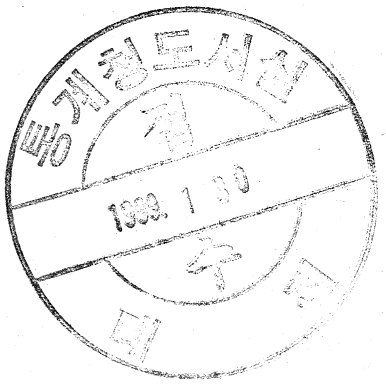
基礎統計教材

(記述統計)

1981



經濟企劃院 調査統計局

目　　次

Ⅰ章 : 序　　論

1.1 統計学의 발전

통계업무와 관련된 일에 종사하는 사람들과 가끔 대화를 나누다 보면 이들중 통계학이 무엇이고 통계학의 사회적 역할이 무엇인가에 대하여 간혹 그릇된 지식을 가진 사람들을 발견하곤 한다. 어떤이는 통계학은 물가, 통화량, 국민총생산과 같은 경제지수를 산출해 내고 여러가지 자료를 집계분류하여 도표를 만들고 평균치를 구해내는 학문으로만 생각하는 사람도 있고, 어떤이는 통계학은 수학의 일부분으로서 확률을 계산해 내고 기대값을 산출해 내는 응용산술이라고 생각하는 사람도 있다.

통계학이 생겨나게 된 근본적인 유래를 살펴볼 때 이와같은 견해를 가지게 되는 것도 무리는 아니다. 통계학이란 말은 원래 국가산술(State arithmetic)이란 말에서 나왔듯이 역사적으로 정치가들이 국가의 살림을 꾸려나가기 위하여 필요한 숫자를 체계적으로 산출해 내는 데서 유래하였다. 그리고 20세기초에 통계학이 발전되기 시작할 무렵, 많은 수학자들이 기여하였으며 수학의 한 부분으로서 발달하기 시작한 것도 또한 사실이다.

대부분의 학문이 그렇듯이 통계학도 역사적으로 발전과 변모를 거듭하였으며 현대적인 통계학은 근대적인 의미에서의 "국가산술"이나 "확률계산"의 영역을 벗어나서 의사결정과학(dacision —

making science)의 큰 몫을 차지하는 학문으로 성장하였다.

통계학의 올바른 정의는 사회, 자연 및 인간생활의 온갖 현상을 연구하기 위하여 불확실성(uncer tainty)이 내포된 자료의 선택, 관찰, 분석 및 추정을 통하여 의사결정에 필요한 정보의 획득과 처리방법을 연구하는 학문이라고 말할 수 있을 것이다. 따라서 자료처리 및 의사결정과학으로서 통계학의 기여는 인문, 사회, 경제, 자연, 의학등 광범위한 영역에 걸쳐 공헌하고 있으며 어떠한 분야이든지 과학적인 조사에서 얻어지는 자료에서 보다 유효한 결론을 도출하기 위하여 통계학에 관한 지식은 필요불가결한 것이다.

통계학의 가장 기본적인 요소는 관측의 결과로서 얻어지는 자료이다. 여기서 말하는 자료란 사회나 개인의 환경 속에 널리 흩어져 있는 모든 임의의 수치를 의미하며 이 자료를 적절한 통계적 방법을 통하여 처리하므로서 유익한 정보를 획득할 수 있는 것이다.

1.2 母集団과 標本

어떤 종류의 자료이든지 간에 그 자료가 얻어지는 대상이 있으며 이 자료대상이 되는 전체집단을 모집단이라고 부른다. 일반적으로 모집단의 크기는 방대하며 모집단 전체를 하나도 빠짐없이 관측한다는 것은 물리적으로나 경제적으로 불가능에 가까운 경우가 많다. 그래서 우리는 모집단 전체를 관측하는 대신에 이 모집단으로 부터 일부분을 고르게 추출하여 이 모집단을 대표하는 標本

(sample)으로 삼고 이 표본자료를 분석하여 모집단에 대한 정보를 얻고 기타 방법으로 취해진 정보와 함께 검토를 가하여 의사를 결정하고 모집단에 대하여 적절한 조처와 행동을 취하는 것이다. 이와같은 역학관계를 그림으로 그려보면 <그림 1.1>과 같다.

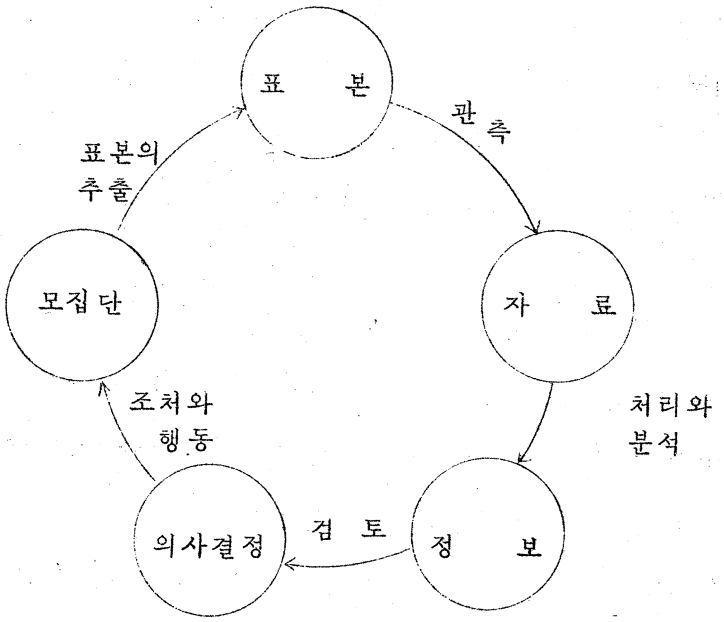


그림 1.1 모집단과 표본과의 관계

여기서 우리가 항상 염두에 두어야 할 사항은 표본을 관측하는 것은 표본에 대한 정보를 얻기 위함이라기 보다는 표본을 관측하여 얻어진 자료를 통하여 원래의 모집단에 대한 정보를 얻고 모집단에 대한 적절한 의사결정과 행동을 취하는데 목적이 있는 것이다. 즉 의사결정과 행동에 대한 대상은 어디까지나 모집단이지

표본이 아니다.

통계학은 여러가지 기준에 의하여 나누어 질 수 있으나 자료의 집단을 취급하는 방법에 따라 記述統計学과 推測統計学으로 대별할 수 있다. 기술통계학에서는 직접 관찰의 대상이 되는 집단이 모집단이든 표본이든간에 이 집단의 기술만을 목적으로 하며 평균, 분산, 비율 등의 통계정보의 획득에 관심이 있으며 이 이상의 일반적 적용은 고려하지 않는 경우이다. 이에 비하여 추측통계학에서는 직접 관찰하여 얻어진 자료는 단순히 표본자료에 지나지 않으므로 그 관찰결과에 근거를 두고 전체집단인 모집단의 성질을 추정하고 검정하려는 분야이다.

실제로 위의 두 분야는 상호 협조적인 관계에 있다고 보아야 할 것이다. 記述的인 방법에 의하여 자료를 연구함이 없이는 올바른 추측은 어려운 일이며 반대로 추측, 예측 등 추론의 여지가 없는 자료의 분석이란 가치가 없는 것이다.

통계학을 다른 각도에서 분류하여 어떤 이론을 대상으로 하는가에 따라서 나눌 수도 있다. 통계학을 수학의 입장에서 순수이론을 대상으로 하는 수리적인 연구의 경우는 數理統計学이라 하고 순수이론보다는 응용이론에 치중하여 통계이론의 응용화와 과학화를 대상으로 연구할 경우는 応用統計学이라 부른다.

모든 학문이 자료를 다루기 때문에 통계학의 응용이 미치지 않는 영역이 없으며 경제에 적용될 경우는 경제통계, 공학에 적용될 경우는 공업통계, 생물학에 적용될 경우는 생물통계, 농학에 적용될

경우는 농업통계라고 부른다. 이들 각 분야의 응용통계는 서로 연관성을 가지면서도 독특한 개별적인 발달을 이루어 왔기 때문에 오늘날의 통계학이 매우 다양화된 원인이 된다고 보아진다.

1.3 概 要

제일 먼저 이 책의 다음 장에서 다룰 내용은 주어진 자료의 정리와 표현방법이다. 기술통계학의 입장에서 이미 얻어진 자료집단에 대하여 어떻게 표나 도표를 만들어서 기술하는 것이 자료의 성질을 알기 쉽게 설명하는 방법인가를 알아보고 자료가 지니는 특성을 평균, 분산 등을 구하여 수치화 시키는 방법을 검토하여 보기로 한다.

제 3 장에서는 불확실성 연구의 바탕이 되는 확률과 확률변수의 분포에 대하여 설명하고 통계적 분석에서 자주 나타나는 확률분포 이론과 응용에 관하여 기술하기로 한다.

제 4 장부터서는 推測統計学의 바탕을 이루는 표본이론에 대한 기초적인 내용을 소개하기로 한다. 주로 표본평균과 표본비율의 분포와 성질에 대하여 관찰하여 보기로 한다.

계속하여 5 장에서는 실제로 모집단으로부터 표본을 추출하는 여러가지 경우와 방법에 대하여 설명하고 각 방법에 있어 標本의 크기를 구하는 방법에 대하여 고찰한다.

다음 장에서는 정부통계에서 널리 사용되는 지수(index number)의

의 산출과 사용법에 관하여 연구하기로 한다.

마지막으로 7장에서는 응용통계학에서 깊이 연구되고 적용되어가는 통계적 기법들인 분산분석을 통한 자료의 변동에 관한 분석, 변수간의 함수관계를 규명하는 방법이며 경제통계에서 많이 연구되는 상관 및 회귀분석 등이다.

이 책에서 취급하는 내용은 앞에서 언급한 바와같이 주어진 자료의 정리와 표현방법을 간략히 기술하고, 중점적으로 강조되는 내용은 주어진 자료가 속해 있는 모집단의 특성을 규명하고 의사결정을 위한 정보를 획득하기 위하여 적은 규모의 표본자료를 정리하고 분석하여 추측하는 방법들이다. 이러한 방법들은 한데 묶어서 통계적 방법이라고 부른다.

이 책은 통계학에 깊은 교육적 배경이 없는 독자를 위하여 만들어 졌으며 통계학의 이론을 공부하는데 역점을 두지않고 이 책의 목적은 기초적인 통계적 방법들을 소개하여 독자들로 하여금 통계학에 대한 개념을 갖도록 유도하고 실질적으로 통계적 방법을 자료분석에 적용할 수 있는 기본적인 능력을 길러주는데 있다.

Ⅱ章 : 資料의 整理方法

2.1 度數分布

우리는 항상 우리의 일상업무 가운데에서 여러가지 형태의 자료를 다루게 된다. 어떤 대상을 관찰한 수치들이 정리되어 있지 않은 상태로 있는 수집된 자료를 원자료(raw data)라고 하고 이러한 원자료를 일정한 기준에 의하여 정리하여 표로 만들어 놓은 것을 統計票라고 한다. 어떤 일정한 기준에 의하여 전체자료가 포함되는 구간을 여러개의 급구간으로 분할하고 자료를 분할된 급구간에 따라 분류하여 양적 분류표를 만들어 놓은 것을 度數分布票라 한다.

실제 예를 들어보기로 하자. <표2.1>은 어떤 공장에서 만들어지는 40왓트 110볼트의 특수 전구들 중에서 전구의 수명을 알아보기 위하여 64개의 전구를 적절한 방법으로 임의 추출하여 표본으로 뽑아낸 것이다. 임의추출 방법에 관해서는 뒤에 상의하기로 한다.

<표 2.1>과 같은 자료는 원자료이며 이처럼 정리되지 않은 자료는 무질서한 수의 나열에 불과하며 아무런 가치가 없는 것이다. 우선 이 자료에서 우리가 알기를 원하는 것이 무엇이 있는가를 생각해 보자.

(가) 전구의 수명은 대개 어떠한 값을 중심으로 분포되어 있는가?

<표 2.1>

40-watt 100volt 특수전구의 수명자료

(단위 : 시간)

1,310	1,262	1,234	1,104	1,105	1,243	1,204	1,103
944	1,343	932	1,055	1,303	1,185	759	1,404
1,248	1,324	1,000	984	1,381	816	1,067	1,252
1,093	1,358	1,024	1,240	1,220	972	1,022	956
1,690	1,302	1,233	1,331	1,157	1,415	1,385	824
1,229	1,079	1,176	1,173	1,109	827	1,209	1,202
609	985	1,233	985	769	905	1,490	918
1,028	1,122	872	826	985	1,075	1,240	985

Source : D.J.Davis,"An Analysis of Some Failure Data," Journal of the American Statistical Association,June 1952,P.142.

(나) 전구의 수명은 어느 일정한 구간에 어느 정도 분포되어 있고 전체적인 변동범위는 어느 정도인가?

(다) 전구의 수명은 회사가 원하는 회사규격에 맞는가? 또한 소비자가 원하는 표준규격에 어느 정도 만족을 시키는가? 만일 그렇지 못하다면 불합격품의 비율은 어느 정도인가?

등등 우리가 원하는 정보는 여러가지가 있다.

