료를 체계적으로 정리, 요약하여 특성이나 구조를 파악하는 데 가장 간단하고 많이 사용하는 방법

범주형 도수분포표 : 자료의 성격을 둘 이상의 범주로 나누어 정리한 도수분포표
1)범주 설정
2)각 범주에 속하는 도수와 상대도수를 표로 정리
-도수 : 각 범주에 속하는 자료의 수 (f)
-상대도수 : 전체 도수에 대한 각 범주에 속한 도수의 비율(f/n or f/n * 100(%))

계급형 도수분포표 : 관측값을 일정한 가격으로 묶어 구간을 나누고, 각 구간에 속하는 자료의 수를 기록함

히스토그램
양적자료의 도수분포표에서 계급의 간격을 밑변으로 하고 계급의 도수를 높이로 하는 직사각형을 좌표평면에 차례로 나타낸 그래프
-연속적인 값으로 나타나는 자료를 표현할 때 사용하므로 막대그래프와 같이 막대의 순서를 마음대로 바꿀 수 없으며 주로 막대 사이에 간격이 없다.

도수분포다각형
히스토그램에서 각 계급 구간의 계급값의 빈도수를 직선으로 연결하여 그린 그림
-양끝은 도수가 0인 계급을 하나씩 추가하여 그 중점을 연결하여 그린 그래프

원그래프
-부채꼴의 중심각 = 360 * 상대도수

줄기-잎 그림
자료의 분포 형태를 쉽게 파악하면서도 각 관측값을 알 수 있는 그림

꺾은선그래프
연속적으로 변화하는 양을 점으로 찍고 그 점을 선분으로 연결하여 나타낸 그림

*퍼센트(%)
-두 개 혹은 그 이상인 수치의 상대적 크기를 명확하게 표현하기 위해서 주로 사용
보통 퍼센트를 백분율이라고 한다.
*퍼센트 포인트(%p)
-퍼센트를 직접 비교하는 경우, 기준이 같다면 퍼센트를 보통의 수치처럼 서로 더하거나 뺄 수 있다. 이때 두 퍼센트의 차이를 퍼센트 포인트라고 한다.

대푯값 : 자료 분포의 중심 위치를 나타내는 값
-산술평균, 중앙값, 절사평균, 최빈값 등이 있다

산술평균
-평균에는 산술평균, 기하평균, 조화평균 등 있음
-일반적으로 평균이라 하면 산술평균을 의미
-가장 보편적으로 사용하는 자료의 중심 위치를 나타내는 측도

표본평균
-변량 X에 대한 표본 n개의 자료가 x1,x2~xn 으로 주어질 때 변량 X의 산술평균
- 1/n(x1+x2+ ~ + xn)

모평균
-변량 X가 모집단에서 얻은 관측값 x1,x2 ~ xN으로 주어질 대, 변량 X의 산수령균

가중산술평균
-변량 X의 자료가 범주형 도수분포표로 주어질 때 가중산술평균을 정의한다.
-전체 도수를 n, I번째 범주에 속하는 도수를 fi라고 할 때

$$x = 1/n(f1x1 + f2x2 + \sim fnxn)$$ (k : 범주의 수)

산술평균의 성질
1. 산술평균에 대한 편차의 합은 0이다.
2. 산술평균은 편차의 제곱의 합을 최소로 한다. 즉, 산술평균에 대한 편차의 제곱의 합은 임의의 수에 대한 편차의 제곱의 합보다 크지 않다.
3. 산술평균은 주어진 자료를 모두 사용하므로 정보 손실이 없고, 특히 표본들의 평균인 표본평균은 모집단을 추론할 때 유용하게 사용된다.
4. 산술평균은 양적자료에 대해서만 구할 수 있으며, 대다수의 자료와 멀리 떨어져 있는 값인 극단값에 매우 민감하게 작용한다.(극단값은 이상점이라고도 한다)

중앙값(중위수)
작은 값부터 크기순으로 배열했을 때, 한가운데 위치한 값

최빈값
변량 X의 자료 중에서 가장 많이 나타나는 값
최빈값은 자료에 따라 두 개 이상일 수도 있고, 없을 수도 있다.