위의 세가지 질문에서 공통성을 이루는 개념은 변동에 관한 개념이다. 이 변동에 관한 개념은 통계학을 이해하는 가장 기본적인 요소이다. 앞의 전구 수명의 예를 들어서 변동에 관한 개념을 설명하여 보자. 전구의 수명을 좌우하는 요소를 열거하여 볼 때 생산과정의 원자재가 되는 금속재료의 품질, 작업원의 숙련도, 생산설비의 품질 등을 들 수 있을 것이다.

하나하나의 전구를 생산하는 과정에서 아무리 표준화된 생산공정을 사용하고 균일한 금속재료를 사용하고 작업원의 숙련도를 일치시키도록 노력한다 할지라도 전반적인 생산과정을 통하여 전구완제품을 만들기 까지에는 완전히 제거할 수 없는 변동적인 요소가 수반되기 마련이다. 이러한 변동적인 요소가 전구의 수명에 차이를 준다. 그리고 생산된 전구의 수명을 측정하는 과정에서도 어느 정도의 측정오차는 불가피하다. 따라서 앞서 나열한 여러 질문에 대답하기 위하여 <표2.1>의 원자료를 정리할 필요성이 있다.

첫단계로 측정치들의 변동성을 파악하기 위하여 <표2.2>와 같은 度數分布票를 만드는 일이다. 원자료를 조사해 보니 최대치가

<표 2.2>

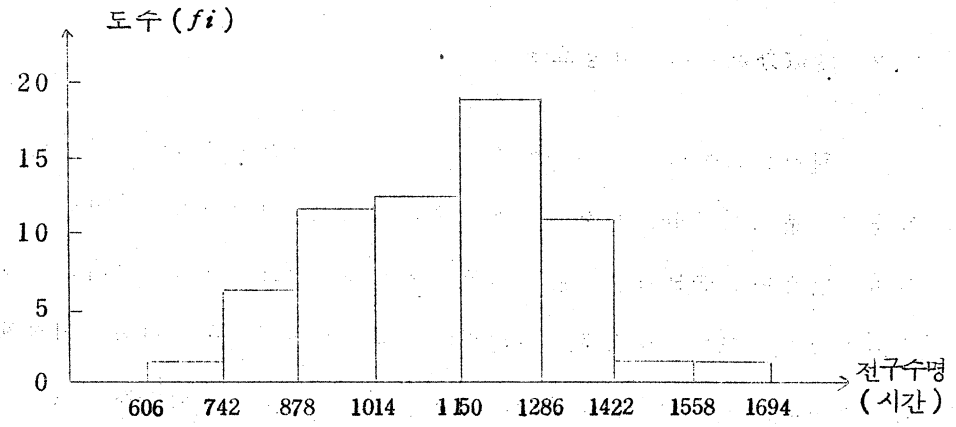
전구수명 측정치의 도수분포표

급의 No	급 구 간	중앙치($\bar{x}i$)	도수체크	도 수(fi)	도수분포율(%)
1	606~ 742 미만	674	/	1	1.6 %
2	742~ 878 미만	810	𝍸 //	7	10.9 %
3	878~1014 미만	946	𝍸 𝍸 //	12	18.7 %
4	1014~1150 미만	1072	𝍸 𝍸 ///	13	20.3 %
5	1150~1286 미만	1218	𝍸 𝍸 𝍸///	18	28.1 %
6	1286~1422 미만	1354	𝍸 𝍸/	11	17.2 %
7	1422~1558 미만	1490	/	1	1.6 %
8	1558~1694 미만	1626	/	1	1.6 %
합 계				64	100 %

최대치가 1690이고 최소치가 609이며 측정치의 수가 64이므로 편리하게 나누어서 급의 폭을 136으로 하고 급의 총수를 8로 하여 등간격 구간으로 나누고 각 구간에 몇개의 측정치가 들어가는가를 <표2.2>에서와 같이 세어 보았다.

이 도수분포표를 기둥그림으로 사용하여 도시하면 <그림 2.1>이 되는데 이런 그림을 度數分布図 또는 히스토그램이라 한다.

이와같이 원자료를 도수분포표나 히스토그램으로 나타내면 한눈에 전구수명 측정치의 분포상태를 알 수 있다. 이와같이 측정치의 값이 어느 급구간에 어느정도의 도수를 가지고 분포되어 있는가를 알아보는 일을 도수분포를 확인하는 작업이라 말할 수 있다.



<그림 2.1> 전구수명 측정치의 히스토그램

위의 도수분포나 히스토그램으로부터 전구의 수명이 대략 742시

간에서 1422시간 사이인 것을 알 수 있고 이 사이의 도수분포율을 합치면 10.9 + 18.7 + 20.3 + 28.1 + 17.2 = 95.2 으로 표본중에서 95.2 %가 여기에 속하는 전구수명을 가지고 있다.

전체적인 변동범위는 <표 2.2 >에서 보면 606 시간에서 1694시간까지이나 대부분이 742시간에서 1422시간 사이의 수명을 갖는 것을 알 수 있다. 만약 소비자가 원하는 전구의 수명이 1000시간 이상이라면 어림잡아서 20.3 + 28.1 + 17.2 + 1.6 + 1.6 = 68.8 %의 전구가 소비자의 요구를 만족시켜 주고 있음을 알 수 있다.

즉 소비자의 입장에서는 31.2 % 정도가 불합격품이 되는 셈이다. (<표 2.2 >에 의하면 정확히는 1014시간 이상이어야 함)

2.2 度數分布表의 작성요령

원자료로부터 도수분포를 작성하고 히스토그램을 그려서 측정치들의 분포상태를 조사할 때에 급구간의 선정방법을 합리적으로 하지 않으면 쓸모가 적은 도수분포표나 히스토그램이 되고 말 경우가 있다. 또한 앞으로 논의하게 될 각종의 통계량을 산출해 내는 데에도 부적당한 도수분포표는 큰 오차를 수반하게 될 가능성이 커진다.

급구간을 선정하는 문제는 어떤 통일된 방법이 있는 것은 아니나 일반적으로 다음의 순서에 기준하여 도수분포표를 만들면 큰 오차는 없을 것이다. 각 순서마다 앞의 예제의 경우를 들어서 설명하겠다.

1. 데이터의 수를 센다. 이 수를 n이라 한다.

$n = 64$

2. 데이터의 최대치 χmax와 최소치 χmin을 구한다.

$\chi max = 1690$, $\chi min = 609$

3. 측정치의 최소단위를 구한다.

측정치의 최소단위 : 1

4. 최대치와 최소치간에 존재할 수 있는 데이터 종류의 최대수를 다음 식에 의하여 구한다.

데이타종류의 최대수 :

$$\left(\frac{\chi max - \chi min}{\text{측정치의 최소단위}}\right) + 1 = \left(\frac{1690 - 609}{1}\right) + 1 = 1082$$

5. 급의 수를 정한다. 급의 수는 일반적으로 데이터의 수에 따라서 다음과 같이 잡아 주는 것이 적절하다.

데이터의 수	적절한 급의 수
40 ~ 100	5 ~ 9
100 ~ 200	8 ~ 12
200 이상	10 ~ 16

이 예제에서는 $n = 64$ 이므로 급의 수를 8로 가정한다.

6. 한 급에 포함될 수 있는 데이터의 종류의 수를 구한다.

$$\frac{\text{데이터의 종류수}}{\text{급의 수}} = \frac{1082}{8} = 135.3 \fallingdotseq 136$$

정수치로 떨어지지 않으면 올림하여 위처럼 큰 수치로 정한다.

7. 급의 구간폭을 구한다.

구간폭=(측정치의 최소단위)×(한 급에 포함될 수 있는 데이터의 종류수)=1×136=136

8. 도수분포용지를 준비한다. 이 용지는 <표 2.2>와 같은 것으로 급의 No, 급구간, 중앙치($\tilde{x}_i$), 도수체크, 도수(f_i), 도수분포율을 기입한 용지이다.

9. 급구간을 결정하여 이 용지에 기입한다. 우선 첫번째 급의 경계치를 다음에 의하여 정한다.

첫번째 급의 하측경계치

$$= \chi min - \frac{(\text{급의수}) \times (\text{한급에포함될 수 있는 데이터의 종류수})}{2}$$

$$-(\text{데이터 종류의 최대수})$$

$$= 609 - \frac{8 \times 136 - 1082}{2}$$

$$= 609 - 3 = 606$$

첫번째 급의 상측경계치

=첫번째 급의 하측경계치+구간폭

$= 606 + 136 = 742$

이 742는 두번째 급의 하측경계치가 되며 만약 어느 측정치의

값이 742라면 이는 두번째 급에 포함시키고 이를 명백히 하기 위하여 첫번째 급의 상측경계치를 〃742미만〃이라고 쓴다.

두번째 급의 상측경계치는

$$742 + 136 = 878$$

이 될 것이다. 이와같이 순차적으로 급구간의 경계치를 정한다.

10. 급의 중앙치는 다음과 같이 구하여 준비된 용지에 기입한다.

$$\text{중앙치}(\tilde{x}_i) = \frac{\text{급의 양 경계치의 합계}}{2}$$

11. 도수를 체크하여 준비된 용지에 기입한다. 도수의 체크는 <표 2.2>에서 처럼 /, //, ///, ////, 등으로 하거나 一, 丅, 下, 正, 正로 하거나 한다.

12. 도수체크를 세어서 도수란에 기입한다.

13. 도수분포율을 구하여 기입한다. 이는 각 급의 도수(f_i)를 데이터의 총수 n으로 나눈 것이다. 또 이 값을 백분율로 표시하는 경우도 많이 있다.

<표 2.2>에서는 백분율로 표시하였다. 따라서 그 총합은 1.0 혹은 100%가 된다.

$$\text{도수분포율} = \frac{\text{도수}(f_i)}{n}$$

2.3 累積度數

주어진 자료를 정리하고 분석하는데 있어서 앞에서 설명한 도수분포표나 히스토그램을 사용하여 데이터의 대략의 특성을 알 수 있으나 만약 어떤 측정치 이하에 속해있는 데이터가 전체 데이터에 비하여 어느정도 인가의 분포를 쉽게 알려면 누적도수를 구하면 될 것이다. 즉 누적도수는 어떤값(일반적으로 급의 상측경계치)이하의 도수가 어느정도 인가를 표시하는 것으로 k번째 급의 상측경계치까지의 누적도수는

$$\sum_{i=1}^{k} f_i$$

가 된다.

각급의 도수 f_i를 n으로 나눈 값을 앞에서 도수분포율이라고 불렀는데 흔히 쓰이는 다른 이름은 相対度數이다. 이 상대도수도 또한 백분율로 표시되는 경우도 많다. 만약 누적도수를 데이터의 총수 n으로 나누어서 표시하면 累積相対度數라고 부르고 백분율로 나타내기도 한다.

<표 2.2>의 자료에 대하여 상대도수, 누적도수, 누적상대도수를 구하여 보면 <표 2.3>과 같다. 이와같은 표를 만들어 놓으면 어떤 측정치 미만이나 이상의 데이터가 얼마나 있는가를 일목요연하게 알 수 있다. 예를 들면 전구수명이 1014시간 미만인 것은 전체 64개 중에서 20개이며 0.312(31.2%)를 차지하고 1014

<표 2.3>

누 적 도 수 분 포 표

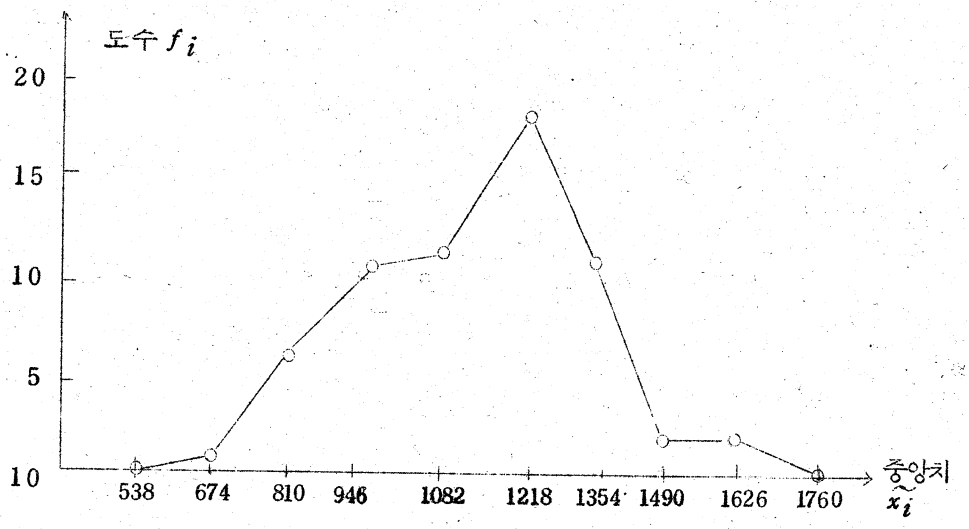
급의 No	급 구 간	중앙치 ($\tilde{x}_i$)	도수 (f_i)	상 대 표 수 f_i/n	누 적 도 수 $\sum_{i=1}^{k} f_i$	누적상태도수 $\sum_{i=1}^{k} f_i/n$
1	606 ~ 742 미 만	674	1	0.016	1	0.016
2	742 ~ 878 미 만	810	7	0.109	8	0.125
3	878 ~ 1014 미 만	946	12	0.187	20	0.312
4	1014 ~ 1150 미 만	1082	13	0.203	33	0.515
5	1150 ~ 1286 미 만	1218	18	0.281	51	0.796
6	1286 ~ 1422 미 만	1354	11	0.172	62	0.968
7	1422 ~ 1558 미 만	1490	1	0.016	63	0.984
8	1558 ~ 1694 미 만	1626	1	0.016	64	1.000
합 계			64	1.000	64	1.000

시간 이상인 것은 64－20＝44개로서 전체의 1－0.312＝0.688 (68.8%)인 것도 쉽게 계산할 수 있다.

<표 2.3>과 같이 누적도수를 표시한 분포를 누적도수분포라 하고 이것을 나타내는 표를 누적도수분포표라 부른다.

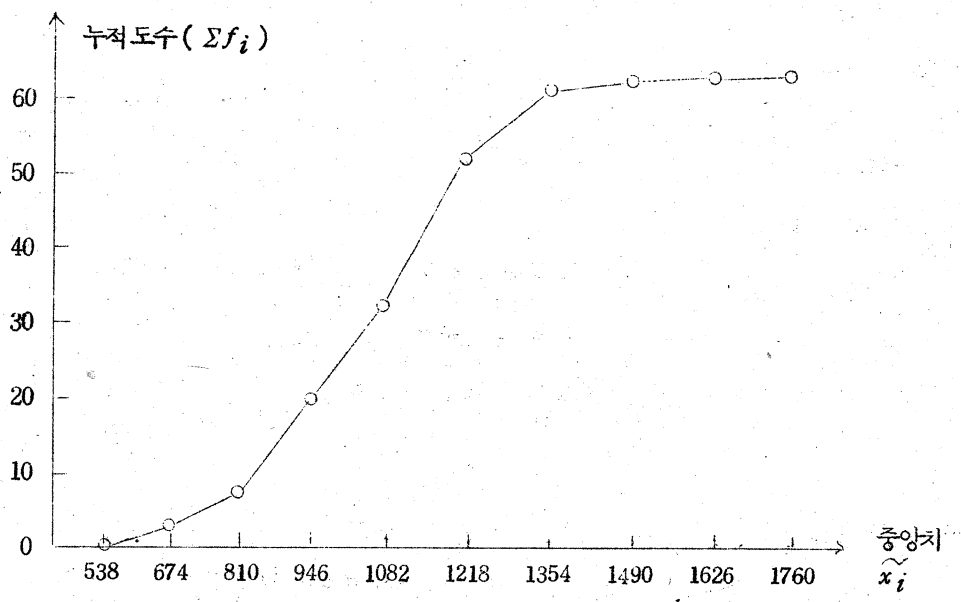
<그림 2.1>에서 각급의 도수 f_i를 기둥그림으로 표시한 히스토그램을 보았는데 도수의 크기를 보여주는 그림으로 도수다각형을 그릴 수도 있다. 이의 작성요령은 히스토그램에서 막대정상의 중앙치들을 구하여 연결하는 것으로 실제로 각급의 중앙치 $\tilde{x}_i$에서 도수 f_i의 크기를 좌표로 찍어서 연결하는 것과 같다.

<그림 2.2>은 전구수명데이터의 도수다각형을 보여주고 있는데 이 그림에서처럼 첫번째 급의 왼쪽에 도수 $f_i=0$을 가진 급을 만들고 마지막 급의 오른쪽에도 마찬가지로 하여 그림이 가로축에 밀착되도록 하여 준다.



<그림 2.2> 도 수 다 각 형

위의 <그림 2.2>는 각급의 중앙치에서 도수를 연결하여 다각형을 만들었는데 만약 각급의 중앙치에서 누적도수를 연결하여 다각형을 만들면 누적도수 다각형이 되고 이를 오자이브(ogive) 또는 누적도수곡선이라고도 부른다. <그림 2.3>은 전구수명자료의 누적도수다각형이다.



<그림 2.3> 누 적 도 수 다 각 형

한가지 첨부하여 말하여 둘 것은 <그림 2.2>와 <그림 2.3>의 도수다각형과 누적도수다각형에서 세로축에 도수대신에 상대도수를 그릴수도 있고 누적도수대신에 누적상대도수를 나타낼 수도 있다. 이는 자료를 정리하는데 있어서 어떤 특성을 나타내 주는 것이 중요한 가에 따라서 임의로 선택할 수 있다.

2.4 平均値와 分散度

정리되지 않은 원자료로부터 도수분포표 및 누적도수분포표를 만들어서 많은 정보를 알아낼 수 있는 것을 앞에서 설명하였다.

그러나 어떤 자료의 특성을 알아내기 위하여 간단한 몇개의 숫자로서 표본의 성질을 분석하고 이로부터 이 표본이 추출된 모집단의 특성을 나타낼 수는 없는가? 실제로 자료분석의 주요 목적은 자료가 얻어진 모집단에 대하여 그의 특성을 몇개의 상수로서 기술할 수 있도록 자료의 현상에 대한 수학적인 표현방법을 찾아내는 일이다.

뒤에 공부하게 될 확률분포는 데이터의 값이 어떤 확률분포에 따라서 발생되는가 하는 분포의 수식적인 표현에 대하여 논의하게 되며 관측한 데이터로부터 이러한 확률분포를 결정짓는 상수를 결정하게 된다. 주어진 표본자료로부터 도수분포를 조사하고 이 분포의 특성을 나타내는 상수를 구해내는 것은 표본자체의 특성을 기술하는 방법이 되는 것은 물론 확률분포의 수학적인 이론을 발전시키는 실마리가 된다.

아래에서 먼저 분포의 중심적 경향을 나타내는 상수들에 대하여 설명하여 보자. 뒤에 분포의 分散度를 나타내는 상수들에 대하여 설명하겠다.

2.4.1 中心的 傾向의 測度

다음의 측도들이 데이터의 중심적 경향을 표시하는 상수로서 흔히 쓰여진다.

가. 산술평균, $\bar{x}$

나. 기하평균, G

다. 중앙치, $\tilde{x}$

라. 최빈수, Mo

마. 범위의 중앙치, Md

가장 흔히 쓰이는 것은 산술평균이며 간단히 평균이라고도 하고 $\bar{x}$로 나타낸다. 만약 n개의 데이터가 있고 이를 $x_1, x_2, \cdots\cdots, x_n$으로 표시하면 산술평균은

$$\bar{x} = \frac{x_1 + x_2 + \cdots\cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (2.1)$$

이 된다. 틀릴 염려가 없을 경우에는 간단히 다음과 같이 쓴다.

$$x = \frac{\Sigma x}{n}$$

기하평균 G는 n개의 데이터를 다음과 같이 처리하여 얻어진다.

$$G = (x_1 \cdot x_2 \cdot \cdots\cdots \cdot x_n)^{\frac{1}{n}} \qquad (2.2)$$

이는 데이터들이 모두 양인 경우에 사용된다.

중앙치 $\tilde{x}$는 데이터를 크기 순서대로 나열하였을 때 데이터의 수 n이 홀수이면 중앙에 위치하는 데이터가 되고 만약 n이 짝수이면 중앙에 위치하는 두개의 데이터의 평균치가 된다.

이와같이 중앙치는 순위를 나열하는 것만으로 구할 수 있어 간단히 구할 수 있으나 중앙의 값만 사용하므로 산술평균에 비하여 전체의 데이터를 활용하는 효율성이 약간 떨어지는 것을 피할 수 없겠다. 그러나 동떨어진 데이터가 있는 경우에 그 영향을 받지 않는다는 이점이 있다. 데이터의 수가 너무 많으면 중앙의 값을 찾는 것이 용이하지 않고 효율도 떨어지므로 n의 크기가 10미만일 경우에만 쓰인다.

최빈수 Mo는 똑같은 값이 반복되는 빈도가 가장 큰 수를 최빈수라 한다. 마지막으로 범위의 중앙치 Md는 데이터 중에서 최대치 χmax과 최소치 χmin의 평균치이다. 즉

$$Md = \frac{\chi max + \chi min}{2} \qquad (2.3)$$

이것은 $\tilde{x}$보다는 계산이 간단하다는 이점이 있으나 양단의 최대치와 최소치만을 사용하므로 효율이 낮고 또 극단적으로 동떨어진 데이터에 민감한 영향을 받으므로 사용할때 조심하여야 할 것이다.

예를들어 앞의 중심적 경향의 측도들을 계산하여 보자. 다음과 같은 8개의 데이터가 있다.

데이터 : 3 , 9 , 6 , 15,5,3,5,5 (2.4)

이 데이터에 대하여 $\bar{x}$, G, $\tilde{x}$, Mo, Md를 각각 구하여 보자.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{3+9+6+15+5+3+5+5}{8} = 6.375$$

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots\cdots \cdot xn} = \sqrt[8]{3 \cdot 9 \cdot 6 \cdot 15 \cdot 5 \cdot 3 \cdot 5 \cdot 5} = 5.558$$

$\tilde{x}$을 구하기 위하여 위의 데이터를 크기 순으로 나열하면

데이터 : 3,3,5,5,5,6,9,15 (2.5)

이 되고 따라서 $\tilde{x}$는 4번째와 5번째 숫자의 평균치인

$$\tilde{x} = \frac{5+5}{2} = 5$$

이 되고 최빈수는 세번 반복이 가장 많으므로 Mo = 5이며 범위의 중앙치는

$$Md = \frac{\chi max + \chi min}{2} = \frac{15+3}{2} = 9$$

가 된다.

데이터의 수 n이 대단히 커서 용이하지 않을 때에는 급으로 나눈 도수분포로부터 중심적 경향의 측도를 계산할 경우가 있다.

또 경우에 따라서는 원자료 없이 도수분포만이 알려져 있어 불가피하게 급을 사용하여 계산하는 때도 허다하다. 예를들어 상당히 많은 통계 보고서가 급을 사용한 형태로서 이루어져 있다. 이런

경우에 산출평균의 계산은

$$\bar{x} = \frac{\sum_{i=1}^{q} f_i \tilde{x}_i}{n} = \frac{\Sigma f\tilde{x}}{n} \qquad (2.6)$$

이 된다. 여기에서 f_i 는 i 번째 급의 도수이고 $\tilde{\chi}_i$ 는 그의 중앙치이며 급의 총수는 q 라고 가정하였다. <표 2.2>의 도수분포에 대하여 $\bar{\chi}$ 를 계산해 보면

$$\chi = \frac{\Sigma f\tilde{\chi}}{n}$$

$$= \frac{(1)(674) + (7)(810) + (12)(946) + \cdots\cdots + (1)(1626)}{64}$$

$$= 1020.2$$

이 된다.

급을 사용하는 경우 중앙치 $\tilde{\chi}$ 의 계산은 다음의 공식을 따른다. 먼저 데이터를 순서에 따라서 나열했을때 끝번째의 데이터가 들어있는 급을 중앙치급이라 하자. 전구수명 데이터에 대해서는 <표 2.3>으로부터 누적도수를 관찰하여 $\frac{n}{2} = \frac{64}{2} = 32$ 번째의 데이터가 4번째의 급구간에 있으므로 중앙치급은 (1014 — 1150 미만)이다.

이 급을 일반적인 경우를 가상하여 k 번째 급이라 하자. 그러면 $\tilde{\chi}$ 에 대한 공식은

$$\tilde{\chi} = L + \frac{\frac{n}{2} - \sum_{i=1}^{k-1} fi}{fk} \cdot C \qquad (2.7)$$

이 되며 여기서 L은 중앙치급의 하측경계치이고 f_i 는 앞에서 정의한 바와 같이 i 번째 급의 도수이며 C는 각급의 간격이다.

<표 2.3>으로 부터 $\tilde{\chi}$를 계산하면

$$k = 4$$

$$\sum_{i=1}^{k-1} f_i = \sum_{i=1}^{3} f_i = 20$$

$$f_4 = 13$$

$$L = 1014$$

$$C = 136$$

이므로

$$\tilde{\chi} = 1014 + \frac{32 - 20}{13} (136)$$

$$= 1139.5$$

이 된다.

위의 $\bar{\chi}$와 $\tilde{\chi}$의 계산에 있어서 우리가 한가지 유의해야 할 것은 이들 계산은 데이터가 각 급구간에서 균일하게 분포되어 있다는 가정하에서 이루어졌다는 것이다. 따라서 각 급구간 내에서 데이터의 분포가 균일하다는 가정이 성립되지 않으면 위의 계산방법

은 무리가 있다고 말할 수 있을 것이다.

도수분포표로부터 $\bar{\chi}$와 $\tilde{\chi}$를 계산해 내는 방법을 설명하였는데 기타 중심적 경향의 측도인 G, Mo, Md 등은 개개의 데이터의 값을 모르므로 도수분포표로부터 직접 계산하기는 곤란하다.

2.4.2 散布度의 測度

앞에서 본 중심적 경향을 나타내는 측도들이 관측된 데이터의 중심이 어디에 위치하고 있는가를 나타내는데 비하여 데이터의 변동성 즉 분포가 그 중심에서 어느정도 퍼져 있는가를 나타내는 측도를 分散度 혹은 散布度라 한다.

평균이 같은 분포라 하더라도 분포도가 다르면 분포의 모양에 큰 차이가 있으며 어떤 분포에 대하여 평균과 분산도를 안다면 그 분포의 모양을 대강 짐작할 수 있다. 분산도를 표시하는데 다음과 같은 측도들이 흔히 쓰인다.