<산술평균, 중앙값, 최빈값 사이의 관계>
도수분포곡선이 단봉형으로 나타나고 극히 비대칭이 아닌 자료 집단의
표본평균(x), 중앙값(Me), 최빈값(Mo) 사이에는 피어슨의 실험공식이 설립한다.

도수분포가 완전히 대칭인 경우 : $x = Me = Mo$
도수분포가 오른쪽으로 치우친 경우 : $x < Me < Mo$
도수분포가 왼쪽으로 치우친 경우 : $x > Me > Mo$

백분위수와 사분위수
자료를 작은 값부터 크기순으로 배열했을 때, $0 \leq p \leq 1$인 p에 대하여 전체 자료를 100p%와 100(1-p)%로 나눈 값을 **제100p 백분위수** 라고 한다.
-자료 수가 n개일 때, 제100p 백분위수는 그 값보다 작거나 같은 자료 수가 np개 이상이고, 그 값보다 크거나 같은 자료 수가 n(1-p)개 이상인 값이다.
-자료 수 n에 p를 곱하여
np 정수 O : np번째로 큰 자료와 (np+1)번째로 큰 자료의 평균을 택
np 정수 X : np의 정수 부분에 1을 더한 값 m을 구하고 m번째로 큰 자료를 택

사분위수
-자료를 4등분하는 위치에 있는 값

-차례대로 Q1, Q2, Q3로 표시하며, 제1사분위수, 제2사분위수(중앙값), 제3사분위수 라고 한다.

절사평균
평균의 장점과 중앙값의 장점을 모두 고려한 대푯값으로 극단값을 제외하고 구한 평균
-전체 데이터 개수에 대하여 상위 몇 퍼센트의 데이터와 하위 몇 퍼센트의 데이터를 배제할 것인가로 절사비율을 결정한다.
- $0 \leq a \leq 0.5$ 인 a에 대하여 자료 수 n에 a를 곱하여
  an이 정수 O : 이 정수에 해당하는 자료 수만큼 양끝에서 제거
  an이 정수 X : an을 넘지 않는 최대 정수에 해당하는 자료 수만큼 제거

산포도
자료의 분포 상태를 알기 위해서는 자료의 중심 위치뿐만 아니라 자료가 흩어진 정도를 함께 고려해야 한다. 이와 같이 자료가 흩어진 정도를 산포도라고 한다.

범위 : 최댓값 – 최솟값

사분위수 범위 : 범위는 자료의 두 극단값의 차이만을 나타내기 때문에 자료의 산포를 나타내기에 불충분하다. 이러한 단점을 일부 보완한 산포도
사분위수 범위 = 제 3사분위수 – 제 1사분위수

분산 : 편차의 제곱의 평균
표준편차 : 분산의 양의 제곱근
분산 또는 표준편차가 작을수록 변량이 평균 주위에 모여 있음을 의미

변동계수
범위, 사분위수 범위, 분산과 표준편차 등은 자료의 흩어진 정도를 나타내는 척도
척도만으로 흩어진 정도를 비교하기 어려울 경우, 평균을 중심으로 상대적으로 흩어진 정도를 측정하는 척도를 사용

CV(변동계수) = 표준편차/평균 X 100(%)

변동계수가 크다는 것은 상대적으로 변동 폭이 크다는 사실 의미
변동계수는 보통 백분율로 나타낸다. 변동계수의 제곱은 상대분산이라고 함

5점요약표시
5개의 통계량 조합으로 나타내는 방법
[최솟값, 제1사분위수, 중앙값, 제3사분위수, 최댓값]

왜도와 첨도
-산술평균이나 표준편차의 값만으로는 분포 형태를 완전히 결정할 수 없다. 왜냐하면 평균

이 같고 표준편차가 같아도 분포 형태가 전혀 다를 수 있기 때문이다.