가. 分散, S^2

나. 標準偏差, S

다. 範囲, R

라. 平均偏差, D_m

마. 変動係数, V_C

표본의 크기가 n인 표본으로부터 관측치 $\chi_1, \chi_2, \cdots\cdots \chi_n$을 얻었을 때 이 데이터의 분산은 다음식에 의하여 구해진다.

$$S^2 = \frac{\sum_{i=1}^{n} (\chi_i - \bar{\chi})^2}{n-1} \qquad (2.8)$$

윗 식의 분자에서 $\chi_i - \bar{\chi}$를 偏差라 부르고 이는 개개의 데이터의 값에서 산술평균을 뺀 차이이다. 이를 제곱하여 n개의 데이터에 대해서 합을 구하면 분자가 되는데 이를 편차제곱의 합이라 부르고 간단히 제곱의 합이라 부르기도 한다. 이 제곱의 합을 좀 더 간단히 계산하기 위하여 다음의 관계식을 사용하면 편리하다.

$$\sum_{i=1}^{n} (\chi_i - \bar{\chi})^2 = \sum_{i=1}^{n} \chi_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} \chi_i\right)^2$$

$$= \Sigma\chi^2 - \frac{1}{n}(\Sigma\chi)^2 \qquad (2.9)$$

물론 분산 S^2이 크면 분산도가 커서 데이터가 퍼져있는 정도가 크다는 얘기이며 만약 $S^2 = 0$이면 이는 모든 데이터의 값이 같다는 말이 된다. 식 (2.5)에 있는 데이터의 분산은

$$\Sigma\chi^2 = 3^2 + 3^2 + 5^2 + \cdots\cdots + 15^2 = 435$$

$$(\Sigma\chi)^2 = (3 + 3 + 5 + \cdots\cdots + 15)^2 = 2501$$

이므로

$$S^2 = \frac{\Sigma\chi^2 - \frac{1}{n}(\Sigma\chi)^2}{n-1} = \frac{435 - (\frac{1}{8}) \cdot (2501)}{7} = 17.48$$

이 된다.

두번째로 표준편차는 분산의 제곱근이므로 표준편차 S는 다음과 같이 나타내 진다.

$$S = \sqrt{\frac{\sum_{i=1}^{n} (\chi_i \sim \bar{\chi})^2}{n-1}} = \sqrt{\frac{\Sigma\chi^2 - \frac{1}{n}(\Sigma\chi)^2}{n-1}} \qquad (2.10)$$

따라서 식 (2.5) 의 데이터에 대해서는 표준편차의 값이

$$S = \sqrt{17.48} = 4.18$$

이 된다.

세번째로 범위 R은 데이터의 최대치와 최소치와의 차이며

$$R = \chi max - \chi min \qquad (2.11)$$

이 측도는 데이터가 어느정도 큰 영역에 걸쳐서 산포되어 있는가를 나타내 준다.

네번째로 평균편차 Dm은 각 데이터의 평균으로부터의 편차의 절대치의 합을 평균한 것이다. 즉,

$$Dm = \frac{\sum_{i=1}^{n} | \bar{\chi}_i - \bar{\chi} |}{n} \qquad (2.12)$$

식 (2.5) 의 데이터의 평균편차는

$$Dm = \frac{|13-6.3751|+|13-6.3751|+|15-6.3751|+\cdots+|15-6.375|}{8}$$

$$= 2.8125$$

이 된다.

마지막으로 변량계수 Vc에 대하여 설명하겠다. 위에 설명된 네 개의 분산도들은 측정단위가 달라지면 같은 데이터라 할지라도 변량에 큰 차가 있을 것이며 측정단위가 다른 두 종류의 데이터를 위의 방법에 의하여 직접 비교하는 것은 곤란할 것이다. 이런 경우의 좋은 방법은 표준편차를 평균으로 나누어서 상대적인 분산도를 관찰하는 것이다.

변량계수 Vc는 다음에 의하여 정의된다.

$$Vc = \frac{S}{\bar{\chi}} \cdot 100\,(\%) \qquad (2.13)$$

변량계수는 일반적으로 퍼센트로 표시된다. 식 (2.5)의 데이터에 대하여 Vc를 계산하면

$$Vc = \frac{4.18}{6.375} \times 100 = 65.6\,\%$$

으로 상당히 큰 상대적 변동이 있음을 알 수 있다.

만약 데이터의 수 n이 크면 위의 분산도의 측정이 복잡한 계산이 되며 이런 경우에는 도수분포표를 만들어 급을 사용하여 계산하면 편리하다. 물론 정확히 같은 값이 나오지 않겠지만 일반적

으로 매우 근사한 값이 나오게 된다. 표<2.2>와 같은 도수분포표로부터 분산을 구하는 공식은 다음과 같다.

$$S^2 = \frac{\sum_{i=1}^{q} f_i(\hat{\chi}_i \sim \bar{\chi})^2}{n-1} \qquad (2.14)$$

여기에서 q는 급의 수를 나타내고 $\bar{\chi}$는 식(2.6)에 의하여 구하여지는 산술평균이다. <표 2.2>의 전구수명측정치에 대하여 분산을 구해보면 $\bar{\chi} = 1020.2$이고 $q = 8$이므로

$$S^2 = \frac{(1)(674-1020.2)^2 + (7)(810-1020.2)^2 + \cdots\cdots}{63}$$

$$\frac{(1)(1626-1020.2)^2}{} = 45107.1$$

이 되며 표준편차는

$$S = \sqrt{45107.1} = 212.4$$

이다. 변동계수는

$$V_C = \frac{S}{\bar{\chi}} = \frac{212.4}{1020.2} \times 100 = 20.8\,\%$$

이며 범위 R과 표준편차 Dm은 개별적인 데이터의 값을 모르므로 계산이 곤란하다.

아직까지 자료의 정리방법으로 원자료로부터 도수분포표, 누적도수분포표 등을 작성하고 히스토그램, 도수다각형, 누적도수다각형 등을 그려서 일차적으로 자료의 특성을 검토하고 이차적으로 데이터의 중심적 경향, 분산도 등을 측정하는 측도들을 자세히 다루어 보았다. 이와같은 자료의 정리와 분석을 통하여 주어진 표본의 자료가 지니는 정보를 획득해 내는 능력은 통계업무에 종사하는 사람이면 알아두는 것이 바람직한 기본적 지식이라 하겠다.

△ 연습문제△

〔2.1〕 다음의 자료는 운동선수 80명의 체중을 측정한 것이다. 이 자료에서 도수분포표를 만들어 보아라

65	54	71	82	67	78	65	79	68	78
73	93	70	83	73	67	85	83	93	66
75	68	68	85	84	73	78	95	74	72
90	84	80	76	74	65	97	75	79	103
88	84	72	75	63	80	77	71	77	62
90	82	93	56	76	75	79	78	94	95
81	68	85	72	62	81	62	89	82	74
101	105	98	83	83	76	71	87	75	87

(a) 만들어진 도수분포표를 사용하여 히스토그램, 도수다각형, 누적도수분포표, 누적도수다각형을 작성하여라.

(b) 만들어진 도수분포표를 사용하여 평균($\bar{\chi}$), 중앙치($\tilde{\chi}$), 분산(S^2), 표준편차(S), 변동계수(Vc)를 구하라.

〔 2.2 〕 위의 (2.1)번의 자료중에서 첫번째 열에 있는 8개의 데이터에 대하여 중심적 경향을 측정하는 측도들을 계산하여 보아라. 즉 평균($\bar{\chi}$), 기하평균(G), 중앙치($\tilde{\chi}$), 최빈수(Mo) 범위의 중앙치(Md)를 구해라.

〔 2.3 〕 위의 (2.1)번의 자료 중에서 첫번째 열에 있는 8개의 데이터에 대하여 분산도를 측정하는 측도인 분산(S^2), 표준편차(S), 범위(R), 평균편차(Dm), 변동계수(Vc)를 계산하라.

Ⅲ章 : 確率과 確率分布

제 2장에서는 주로 많은 데이터가 얻어졌을 경우 이 데이터를 통계적인 관점에서 어떻게 처리하여 필요한 정보를 획득하는가에 대한 것이었다. 제 3 장부터는 불확실성하에서의 의사결정을 함에 있어서 통계학의 지식과 방법이 어떻게 쓰여지는가를 보여주는 통계학의 새로운 면에 대하여 다루기로 한다.

그런데 불확실성이 개입된 문제에서는 언제든지 확률의 개념이 따르기 마련이며 확률론에 근거를 둔 해결방안을 모색하게 될 것이다. 먼저 기본적인 확률이론을 검토하여 보자.

3.1 確 率

실험의 결과가 우연에 의하여 결정지어지며 이러한 실험이 똑같은 조건하에서 여러번 반복되어질 수 있는 실험을 시행이라 한다. 시행때마다 나타날 수 있는 가장 기본적인 결과를 그 시행의 근원사상이라고 한다. 근원사상은 더이상의 기본단위로서 분해되지 않으며 근원사상이 한개 이상 모여서 이루어지는 집합을 사상이라고 한다. 물론 하나의 근원사상도 사상이 될 수 있다.

어떤 사상 A가 일어나리라고 기대되는 확실성을 수량적으로 표시한 값을 사상 A의 確率이라고 부른다. 그리고 어떠한 시행과 관련하여 생각할 수 있는 모든 가능한 근원사상을 한데 묶어논 집합을 標本空間이라고 한다.

예를 들어 동전을 두번 연달아 던지는 시행에 있어서 앞면이 나오면 H로 표시하고 뒷면이 나오면 T로 표시하기로 하자. 이 실험에서 4개의 근원사상이 필요하며 따라서 표본공간을 S로 표시하면 이것은

$$S = \{(H, T), (T, H), (H, H), (T, T)\} \quad (3.1)$$

로 표시된다. 예로서 (H, T)의 의미는 첫번째 동전을 던져서 앞면이 나오고 두번째는 뒷면이 나온 근원사상이란 뜻이다.

만약 첫번째 동전에서 앞면이 나오는 사상을 A로 표시하면 이 A라는 사상을 이루는 근원사상은 (H, T)(H, H)의 두개가 있으며

$$A = \{(H, T), (H, H)\}$$

로 표시된다. 만약 동전의 앞면과 뒷면이 나올 확률이 똑같이 $\frac{1}{2}$씩이라면 식 (3.1)에 있는 4개의 근원사상은 각기 일어날 확률이 동일하므로 각 근원사상의 확률은 $\frac{1}{4}$라고 하겠다. 따라서 사상 A는 $\frac{1}{4}$확률을 가진 두개의 근원사상으로 이루어졌으므로 사상 A의 일어날 확률은

$$P(A) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

이다.

확률 P(A)와 같은 표현은 A가 집합이면 P(A)는 A에 대한 확률함수이므로 이를 確率集合函數라고 부르고 다음과 같은

공리에 의하여 정의한다.

(가) 표본공간 S의 확률은 1이다. 즉

$$P(S) = 1$$

(나) 표본공간 S안에서 정의되는 모든 사상 A에 대하여

$$0 \leq P(A) \leq 1$$

(다) 표본공간 S안에서 임의의 사상 A_i와 $A_j (i \neq j)$가 주어졌을 때 만약 A_i와 A_j가 서로 공통적인 근원사상을 가지고 있지 않으면 다음이 성립한다.

$$P(A_i \text{ 또는 } A_j) = P(A_i) + P(A_j)$$

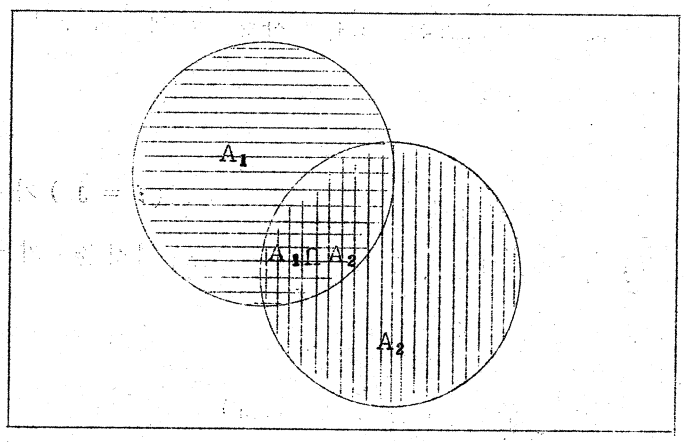
위에서 $P(A_i \text{ 또는 } A_j)$는 사상 A_i 또는 A_j중 최소한 어느 하나가 일어날 확률을 나타내며 보통 $P(A_i \cup A_j)$라고 쓰기도 한다. 이에 비하여 A_i과 A_j가 동시에 일어날 확률을 $P(A_i \cap A_j)$라고 쓴다. 위의 공리들은 앞으로 다룰 확률론의 바탕이 될 것이다.

3.2 確率法則

두개의 사상 A와 B가 어떤 표본공간에서 정의되었을때 다음의 가법정리가 성립한다.

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \qquad (3.2)$$

이 관계를 설명하기 위하여 울러의 그림으로 나타내면 <그림 3.1>이 된다.



<그림 3.1> $P(A_1 \cup A_2)$

즉 A_1 또는 A_2에 속하는 상대적 빈도는 A_1에 속하는 상대적 빈도와 A_2에 속하는 상대적 빈도의 합에서 A_1과 A_2에 동시에 속하는 상대적 빈도를 뺀것과 같음이 명백하다.

두 사상 A_1, A_2가 동시에 일어날 수 없는 경우를 상호배반이라 하고 이들을 배반사상이라 한다. A_1과 A_2가 서로 배반인 경우는 $P(A_1 \cap A_2) = 0$이며 따라서 식 (3.2)로부터

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

이다.

두 사상 A_1, A_2가 동시에 일어날 때 그중 한 사상이 일어나는 확률이 다른 사상이 일어나는 확률에 아무런 영향을 미치지 못할

때 두 사상은 서로 독립이라고 하고 이 경우에는 다음의 식이 성립한다.

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2) \qquad (3.3)$$

만약 어떤 임의의 두 사상 A_1, A_2가 식 (3.3)을 만족시키면 A_1, A_2는 서로 독립사상이 된다. 두 사상이 독립이 아니면 이들 사상을 종속사상이라고 한다.

사상 A_1이 일어나는 것을 조건으로 하여 사상 A_2가 일어나는 확률을 A_1을 조건으로 하는 A_2의 조건부확률이라 하고 $P(A_2 / A_1)$이라고 한다. 이 확률을 계산하는 방법은

$$P(A_2 \; A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)} \qquad (3.4)$$

에 의하여 구하여진다.

두 사상 A_1, A_2가 동시에 일어나는 확률 $P(A_1 \cap A_2)$는 다음에 의하여도 구해질 수 있다.

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1) \cdot P(A_2 \; A_1) \\ &= P(A_2) \cdot P(A_1 \; A_2) \end{aligned} \qquad (3.5)$$

이와같은 관계식을 흔히 승법정리이라 부른다.

<예제 3.1> 어떤 회사의 100의 종업원에 대하여 연령과 성별로 분류하여 다음 표를 얻었다. <표 3.1>에 있는 것처럼 연령에 따라서 A_1, A_2, A_3로 분류하고 성별에 따라 B_1, B_2로 나

누자.

<표 3.1> 공무원의 연령, 성별분류

연령	성별		합계
	남 (B_1)	여 (B_2)	
34세이하 (A_1)	21	9	30
35~54 (A_2)	42	18	60
55세이상 (A_3)	7	3	10
합계	70	30	100

만약 임의로 한사람의 종업원을 선택하였을 경우 이사람의 나이가 34세이하이거나 또는 남자일 확률은 $P(A_1 \cup B_2)$인데 식 (3.2)를 이용하면

$$P(A_1 \cup B_1) = P(A_1) + P(B_1) - P(A_1 \cap B_1)$$

이며

$$P(A_1) = 30/100 = 0.3$$

$$P(B_1) = 70/100 = 0.7$$

$$P(A_1 \cap B_1) = 21/100 = 0.21$$

이므로

$$P(A_1 \cup A_2) = 0.3 + 0.7 - 0.21 = 0.79$$

이다.

<표 3.1>에서 A_1, A_2, A_3는 서로 동시에 일어날 수 없으므로 (예로서 한사람이 동시에 34세 이하도 되고 35-54세에 속할 수는 없다.) A_1, A_2, A_3는 서로 상호배반관계에 있다. B_1, B_2도 상호배반관계에 있다.

다음으로 조건부 확률을 보자. 만약 임의로 뽑힌 사람이 여자라는 것이 판명되었을 때 이 여자가 35～54세에 속할 확률을 구해보자. 이 확률은 $P(A_2\ B_2)$라고 표시할 수 있고 식(3.4)로 부터

$$P(A_2\ B_2) = \frac{P(A_2 \cap B_2)}{P(B_2)}$$

$$= \frac{18/100}{30/100} = 0.6$$

이 됨을 알 수 있다. 다음으로 식 (3.5)의 승법정리를 알아보자. 임의로 뽑힌 사람의 연령이 35～54세에 속하고 동시에 여자일 확률 $P(A_2 \cap B_2)$는 식 (3.5)를 이용하면

$$P(A_2 \cap B_2) = P(B_2) \cdot P(A_2\ B_2)$$

$$= \left(\frac{30}{100}\right)(0.6) = 0.18$$

이 된다.

3.3 確率変数

어떤 시행의 표본공간을 S라고 하고 이 표본공간의 임의의 근원사상을 W라 하자. 근원사상 W에 대하여 어떤 일정한 규칙에 따라 하나의 실수를 대응시키고 이 관계를 X(W)로 표시하면 이는 수학적인 의미에서의 함수가 되며 이러한 함수를 확률변수라 한다. 확률변수는 X, Y, Z등의 대문자로 표시하는 것이 통례이다.

<예제 3.2> 동전을 두번 던지는 시행을 생각하여 보자. 여기서 근원사상은 식(3.1)에서 본 바와 같이 전부 4개가 있으며 표본공간은

$$S=\{(H,T),(T,H),(H,H),(T,T)\}$$

이다. 각 근원사상에 대하여 앞면(H)의 수를 세어 이 수를 대응시키면 $X((H,T))=1$, $X((T,H))=1$, $X((H,H))=2$, $X((T,T))=0$가 되며 이런 함수관계를 표시하는 변수 X를 시행과 관련하여 생각할 수 있는 확률변수의 한가지이다.

확률변수는 이산형과 연속형으로 대별된다. 동전을 여러번 던졌을 때 앞면의 수, 생산공정에서 한시간에 나오는 불량품의 개수 등은 이산확률변수이며, 자택에서 근무처까지의 통근소요시간, 제품의 중량처럼 연속적인 값을 취할 수 있는 확률변수는 연속확률변수라 한다. 연속확률변수의 표본공간은 실수전체가 되든가 어떤 구간이 되든가 한다.

확률변수가 취할 수 있는 각각의 값 x에 대하여 일정한 확률이 대응되도록 함수 $P(x)$를 정의하면 이를 확률변수 X의 確率密度함수라 하고 약자로 $p.d.f.$라고 표시한다. 일반적으로 이산확률변수 X에 대하여서는 $P(x)$로 표시하나 연속확률변수에 대하여서는 $f(x)$로 나타내 준다.

앞에서 논의한 동전을 두번 던지는 시행에 있어서 앞면을 세어서 확률변수 X의 값으로 정해줄 경우에는

$$P(X=0)=1/4$$

$$P(X=1)=1/2$$

$$P(X=2)=1/4$$

이 됨을 알 수 있다. 이를 다음과 같이 표시할 수 있다.

$$\begin{aligned} p(x) &= 1/4 \text{ , 만약 } x=0 \\ &= 1/2 \text{ , 만약 } x=1 \\ &= 1/4 \text{ , 만약 } x=2 \\ &= 0 \quad \text{ , 만약 } x \neq 0,1,2 \end{aligned} \tag{3.6}$$

이를 좀더 관찰해 보면 다음과 같은 두가지 성질을 알 수 있다.

(가) $p(x) \geq 0$, 모든 x의 값에 대하여

(나) $\sum_x p(x)=1$, 모든 x의 값에 대하여 $p(x)$를 합칠 경우

위의 두가지 특성은 모든 이산확률변수의 $p.d.f.$가 갖는 고유의 성질이다.

연속확률변수의 표본공간은 어떤 구간이기 때문에 $p.d.f.$인

$f(x)$로 부터 확률을 구하기 위해서는 확률변수 X가 어떤 구간 (a,b)에 속할 확률이 무엇인가와 같이 구간을 정해주어야만 한다. 구간 (a.b)에서 X가 값을 취할 확률은

$$p\ (\ a \leq X \leq b\ ) = \int_a^b f(x)dx \qquad (3.7)$$

라고 표시된다.

연속확률변수 X의 확률밀도함수 $f(x)$는 다음의 성질을 가진다.

(가) $f(x) \geq 0$, 모든 x의 값에 대하여

(나) $\int_{-\infty}^{\infty} (x)dx = 1$

다음으로 확률변수 X와 관련하여 자주 쓰이는 것으로 확률분포함수가 있다. 이 함수를 $F(x)$로 나타내고 저자에 따라서는 간단히 분포함수 또는 누적분포함수라고도 부른다. $F(x)$는 다음 식으로 정의된다.

$$F(x) = P\{\ X \leq x\ \} \qquad (3.8)$$

확률분포함수는 다음과 같은 성질을 갖는다.

(가) 만일 $x_1 < x_2$이면 $F(x_1) \leq F(x_2)$이다. 즉 $F(x)$는 비감소함수이다.

(나) $F(-\infty) = 0$, $F(\infty) = 1$

(다) $F(x)$는 우측으로부터 연속이다.

동전을 두번 던져서 앞면을 세는 문제를 생각해 보자. 확률변수 X가 앞면의 수라면 식 (3.6)에 있는 확률을 사용하여

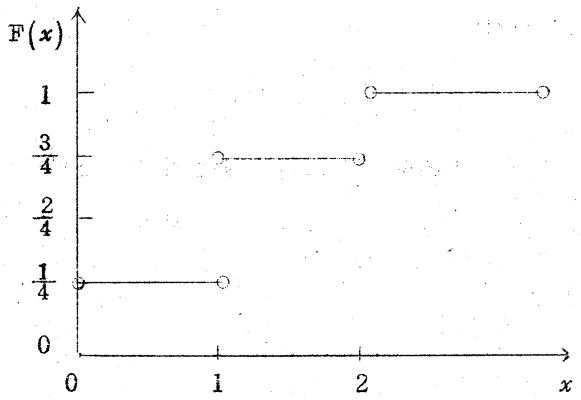
$$F(x) = P\ (\ X \leq x\ )$$

$= 0$, 만약 $x < 0$

$= 1/4$, 만약 $0 \leq x < 1$

$= 3/4$, 만약 $1 \leq x < 2$

$= 1$, 만약 $x \geq 2$



<그림 3.2> 확률분포함수

이 되며 이들 그림으로 그리면 <그림 3.2>가 된다.

이 그림에서 $F(x)$는 마치 계단과 같은 모양을 한 함수임을 알 수 있다. 이러한 모양은 이산확률변수의 분포함수에서 나타나며 연속확률변수의 경우에는 <그림 3.2>에 있는 누적도수다각형과 흡사한 연속적인 그림이 나올 것이다.

3.4 期待値와 分散

확률변수의 분포와 관련된 문제중에서 가장 중요한 개념의

하나는 수학적 기대치 또는 간단히 기대치에 관한 것이다. 확률변수 X가 이산형인 경우에는 X가 취할 수 있는 모든 값과 그 값이 취하는모든 확률치와의 곱의 합을 기대치로 정의한다. 즉 X의 기대치를 E(X)라고 써서

$$E(X)=\sum_{x} x\, p(x) \qquad (3.9)$$

로 나타내며 이것은 X의 이론적 분포의 평균치가 된다.

X가 연속형일 경우에는 다음에 의하여 정의한다.

$$E(X)=\int_{-\infty}^{\infty} x\, f(x)\,dx \qquad (3.10)$$

다음으로 X의 分散을 정의하여 보자. X의 분산을 σ^2으로 표시하고 X의 기대치 E(X)를 μ로 나타내면

$$\sigma^2 = \sum_{x} (x-\mu)^2 p(x) \qquad (3.11)$$

이 되는데 이것은 이산확률변수 X의 분산이 되고

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \qquad (3.12)$$

은 연속확률변수의 분산이다. 예제를 들어 설명하여 보자.

<예제 3.3> 동전을 던져 앞면과 뒷면이 나타나는 사항을 관찰하는 문제를 생각하여 보자. 이때 동전은 세번 던진다. 근원사상은 모두 8개이며 표본공간을 써 보면

$$S = \{(H,H,H),(H,H,T),(H,T,H),(T,H,H),(H,T,T),(T,H,T),$$
$$(T,T,H),(T,T,T)\} \qquad (3.13)$$

이며 확률변수 X를 앞면의 수로 규칙을 정한다면 다음과 같은 $p.d.f.$ $p(x)$를 얻는다.

<표 3.2> 확률밀도함수($p.d.f$)

x의 값	P(x)
0	1／8
1	3／8
2	3／8
3	1／8

이 경우는 이산확률변수 X가 되며 X의 기대치는 식 (3.10)으로 부터

$$E(X)=\sum_x x\,p(x)$$

$$=(0)\left(\frac{1}{8}\right)+(1)\left(\frac{3}{8}\right)+(2)\left(\frac{3}{8}\right)+(3)\left(\frac{1}{8}\right)$$

$$=1.5$$

이 되며 X의 분산은 식 (3.11)에서

$$\sigma^2=\sum_x (x-1.5)^2 p(x)$$

$$=(0-1.5)^2\left(\frac{1}{8}\right)+(1-1.5)^2\left(\frac{3}{8}\right)+(2-1.5)^2\left(\frac{3}{8}\right)$$

$$+(3-1.5)^2\left(\frac{3}{8}\right)$$

$$=0.75$$

<예제 3.4> 어떤 확률변수 X가 구간 (a,b)에서 균일한 분포를 갖는다고 하자. 이 경우의 $p.d.f.$ 는

$$f(x)=\frac{1}{b-a} \quad , \text{만약} \quad a \leq x \leq b \qquad (3.14)$$

$$= 0 \ , \text{만약} \quad x > b \quad \text{또는} \quad x < a$$

이 되며 확률변수 X가 구간 (a,b)에서 평등분포 또는 구형분포를 갖는다고 말한다. X의 기대치는 식 (3.10)에서

$$E(X)=\int_{-\infty}^{\infty} xf(x)dx = \int_{b}^{a} x\left(\frac{1}{b-a}\right) dx$$

$$= \left(\frac{1}{b-a}\right)\left[\frac{x^2}{2}\Big|_{a}^{b}\right] = \frac{1}{b-a}\left(\frac{1}{2}(b^2-a^2)\right)$$

$$= \frac{a+b}{2}$$

이 되며 X의 분산은 식 (3.12)로 부터

$$\sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx$$

$$= \int_{b}^{a}\left(x-\frac{a+b}{2}\right)^2\left(\frac{1}{b-a}\right) dx$$

$$= \frac{(a-b)^2}{12}$$

이 된다.