왜도 : 분포의 대칭이나 비대칭 정도를 표시하는 척도 (비대칭도)
a=0 (대칭) / a>0(왼쪽으로 치우침) / a<0 (오른쪽으로 치우침)

첨도 : 뾰족함의 정도를 나타내는 척도
b=3(표준정규분포와 같다) / b>3(정점이 높고 뾰족) / b<3 (정점이 낮고 완만)

상자그림 : 수치 자료를 표현하는 그래프
상자그림 작성 순서
1. 사분위수 값 Q1, Q3과 중앙값 Me를 결정한다.
2. Q1과 Q3을 상자 형태로 연결하고, 중앙값 Me의 위치에 선을 표시한다.
3. 사분위수 범위(Q3-Q1)를 계산하고, Q1과 Q3으로부터 각각 오른쪽, 왼쪽으로 1.5(Q3-Q1) 크기 범위 내의 인접 값을 실선으로 연결하여 표시한다.
4. 안 울타리로부터 1.5(Q3-Q1) 크기의 범위를 바깥 울타리로 표시한다. 안 울타리와 바깥 울타리 사이의 값을 보통 극단값이라 하고, 그 값이 존재하면 O으로 표시한다.
5. 바깥 울타리 경계를 벗어난 값을 *로 표시하고, 이 점을 극단값으로 판정한다.

ii 확률

표본공간
어떤 실험 또는 시행에서 일어날 수 있는 모든 가능한 결과의 집합

사건 : 표본공간의 임의의 부분집합
표본점 : 표본공간을 구성하는 개개의 원소
근원사건 : 표본공간의 부분집합 중에서 한 원소만으로 이루어진 사건
전사건 : 표본공간의 모든 원소를 포함하는 사건
공사건 : 표본점을 하나도 포함하지 않는 사건

이산 표본공간 : 표본공간에 속하는 원소 개수가 유한하거나 그 수를 셀 수 있는 경우
연속 표본공간 : 표본공간에 속한 원소가 실수 구간으로 이루어진 경우와 같이 수학적으로 셀 수 없는 경우

라플라스 확률(수학적 확률)
일반적으로 1회의 시행에서 일어날 수 있는 경우의 수가 N인 표본공간에서 각 근원사건이 발생할 가능성이 같을 경우, 사건 A가 일어나는 경우의 수를 n이라고 하면, A의 확률은

$$P(A) = n/N$$

콜모고로프의 확률(통계적 확률)
시행을 반복하면 한 사건이 일어나는 상대도수가 일정한 극한값에 수렴한다.
확률의 공리 : 각 근원사건이 일어날 가능성이 다른 경우와, 주관에 따라 정하는 주관적 확률의 개념도 수용할 수 있다.

*확률의 공리적 정의
1. 표본공간(S)에서 임의의 사건 A에 대하여 $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. A1, A2, ~ 가 각각 서로 배반사건이면
$P(A1 \cup A2 \sim) = P(A1) + P(A2) + \sim$
이 세 가지 조건을 만족하는 P(A)를 사건 A의 확률이라고 한다.

표본공간이 유한집합이고 각 원소가 일어날 가능성이 서로 같을 것으로 기대되면 수학적 확률이 공리를 만족하므로 수학적 확률로 확률을 구하면 되고, 그렇지 않으면 통계적 확률로 구하면 된다.

*확률의 성질
1. P(공집합) = 0 (공사건의 확률)
2. P(A의 여집합) = 1 – P(A) (여사건의 확률)
3. 임의의 사건 A, B에 대하여, A가 B에 속하면 $P(A) \leq P(B)$
4. 임의의 사건 A, B에 대하여,
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. 임의의 사건 A, B에 대하여,
$P(A \cup B) \leq P(A) + P(B)$

조건부 확률
사건 B가 일어났다는 조건 아래에서 사건 A가 일어날 확률을 사건 B에 대한 사건 A의 조건부 확률이라고 한다. P(A|B) 로 표기한다.
$P(A|B) = P(A \cap B)/P(B)$. $P(B)!=0$

조건부 확률의 공리
p(B)=/=0인 사건 B에 대하여 다음이 성립한다.
1. 임의의 사건 A에 대하여 $0 \leq P(A|B) \leq 1$
2. $P(B|B) = 1$
3. $A1 \cap A2$ = 공집합 이면 $P(A1 \cap A2 \mid B) = P(A1|B) + P(A2|B)$

곱셈정리
$P(A|B) = P(A \cap B)/P(B)$, $P(B)>0$
$P(B|A) = P(A \cap B)/P(A)$, $P(A)>0$

$P(A \cap B) = P(B) \times P(A|B) = P(A) \times P(B|A)$

독립사건과 종속사건
두 사건 A, B가 다음을 만족하면 두 사건 A, B는 독립이라고 한다.
$P(A \cap B) = P(A)P(B)$
그리고 독립이 아닌 두 사건을 종속이라고 한다.