분산의 공식 (3.11)과 (3.12)를 동시에 표현하는 분산의 공식은

$$\sigma^2 = E\left[(X-\mu)^2\right] \qquad (3.15)$$

이라 쓸 수 있고 우리는 흔히 다음과 같은 관계식을 이용하여

계산을 간단히 한다.

$$\begin{aligned}\sigma^2 &= E[(X-\mu)^2] \\ &= E[X^2-2\mu X+\mu^2] \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E[X^2] - \mu^2 \end{aligned} \tag{3.16}$$

위에서 $E[X^2]$은 이산형 X에 대해서는

$$\sum_x x^2 p(x)$$

이며 연속형 X에 대해서는

$$\int_{-\infty}^{\infty} x^2 f(x)dx$$

이다.

3.5 確率変數의 分布

확률밀도함수($p.d.f.$)는 이론적으로 무한개가 있다고 볼 수 있으며 그중에서 잘 알려지고 응용이 많이 되는 몇가지에 대하여 상의하여 보자. 먼저 이산확률변수 X에 대한 $p.d.f.$는 다음의 것들이 많이 쓰여진다.

(가) 베르눌리 분포

(나) 이항 분포

(다) 포아송 분포

(라) 초기하 분포

연속확률변수의 분포로는 다음의 것들을 들 수 있겠다.

(가) 정규 분포

(나) 지수형 분포

(다) t -분포

(라) x^2- 분포

(마) F -분포

3.5.1 베르눌리 분포

어떤 시행에 있어서 결과가 둘중의 하나일 때 이와 관련된 이산확률분포는 베르눌리 확률밀도함수를 가지며 X가 취할 수 있는 값을 0와 1으로 분류한다. 제품을 검사해서 합격, 불합격을 판정하는 경우 $x=0$를 합격으로, $x=1$을 불합격으로 놓거나, 어떤 일의 성공과 실패를 $x=0$를 성공으로, $x=1$을 실패로 놓을 수 있겠다. 이처럼 두개의 정상적인 결과를 $x=0, 1$으로 놓을 경우이다. 관련된 확률변수 X는 베르눌리 분포를 하고 $p.d.f.$는

$$p(x)=p^x\ q^{1-x}, x=0,1 \qquad (3.17)$$

으로 표시된다. 여기서 p는 $x=1$이 될 확률이고 $q=1-p$이다.

이 확률분포의 기대치와 분산을 구해보자. 기대치는

$$E(X)=\sum_x x\,p(x)$$

$$=0\cdot P(X=0)+1\cdot P(X=1)$$

$$=0+P \quad =P$$

이며 분산은

$$\sigma^2 = \sum_x (x-p)^2 p(x)$$

$$= (0-p)^2 \cdot q + (1-p)^2 \cdot p$$

$$= p^2 q + q^2 p$$

$$= pq$$

이다. 만약 공업제품의 생산과정에서 검사를 하는데 합격될 확률이 0.95이고 불합격될 확률이 0.05이면 $x=0$를 합격으로 놓으면 $q=0.95$, $p=0.05$이며 베르눌리 분포는

$$p(x) = (0.05)^x (0.95)^{1-x}$$

가 되며 X의 기대치는 $p=0.05$이고(즉 제품이 $x=1$(불합격)이 될 기대치가 100개중 5개라는 의미) X의 분산은 $\sigma^2 = pq = (0.05)(0.95) = 0.0475$이 된다.

3.5.2 二項 分布

위에서 설명한 베르눌리 시행을 똑같은 조건하에서 n회 반복시행하게 되면 X가 취할 수 있는 값은 $x=0,1,\cdots\cdots,n$이 되며 이러한 확률변수 X는 이항분포를 갖는다고 한다. 이 분포의 $p \cdot d \cdot f$는

$$p(x) = \binom{n}{x} p^x q^{n-x}, \quad x=0,1,\cdots\cdots,n \qquad (3.18)$$

으로 나타내진다. 여기서 $\binom{n}{x}$는 n개 중에서 x개를 뽑아내는

조합의 수를 말한다. $\binom{n}{x}$는 $\frac{n!}{x!(n-x)!}$ 로 부터 구해진다.

동전을 한번 던졌을때 나타나는 결과는 앞면과 뒷면이므로 이 시행은 베르눌리 시행이며 동전을 예로들어 세번 던지는 경우는 이항분포를 따르는 시행이 되며 $p = q = 0.5$ 이고 $n = 3$ 이 된다.

앞면이 나오는 가능한 수는 $x = 0,1,2,3$ 이며 이의 확률을 식 (3.18)에 의해 구해보자.

$$P(X=0) = \binom{3}{0} (\tfrac{1}{2})^0 (\tfrac{1}{2})^3 = \frac{3!}{0!(3-0)!} (\tfrac{1}{2})^3 = \frac{1}{8}$$

$$P(X=1) = \binom{3}{1} (\tfrac{1}{2})^1 (\tfrac{1}{2})^{3-1} = \frac{3!}{1!(3-1)!} (\tfrac{1}{2})^3 = \frac{3}{8}$$

$$P(X=2) = \binom{3}{2} (\tfrac{1}{2})^2 (\tfrac{1}{2})^{3-2} = \frac{3!}{2!(3-2)!} (\tfrac{1}{2})^3 = \frac{3}{8}$$

$$P(X=3) = \binom{3}{3} (\tfrac{1}{2})^3 (\tfrac{1}{2})^{3-3} = \frac{3!}{3!(3-3)!} (\tfrac{1}{2})^3 = \frac{1}{8}$$

위의 확률의 값들은 실제로 <표 3.2>에서 구해본 확률밀도함수와 같은 것임을 알 수 있다.

이항분포의 확률변수 X의 거래치와 분산을 구해보자. 이항분포는 베르눌리시행을 n번하여 얻은 것이며 매 시행마다 독립이라는 가정하에 있으므로 이항확률변수의 기대치는 베르눌리 확률변수 기

대치의 n 배이다. 즉

$$E(X) = np$$

이며 분산도 마찬가지로 베르눌리 확률변수의 분산 pq 에 n 배하여

$$\sigma^2 = npq$$

를 얻을 수 있다.

<예제 3.5> 10개의 제품을 검사할 때에 불합격품이 될 확률이 각 제품마다 $\frac{1}{10}$ 이면 10개를 전부 검사했을 경우 불량품의 기대되는 숫자를 구하라. 또 불량품의 숫자가 2개 이하가 되는 확률을 구하여라.

<풀이> 불량품의 갯수를 나타내는 변수는 이항확률변수가 되며 이의 분포는

$$p(x) = \binom{10}{x} \left(\frac{1}{10}\right)^x \left(\frac{9}{10}\right)^{10-x}, \quad x = 0, 1, \cdots\cdots, 10$$

이다. 불량품의 기대치는

$$E(X) = np$$

$$= (10)\left(\frac{1}{10}\right) = 1$$

이며, 두개 이하가 되는 확률을 구하기 위하여 $X = 0, 1, 2$ 가 되는 확률을 구하여 합하면 될 것이다.

$$P(X = 0) = \binom{10}{0}\left(\frac{1}{10}\right)^0\left(\frac{9}{10}\right)^{10} = 0.349$$

$$P(X=1)=\binom{10}{1}\left(\frac{1}{10}\right)^{1}\left(\frac{9}{10}\right)^{9}=0.387$$

$$P(X=2)=\binom{10}{2}\left(\frac{1}{10}\right)^{2}\left(\frac{9}{10}\right)^{8}=0.194$$

따라서 두개 이하가 되는 확률은

$$P(X\leq 2)=P(X=0)+P(X=1)+P(X=2)$$
$$=0.349+0.387+0.194$$
$$=0.930$$

이며 양품이 8개 이상 나올 확률이 또한 0.930이다.

3.5.3 포아송分布

포아송 분포는 확률밀도함수를

$$p(x)=\frac{\mu^{x}e^{-\mu}}{x!},\quad x=0,1,2,\cdots\cdots \qquad (3.19)$$

으로 가지며 μ는 X의 기대치이다. 그리고 e의 값은 약 2.718이며 자연대수에서 기준으로 사용하는 상수이다.

따라서 포아송 분포는 μ의 값만 주어지면 X가 취할 수 있는 모든 x의 값에 대한 확률을 구할 수 있다.

전화교환대에 매분마다 걸려오는 전화의 수나 식료품상에 어느 시간당 들어오는 사람의 수 등은 일반적으로 포아송 분포를 한다.

포아송 확률변수의 기대치는

$$E(X)=\mu$$

이고, 분산은

$$\sigma^2=\mu$$

이 되는 특이한 성질을 갖는다. 기대치와 분산이 같은 값을 갖는 분포는 매우 드무나 이 포아송 분포가 이러한 성질을 갖는다.

<예제 3.6> 어느 고속도로 상에서 일주일에 일어나는 사고의 횟수가 기대치 3을 갖는 포아송 분포를 한다고 알려져 있다. 어느 일주일 구간을 임의로 선택하였을 때 사고의 횟수가 4번 이상이 되는 확률을 구하라.

<풀이> $\mu=3$인 포아송 확률밀도함수는

$$p(x)=\frac{3^x e^{-3}}{x!},\quad x=0,1,2,\cdots\cdots$$

이므로 x의 값을 0,1,2,……등으로 대입하여 계산하여 표를 만들면 <표 3.3>이 되므로 사고의 횟수 X가 4이상일 확률은 다음과 같이 구해진다.

$$\begin{aligned}P(X\geq 4) &= 1-P(X\leq 3)\\ &= 1-[P(X=0)+P(X=1)+P(X=2)+P(X=3)]\\ &= 1-[0.0498+0.1494+0.2240+0.2240]\\ &= 1-0.6472\\ &= 0.3528\end{aligned}$$

<표 3.3> $\mu = 3$ 인 포아송분포의 확률

$X = x$	$p(x)$
0	0.0498
1	0.1494
2	0.2240
3	0.2240
4	0.1680
5	0.1008
6	0.0504
7	0.0216
⋮	⋮

3.5.4 超幾何 分布

초기하 분포를 설명하기 위하여 다음의 문제를 풀어보자. 어느 위원회의 위원수는 전부 10명인데 이중 여자의 비율은 40%이다. 이들 10명의 위원중에서 3명을 임의로 선택할 때 두명의 여자위원이 뽑히고 한명은 남자가 뽑힐 확률을 구하라.

10명의 위원중에서 6명은 남자이고 4명은 여자이므로 3명을 뽑을때 2명이 여자이고 1명이 남자가 될 경우의 수는

$$\binom{4}{2}\cdot\binom{6}{1}=\frac{4!}{2!(4-2)!}\cdot\frac{6!}{1!(6-1)!}$$

$$=6.6=36$$

이며 10명중에서 3명을 뽑는 모든 경우의 수는

$$\binom{10}{3}=\frac{10!}{3!(10-3)!}=120$$

이므로 구하고져 하는 확률은

$$\frac{\binom{4}{2}\cdot\binom{6}{1}}{\binom{10}{3}}=\frac{36}{120}=0.3$$

이 된다.

일반적인 경우에 대하여 초기하 분포를 확률밀도함수 (p,d,f)로 나타내면

$$P(x)=\frac{\binom{NP}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}},\qquad \begin{matrix}x\leq NP\\ x\leq n\end{matrix}\qquad (3.20)$$

이 되며 위의 문제의 경우에는 N을 전체 위원의 수, n은 임의로 뽑는 위원의 수, P는 여자의 비율(여기서는 0.4)이다.

초기하 확률변수 X는 n명 중에서 뽑히는 여자위원의 수를 나타내는 변수이다. X의 값은 NP나 n보다 더 큰 수일 수는 없

으므로 $x \leq NP$, $x \leq n$이라는 조건이 붙는다. 여기서는 $NP=(10)(0.4)=4$이고 $n=3$이므로 X가 취할 수 있는 값은 0,1,2,3, 이다.

구하고져 하는 확률은 $X=2$가 되는 확률이므로

$$P(X=2)=\frac{\binom{(10)(0.4)}{2}\binom{10(1-0.4)}{3-2}}{\binom{10}{3}}$$

$$=0.3$$

이 되서 앞에서 풀은 것과 일치한다.

3.5.5 正規分布

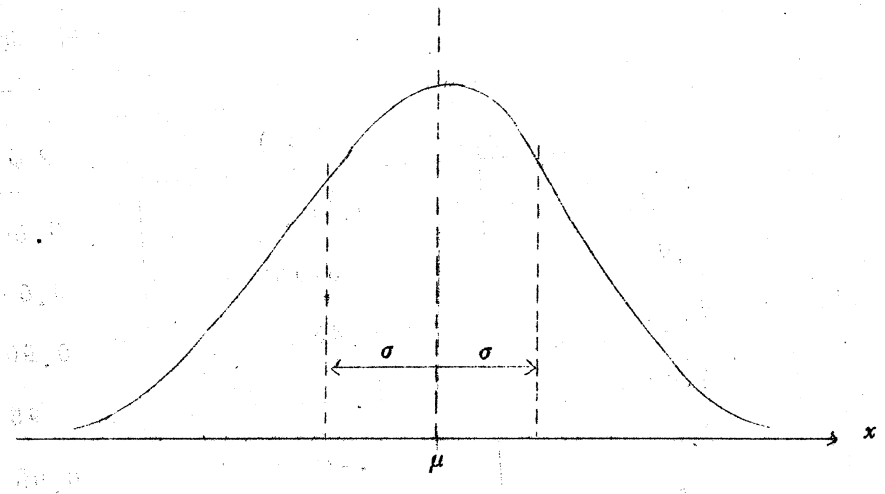
앞에서 설명된 4개의 분포는 모두 이산형 분포였다. 다음으로 연속형 분포를 다루어 보자. 먼저 가장 광범위한 응용력을 가지고 있고 통계적 추정과 검정에서 분포이론의 기초가 되는 정규분포를 알아보자

정규분포의 확률밀도함수는 μ를 확률변수 X의 기대치라고 하고 σ^2을 분산이라 하면

$$f(x)=\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad (3.21)$$

으로 표현되며 e는 자연대수의 기준이다. 정규확률변수 X가 취할

수 있는 값은 영역은 $+\infty$에서 $-\infty$까지이며 $f(x)$를 그려보면 <그림 3.3>과 같다.



<그림 3.3> 정규확률밀도함수의 곡선

정규분포의 모양은 두개의 모수인 μ와 σ^2에 의해서 정해진다. μ를 중심으로 좌우 대칭이며 μ의 값은 곡선중심의 위치를 정하고, σ의 값은 곡선의 모양을 정하는데 이 분포를 $N(\mu, \sigma^2)$이라고 나타내기도 한다. $f(x)$의 값은 $X=\mu$에서 가장 크며 X와 μ의 차가 크면 급속하게 대칭으로 감소하여 0에 가까워진다. σ가 커지면 정규분포곡선이 납작해지며 σ가 작으면 가운데가 뾰족해지며 X의 값이 대부분 μ의 근처의 값을 취하게 된다.

만약 $\mu=0$, $\sigma^2=1$이면 이러한 분포를 標準正規分布라고 하고 $N(0,1)$로 표시한다. 이 경우의 p, d, f는 간략히

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}Z^2} \qquad (3.22)$$

라고 표시되며 표준정규확률변수 Z가 어떤 구간에 있게 될 확률을 구하기 용이하도록 다음과 같은 표는 거의 모든 통계책에서 발견할 수 있다.

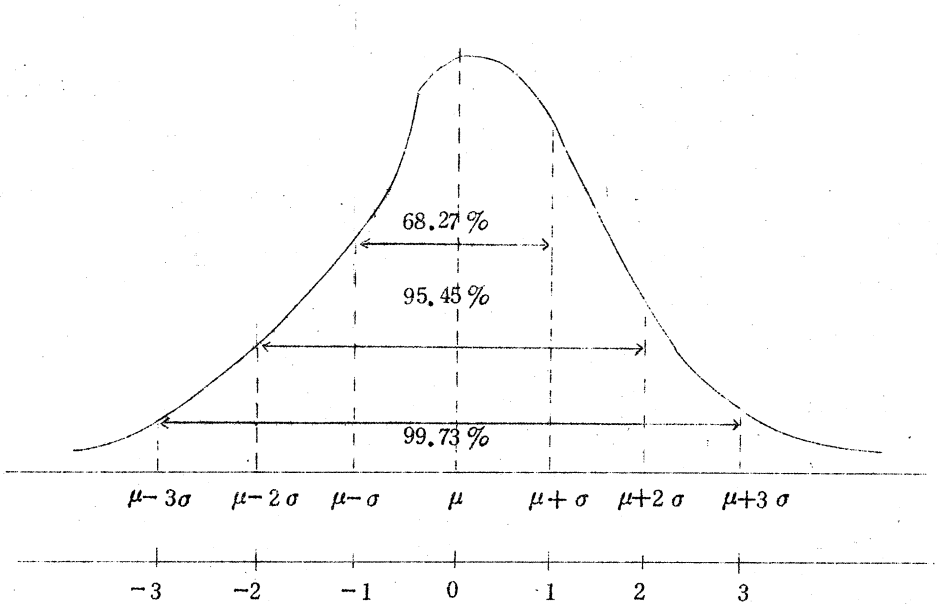
<표 3.4> $\phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{1}{2}z^2} dz$ 의 값

z	$\phi(z)$	$2\phi(z)$
0.675	0.25	0.50
1.0	0.3413	0.6827
1.645	0.45	0.90
1.96	0.475	0.95
2.0	0.4772	0.9545
2.58	0.495	0.99
3.0	0.49865	0.9973

Z의 여러가지 값에 대한 표준정규분포의 값을 그림으로 나타낸 것이 <그림 3.4>이다.

정규분포 $N(\mu, \sigma^2)$을 갖는 확률변수 X에서 μ를 빼고 σ로 나누면 Z가 되며

$$Z = \frac{X - \mu}{\sigma} \qquad (3.23)$$



<그림 3.4> 표준정규분포

이는 앞에서 얘기한 N (0 , 1)으로 표준정규분포이다. 따라서 Z는 변수 X의 평균에서의 편차를 표준편차의 단위로 표시한 것이므로 정규곡선에서는 평균치 μ를 중심으로 하여 표준편차의 ±1의 범위안에는 68.27 %, ±2배의 범위안에는 95.45 %, 그리고 ±3배의 범위 안에는 99.73 %의 상대도수가 포함된다.

정규분포에서 변수 X가 어느 구간 (a , b)사이에 들어갈 확률은 다음과 같은 요령에 의하여 계산한다.

$$P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma})$$

$$= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \qquad (3.24)$$

변수 Z는 N(0, 1)이므로 <표 3.4>와 같은 표를 사용하여 확률을 구할 수 있는데 그 계산방법은 다음과 같다.

먼저

$$\frac{a-\mu}{\sigma} = Z_1$$

$$\frac{b-\mu}{\sigma} = Z_2$$

라 놓고 다음의 세가지 경우를 분류하여 생각하여 보자.

(가) $z_1 > 0$, $z_2 > 0$인 경우

이 때에는 <그림 3.5>가 되며 구하고져 하는 확률은 <표 3.4>에 있는 함수의 정의를 사용하여

$$P(a \leq X \leq b) = P(z_1 \leq Z \leq z_2)$$

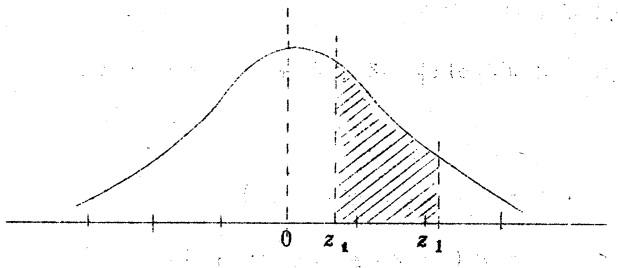
$$= \phi(z_2) - \phi(z_1)$$

이 된다. 왜냐하면 <그림 3.5>에서 빗금친 부분의 면적은 가로축에서 $(0, z_2)$사이의 면적으로부터 $(0, z_1)$사이의 면적을 빼면 되기 때문이다.

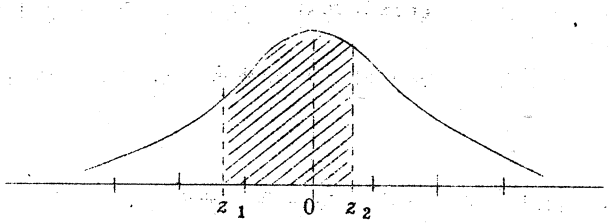
(나) $z_1 < 0$, $z_2 > 0$인 경우

이 경우에는 <그림 3.6>에서 보면 명백히

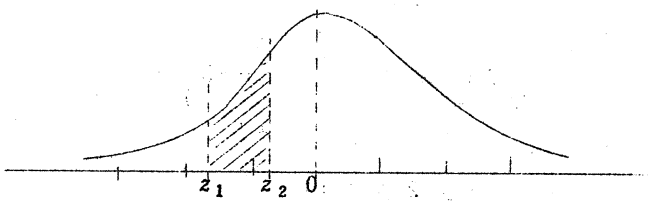
$$P(a \leq X \leq b) = P(z_1 \leq Z \leq z_2)$$



<그림 3.5> $z_1 > 0$, $z_2 > 0$



<그림 3.6> $z_1 < 0$, $z_2 > 0$



<그림 3.7> $z_1 < 0$, $z_2 < 0$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{1}{2}z^2} dz$$

$$= \phi(z_1) + \phi(z_2)$$

가 된다.

(다) $z_1 < 0$, $z_2 < 0$인 경우

앞의 (가)의 경우와 반대쪽이나 정규분포가 대칭이므로 이 대칭성을 사용하여

$$P(a \leq X \leq b) = P(z_1 \leq Z \leq z_2)$$
$$= P(-z_2 \leq Z \leq -z_1)$$
$$= \phi(-z_1) - \phi(-z_2)$$

가 됨을 알 수 있다.

<예제 3.7> 어느 대학교의 학생들의 체중이 근사적으로 정규분포를 한다고 하자. 만약 평균체중이 65 *kg*이고 표준편차가 10 *kg*이라고 알려져 있을 경우 한명의 대학생을 임의로 선택하면 이 학생의 체중이

(가) 75 *kg*에서 85 *kg*사이에 있을 확률을 구하라.

(나) 55 *kg*에서 75 *kg*사이에 있을 확률을 구하라.

<풀이> 이 학생의 체중 X는 $N(65, 10^2)$이므로 식(3.24)의 절차를 따라서

$$P(75 \leq X \leq 85) = P\left(\frac{75-65}{10} \leq Z \leq \frac{85-65}{10}\right)$$
$$= P(1 \leq Z \leq 2)$$
$$= \phi(2) - \phi(1)$$

이며 <표 3.4>를 이용하면 $\phi(2)=0.4772$, $\phi(1)=0.3413$ 이므로 확률은 $0.4772-0.3413=0.1359$이다. 두번째 질문에 대한 답은

$$P(55\leq X\leq 75)=P\left(\frac{55-65}{10}\leq Z\leq\frac{75-65}{10}\right)$$

$$=P(-1\leq Z\leq 1)$$

$$=\phi(1)+\phi(1)$$

$$=0.3412+0.3413$$

$$=0.6826$$

이 됨을 알 수 있다.

3.5.6 指型數 分布

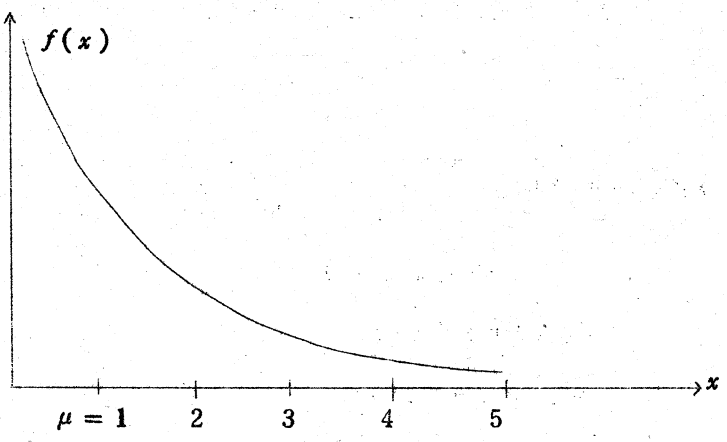
지수형 분포는 변수의 값이 0에서 ∞까지 어느 값이나 취할 수 있는 연속확률변수의 분포이며 이의 p, d, f는

$$f(x)=\frac{1}{\mu}e^{-\frac{x}{\mu}}, \quad x\geq 0 \qquad (3.25)$$

으로 표시되며 μ는 X의 기대치이며 e는 자연대수의 기준이다.

$\mu=1$일 때 지수형분포의 $p\cdot d\cdot f$는 대략 다음의 모양을 갖는다.

<그림 3.8>에서 보는 바와 같이 $p\cdot d\cdot f$는 감소함수이다. 일반적으로 전구의 수명이나 써비스기관에서 한명의 고객을 써비스하



<그림 3.8> 지수형 확률분포, $\mu = 1$

는데 걸리는 시간 등은 지수형 분포를 하는 경우가 많다. 지수형 확률변수 X의 기대치와 분산은

$$E(X) = \mu$$

$$\sigma^2 = \mu^2$$

이다.