복원추출과 비복원추출
복원추출 : 추출한 것을 제자리에 되돌려 놓고 다음 것을 추출하는 방법
비복원추출 : 추출한 것을 제자리에 되돌리지 않고 다음 것을 추출하는 방법

베이즈 정리
주어진 조건에서 어떤 사건이 실제로 일어날 확률을 산출하는 기법을 정리한 것으로, 불확실한 상황에서 의사결정 문제를 다룰 때 중요하게 사용한다.

공집합이 아닌 어떤 집합 A에 대하여 집합 A의 부분집합 A1, A2, ~, An을 원소로 하는 **집합족 {A1, A2,~, An}** 이 다음을 만족할 때, {A1, A2, ~,An}을 **집합 A의 분할**이라고 한다.

1. $A_i \subset A$ (i=1,2,~n)
2. $A_i \cap A_j$ = 공집합 (i=/=j , I, j = 1,2,3~n)
3. $A1 \cup A2,\sim,\cup An = A$

$P(A_i \mid B) = P(A_i \cap B)/P(B) = P(A_i)P(B|A_i)/\sum P(A_k)P(B|A_k)$ (i=1,2,3~n)

iii 확률변수와 확률분포
확률변수
어떤 시행에서 표본공간의 각 원소에 하나의 실숫값을 대응하는 함수 X

-확률변수에는 이산확률변수와 연속확률변수가 있다.

이산확률변수
확률변수 X의 치역이 셀 수 있는 이산 값으로 주어지는 확률변수 X
-이산점 : 확률변수 X가 갖는 각 값 x

연속확률변수
어떤 연속하는 범위 안에서 모든 실숫값을 가지는 확률변수 X

확률질량함수의 성질
-이산확률변수 X의 확률질량함수 f(x)에 대하여 다음 성질이 성립한다.

1. 모든 실수 x에 대하여 $0 \le f(x) \le 1$
2. $\sum f(x) = 1$
3. 임의의 $A \subset R$에 대하여 $P(X \in A) = \sum f(x)$

확률밀도함수
연속확률변수 X에 대하여 $a \le X \le b$일 확률을 다음과 같이 표현할 때, 확률변수 X는 연속확률분포를 따른다고 한다.
$P(a \le X \le b) = \int f(x)dx$

이때 연속함수 f(x)를 확률변수 X의 확률밀도함수라고 한다.

확률밀도함수의 성질
1. 모든 실수 x에 대하여 $f(x) \geq 0$
2. $\int f(x)dx = 1$
3. $P(a \leq X \leq b) = \int f(x)dx$

확률분포
[1] 이산확률분포
분포함수 : $F(x) = P(X \leq x)$,F(x)를 확률변수 X의 분포함수라고 한다.
(누적분포함수라고도 한다.)

분포함수의 성질
1. 모든 실수 x에 대하여 $0 \leq F(x) \leq 1$
2. F(x)는 증가함수이다.
3. $F(\infty) = \lim F(x) = 1$, $F(-\infty) = \lim F(x) = 0$
4. X가 이산확률변수인 경우, $P(X=x)=F(x)-F(x-1)$

확률변수의 기댓값과 분산

이를 통해 이산형확률분포의 기대값E(X)은
$E(X) = \sum xf(x)$ 이 됩니다.

연속형확률분포의 기댓값

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

$$E(Y) = E(g(x)) = \sum g(x)f(x)$$

$$E(Y) = E(g(x)) = \int g(x)f(x)dx$$

$E(aX+b) = aE(X)+b$ (a,b는 상수)
$E(X+Y) = E(X)+E(Y)$
$E(XY) = E(X)E(Y)$ (확률변수 X Y는 독립일 때)

$Var(X) = \sigma^2 = E[(X-\mu)^2]$ (* $\mu=E(X)$)
위 확률변수의 함수의 기댓값을 구하는 방법을
이용하면 아래와 같이 유도가 가능합니다.

$Var(X) = E[(X-\mu)^2] = \sum (x-\mu)^2 f(x)$ (이산형)

$Var(X) = E[(X-\mu)^2] = \int (x-\mu)^2 f(x)$ (연속형)

분산의 특징

$Var(aX+b) = a^2 Var(X)$ (a,b 상수)

Var(X) = E{(X-μ)^2} = E(X^2) - μ^2

분산의 성질

Var(k) = 0

Var(X±k) = Var(X)

Var(kX) = k^2Var(X)

적률생성함수

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

결합확률분포

공분산

두 확률변수 X와 Y가 가지는 값이 동일한 방향으로 변화하는지, 반대 방향으로 변화하는지의 결합 산포 정도를 나타내는 척도

상관계수

두 확률변수 X와 Y의 공분산은 확률변수의 단위에 따라 그 크기가 다르게 나타나므로, 공분산을 이용하여 두 자료를 비교하기 곤란할 수 있다. 따라서 단위와 무관하게 두 자료를 비교하기 위한 척도

이산균등분포
확률변수 X가 가지는 값(이산점)의 확률이 모두 같은 확률분포

베르누이시행
어떤 시행을 독립적으로 반복할 때, 발생할 수 있는 결과가 오직 두 개뿐인 경우에 대하여 다음 사항을 만족하는 시행

베르누이분포
f(x) = p^x(1-p)^(1-x) 와 같은 확률질량함수를 갖는 확률변수 X의 확률분포

이항분포
성공률이 p인 π회의 베르누이시행에서 성공횟수를 X라고 할 때, 확률변수 ε의 확률분포
f(x) = nCr p^r * p^(n-r)

이항분포 평균과 분산
확률변수 X가 B(n,p)를 따를 때,
평균 : E(X) = np / 분산 : Var(X) = np(1-p)

초기하분포
무한 모집단에서 표본을 비복원추출하거나 유한 모집단에서 복원추출하는 경우에는 베르누이시행의 조건을 만족하므로 이항분포를 사용할 수 있다.
그러나 유한 모집단에서 비복원추출할 때는 각 시행이 서로 독립이 아니므로 베르누이시행의 조건을 만족하지 못한다. 따라서 이런 경우에는 이항분포를 사용할 수 없다.
초기하실험에서 추출한 성공 개수를 X라 하면, 확률변수 X는 초기하분포를 이룬다고 하고, H(N,M,n)으로 표기한다.
주로 품질관리에서 사용한다.
불량률에 대한 함수를 곡선으로 나타내는 것 : 검사특성곡선

<표본추출법>
확률추출법 : 모집단의 개체가 표본으로 뽑힐 가능성이 모두 동등하다는 조건에서 객관적으로 표본을 추출하는 방법 (=임의추출법)
이때 확률추출법에 의하여 추출된 표본을 확률표본 또는 임의표본이라고 한다.
사실 표본을 선정할 때 가장 좋은 방법이다.

비확률추출법 : 모집단으로부터 표본을 추출할 때, 주관적으로 모집단의 대표를 정하여 표본을 추출하는 방법
이때 각 개체를 추출할 확률을 알 수 없으므로 표본으로부터 모집단에 대한 어떠한 결론을 내릴 때 신뢰 수준을 측정할 수 없다.

<확률추출법의 종류>
단순임의추출법 : 표본을 동등한 확률로 추출하는 방법

층화임의추출법 : 모집단을 성질이 비슷한 것끼리 층으로 분류하고, 각 층으로부터 임의로 표본을 추출하는 방법(평균소득을 조사할 때 고소득층, 중간층, 저소득층으로 나눈 후에 각 층에서 표본을 추출하는 방식)
-추정의 정도를 높일 수 있으며, 층별로 추정할 수 있다는 장점이 있다.

계통임의추출법 : 모집단으로부터 일정한 간격을 두고 추출하는 방법
임의로 1개의 숫자를 뽑아 I라고 할 때, 이 i를 임의출발점이라 하고, 다음과 같이 I에 차례로 k만큼씩 더한 수에 해당하는 모집단의 단위를 표본으로 추출하는 방법
-임의출발점을 정하고 그로부터 k번째마다 표본을 뽑는 방법
(이때, k를 추출간격이라 한다.)
-대규모 조사에서 주로 사용하는 추출법이며, 단순임의 추출법보다 추출 작업이 쉽고 표본의 정밀도가 높기 때문에 실제 조사에서 널리 사용한다.

집락추출법 : 단순임의추출법, 층화임의추출법, 계통임의추출법은 조사 단위 자체를 추출 단위로 하는 추출 방법이다. 이와 달리 단위 또는 집계 단위를 모은 집락을 추출 단위로 하는 추출법

-확률비례추출법 : 집락을 추출 단위로 하는 집락추출법 중 하나로, 집락의 크기가 현저히 다른 경우에 이용
->모집단이 크고 추출 단위의 리스트를 작성하기 어려운 경우, 집락을 추출 단위로 하면 추출 작업이 편리하다. 또한 단순임의추출법보다 조사비용이 크게 절약된다.

다단추출법 : 집락추출법에서 집락의 크기가 매우 클 때는 추출한 표본 집락에서 다시 표본을 추출하여 조사하는 것이 비용 절약, 정밀도 면에서 유리하다.
이러한 추출법을 2단 추출법 또는 부차추출법이라 한다.
이때 제 1단계의 추출 단위를 제1차 추출 단위 ~ 2단 추출법 이상을 다단추출법이라고 한다.
참고로 단순임의추출법, 층화임의추출법, 계통임의추출법은 1단 추출법이다.

모수 : 모집단의 특성을 나타내는 값
추정 : 모집단으로부터 추출한 표본에서 얻은 통계량을 이용하여 모집단의 특성을 나타내는 값인 모수를 파악하는 것
-모수의 추정은 일반적으로 표본으로부터 정의되는 함수를 통해 이루어지므로, 표본이 달라지면 추정 결과도 달라진다.

점추정 : 최적의 추정값을 구하는 과정
-점추정에 의한 모수의 추정은 표본이 어떻게 선정되느냐에 따라 잘못 추정하는 오류를 범할 수 있다.
구간추정 : 오류를 범하지 않기 위해 어느 정도 확신을 가지고 모수의 참값이 포함되었을 구간을 추정하는 방법

어떤 모수에 대하여 점추정할 때는 점추정량으로 표본평균, 중앙값, 최빈값 등을 사용할 수 있다. 이때 가장 좋은 추정량을 선택하는 것이 중요하다.
미지의 모수를 세타라 할 때 세타를 추정하는 데 통계량 세타^를 사용했다면, 세타^를 세타의 추정량이라 하고 세타의 통계 값을 세타의 추정값이라 한다.

좋은 추정량의 성질에는 비편향성, 유효성, 일치성, 충분성 등이 있다.
<추정량 종류>
1. 비편향추정량
2. 유효추정량
(모수 세타의 가능한 모든 비편향추정량 중 분산이 가장 작은 추정량을 최소분산 비편향추정량 또는 최량유효추정량 이라고 한다.)
3. 일치추정량 (추정량의 일치성을 밝히는 데에는 체비셰프부등식이 유용하다)
4. 충분통계량

<점추정>
점추정량을 구하는 방법에는 적률법, 최대우도법, 최소제곱법 등이 있다.
1. 적률법
-모집단으로부터 추출한 확률표본의 표본적률을 모집단에 대응하는 적률과 같게 둠으로써 방정식을 얻을 수 있다.
-이 방정식을 미지의 모수에 관해 풀어서 추정값을 구하는 방법을 적률법이라 한다.
2. 최대우도법
-우도함수와 결합확률밀도함수는 같으나 함수의 모수 세타의 함수로 볼 때 이를 우도함수라고 부른다.
-우도함수를 이용하여 주어진 표본의 값이 나올 확률을 최대로 만드는 추정량을 구하는 방법을 최대우도법이라 한다.

-일반적으로 모집단의 모수인 모평균, 모분산, 모표준편차, 모비율을 추정하는 추정량으로 보통 표본평균, 표본분산, 표본표준편차, 표본비율을 각각 사용하는데, 이들은 앞에서 살펴본 추정량의 성질을 대부분 만족한다.

<구간추정>
-구간추정은 미지의 모수 세타의 참값이 속할 것이라 믿는 구간을 추정하는 것