3.5.7 χ^2, t, F分布

표준정규분포를 하는 n개의 독립된 확률변수가 각각 $Z_1, Z_2, \cdots\cdots Zn$이라 하고 그 제곱의 합을 χ^2 즉

$$\chi^2 = \sum_{i=1}^{n} Z_i^2 \qquad (3.26)$$

로 표시하면 χ^2도 또한 하나의 확률변수로서 그 확률밀도함수는 다음 식으로 주어진다.

$$f(\chi^2) = \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} (\chi^2)^{\frac{n}{2}-1} e^{-\frac{1}{2}\chi^2}, \quad \chi^2 \geq 0 \qquad (3.27)$$

윗 식에서 Γ는 감마함수를 표시하는 기호로서 $n > 2$이면

$$\Gamma\left(\frac{n}{2}\right) = \left(\frac{n}{2}-1\right)! = \left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\cdots 3\cdot 2\cdot 1, \quad n\text{이}$$

짝수일 때

$$= \left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\cdots \frac{3}{2}\cdot\frac{1}{2}\cdot\sqrt{\pi}, \quad n\text{이}$$

홀수일 때

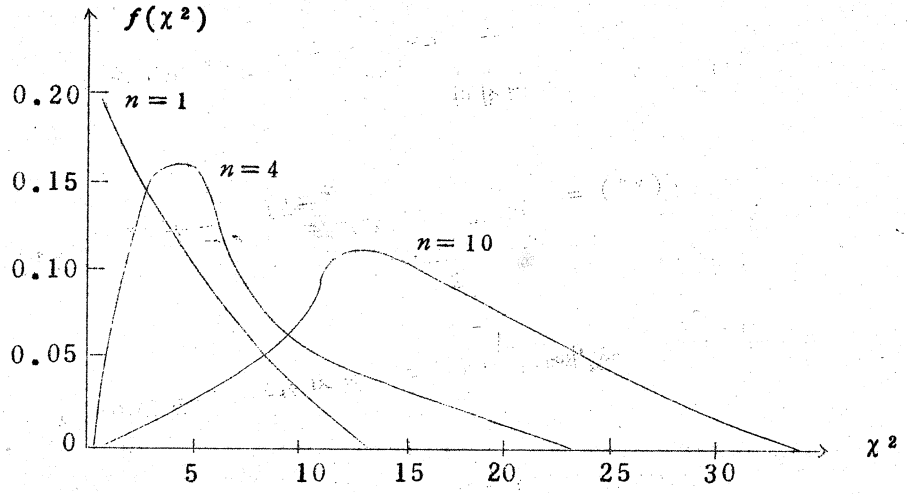
이며 또한 $\Gamma(1) = 1$, $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$이다. 확률밀도함수 식(3.27)에 의하여 정의되는 분포를 χ^2-분포라 한다.

χ^2-분포는 식 (3.27)에서 보는 바와 같이 n에 의하여 완전히 결정된다. n은 이 분포의 자유도라 하고 n의 값에 따라 χ^2-분포는 <그림 3.9>에 도시된 것과 같은 분포를 한다.

χ^2확률변수는 제곱의 합이므로 $\chi^2 \geq 0$이며 이의 기대치와 분산은 다음과 같다.

$$E(\chi^2) = n$$

$$\sigma^2_{\chi^2} = 2n$$



<그림 3.9 > $n=1,4,10$에 대한 χ^2분포의 $p\cdot d\cdot f$

실제적인 면에서 진공관의 수명이나 전구의 수명등이 경험적인 관찰에 의하면 χ^2 분포를 하는 경우가 있다.

다음으로 가설검정에서 많이 쓰이게 될 t 분포를 알아보자. Z를 표준정규분포, N (0 , 1)를 하는 확률변수라 하고 χ^2 를 Z와 독립인 자유도 n인 χ^2 확률변수라 하면 다음에서 정의하는 확률변수 T는 t 분포를 한다고 한다.

$$T=\frac{Z\sqrt{n}}{\sqrt{\chi^2}} \qquad (3.28)$$

확률변수 T의 확률밀도함수는 다음 식으로 주어진다.

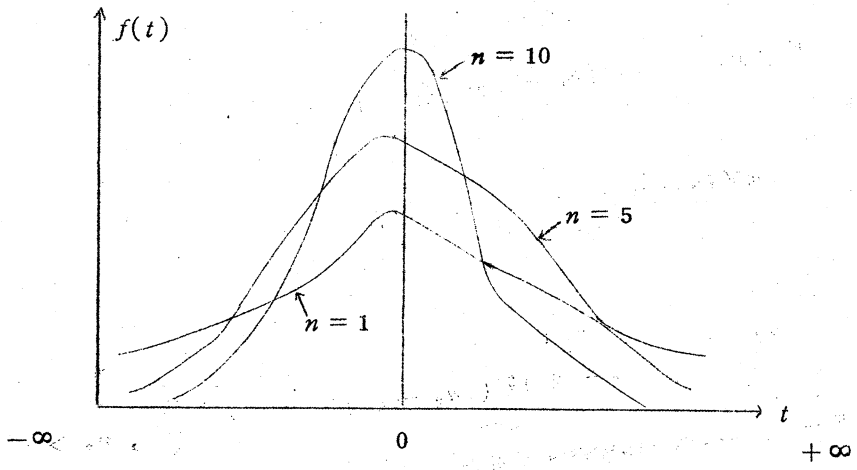
$$f(t)=\frac{1}{\sqrt{n\pi}}\,\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad (3.29)$$

이며, t의 값의 영역은 $-\infty < t < \infty$이며 식 (3.29)에서 보는 바와 같이 t분포의 $p \cdot d \cdot f \cdot$는 χ^2 확률변수의 자유도 n에 의하여 완전히 정의된다. T확률변수의 평균치와 분산은 각각 다음과 같다.

$$E(T) = 0$$

$$\sigma^2_T = \frac{n}{n-2}, \quad n > 2$$

<그림 3.10>에서 보는 바와 같이 t—분포는 그 평균치 0을 중심으로 대칭이며 자유도 n이 무한대로 커짐에 따라 표준정규분포에 수렴한다.



<그림 3.10> $n = 1, 5, 10$에 대한 t—분포의 $p.d.f$

t—분포의 용도에 대해서는 뒤에 다시 언급하기로 하겠다. 앞으로 설명될 표본분포이론, 통계적 추정 및 검정에서 자주 나오는

분포의 하나가 이 t 一분포이다.

다음으로 F 一분포를 알아보자.

χ_1^2를 자유도 n을 가진 카이제곱(χ^2) 분포를 하는 확률변수라 하고 χ_2^2을 자유도 $n2$를 가진 χ^2분포를 하는 확률변수라고 하자. 만약 이들 두 확률변수가 서로 독립된 확률변수라 하면 다음에서 정의되는 확률변수 F는 F 一분포를 한다.

$$F = \frac{\chi_1^2 / n1}{\chi_2^2 / n2} \qquad (3.30)$$

확률변수 F의 확률밀도함수는 그 표현식이 복잡하므로 여기서는 생략하기로 한다. F 一분포의 확률밀도함수는 두 카이제곱 확률변수의 자유도 $n1$과 $n2$에 의하여 완전히 결정된다. 따라서 이와 같은 관계를 표시하기 위하여 식 (3.30)에 있는 확률변수를 $F(n1, n2)$로 표시하기도 한다.

F 一분포의 평균치와 분산은 다음과 같다

$$E(F) = \frac{n_2}{n_2 - 2}, \quad n_2 > 0$$

$$\sigma_F^2 = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}, \quad n_2 > 4$$

F 一분포의 용도에 대해서는 뒤에 다시 언급하기로 한다. 주로 가설검정과 분산분석 등에 널리 쓰인다.

△ 연습문제 △

(3.1) 동전 하나와 주사위 하나가 동시에 던져졌다고 하자. 동전의 앞면을 H라고 하고 뒷면을 T로 하고, 주사위의 눈을 1, 2, 3, 4, 5, 6으로 표시하면 이 시행에서 몇개의 가능한 근원사상이 있느냐? 표본공간을 정의하라. 또한 다음 질문에 답하라

(a) 사상 A, B, C를 다음과 같이 정의하자.

A = {H와 맞음 짝수가 나오는 경우}

B = {주사위의 홀수가 나오는 경우}

C = {T와 홀수가 나오는 경우}

위의 A, B, C사상에서는 각각 어떠한 근원사상이 있는가를 밝혀라.

(b) 사상 A, B, C를 위처럼 정의할 때 확률 $P(A \cup B)$, $P(A \cap B)$, $P(A \cup C)$, $P(B \cup C)$, $P(A \cap C)$, $P(B \cap C)$를 구하라.

(c) 사상 A, B, C에서 어떤 사상끼리 서로 배반인가? 독립인가?

(d) 조건부 확률 $P(A/B)$, $P(B/A)$를 구하고 그 뜻을 설명하라.

(e) 조건부 확률 $P(B/C)$, $P(C/B)$를 구하고 그 뜻을 설명하라.

(3.2) 어떤 복권의 상금액을 확률변수 X라고 하자. 추첨결과

만일 1등이 되면 9백만원이 될 수도 있고 떨어지면 0원이 될 수도 있다. 이 복권의 제도는 100,000번에서 199,999번까지의 10만가지를 1조로 하여 5조를 매출, 추첨에 의하여 다음과 같은 당첨금을 주기로 하였다고 하자.

(등 급)	(당첨금)	(개수)	(확 률)
1등	9,000,000원	1	1/50만
애석상(1등과 조가 다름)	10,000원	4	4/50만
2등	1,000,000원	5	5/50만
3등	10,000원	50	50/50만
장 려 상	1,000원	500	500/50만

(a) 이 복권을 한장 샀을 경우 상금액 X의 기대치를 계산하라.

(b) 상금액 X의 표준편차를 구하라.

(c) 50만장을 다 팔아서 당첨금을 분배한 후에 순이익금을 2천만원으로 하고 싶다면 복권 한장에 대략 얼마씩 받아야 하겠는가?

(3.3) 어떤 사격수가 사격을 하여 목표물을 맞힐 확률은 매 사격마다 똑같이 3/4이라고 한다. 만약 이 사격수가 7번의 사격을 할 경우 다음의 질문에 답하라.

(a) 최소한 5번이상 명중시킬 확률을 구하라.

(b) 최소한 2번이상 목표물을 명중시키지 못할 확률을 구하라.

(c) 최소한 6번 이상 명중시키면 A라는 사람이 B에게 1,000원을 주고 4번이상 명중시키지 못하면 B가 A에게 1,000원을 주기로 하였다면 누구에게 더 이로운 내기가 되겠는가?

(3.4) 한여름 낮의 온도 T가 평균치 28℃이고 표준편차 4℃를 갖는 정규분포를 한다고 하자. 어느 한여름 낮을, 임의로 선택하였을때 다음의 질문에 답하라.

(a) 온도 T가 24℃에서 28℃사이에 있을 확률을 구하라.

(b) 온도 T가 36℃ 이상이 될 확률을 구하라.

(c) 온도 T가 32℃에서 36℃사이일 확률을 구하라.

Ⅳ章 : 標 本 分 布

4.1 標本理論의 基礎概念

표본은 모집단의 성질을 잘 대표할 수 있어야만 모집단에 대한 어떤 추정을 할 때에 보다 정확성을 기할 수 있을 것이다. 그러나 표본이 모집단의 성질을 잘 지니고 있다 할지라도 표본의 자료가 대표하는 어떤 특정치의 정확성에 관한 지식이 확실하지 않으면 그 정보의 이용에 어려움이 있을 것이다. 표본의 정보가 표본이론에 근거하여 어느정도의 정확성을 가지고 얻어질 수 있다면 표본의 이용도는 상당히 증가할 것이다.

확률표본은 정확성을 산출할 수 있는 근거를 마련해 주는 표본이며 보통 표본이라고 하면 이러한 확률표본을 의미한다. 확률표본 또는 임의표본이란 모집단을 구성하는 모든 단위에 일정한 확률이 부여되고 그 확률에 따라 뽑히는 표본을 말한다.

표본을 추출할 때에 모든 구성단위에 동일한 확률을 부여할 필요는 없지만 동일한 확률을 주는 경우가 가장 간단하며 이러한 경우에 얻어진 표본을 단순확률표본이라고 한다. 모집단의 구성원소의 수가 유한개일 때에는 이 모집단을 有限母集団이라 하고 무한개가 모집단을 형성하고 있다고 간주할 때에는 無限母集団이라 한다.

표본을 뽑을때 동일한 단위가 2회이상 뽑히는 것을 허용하는

경우에는 먼저 뽑힌 것을 다시 제자리에 넣는 경우로 이러한 방법을 復元抽出이라고 하며 이 경우에는 각 구성원소가 갖는 추출될 수 있는 확률은 항상 일정하다.

이에 비하여 거듭 뽑히는 것을 허용하지 않는 경우에는 일단 뽑힌 것은 다시 모집단 속에 넣지 않는 것이며 이러한 추출방법을 非復元抽出이라고 한다. 비복원추출인 경우에는 모집단의 각 구성원소가 표본에 뽑힐 확률이 항상 동일하게 유지되지는 않는다.

모집단 전체를 관찰하는 것을 全數調査라고 하며, 모집단의 규모가 적을 경우에 흔히 사용되고 인구센서스와 같은 것은 큰 규모의 모집단에서도 사용된다.

전수조사에 비하여 모집단의 일부분인 표본만을 관찰하는 조사를 標本調査라고 부른다.

표본조사는 전수조사에 비하여 다음과 같은 장점을 가지고 있다.

(가) 비용의 절감

(나) 신속한 처리

(다) 정확성

(라) 정보의 양

(마) 조사작업이 모집단 구성요소의 파괴를 초래할 경우 표본조사만이 가능한 조사임.

위와같은 이점을 표본조사가 가지고 있으나 모집단의 일부에 관한 조사이므로 전수조사와는 차이가 생기는데 이러한 차이를 標本誤差라고 하며, 이 오차를 줄이는 방법은 많이 연구되고 있으나

발생하지 않도록 하기는 어렵다.

표본조사이든 전수조사이든 관계없이 각각의 조사단위의 부정확한 관찰때문에 발생하는 차이를 非標本誤差라고 한다.

전수조사에서는 모집단의 규모가 커짐에 따라서 일반적으로 비표본오차가 증가한다. 전수조사에서는 언제나 이 비표본오차를 줄이는 것이 큰 과제가 되겠다.

그러나 표본조사에서는 조사단위가 비교적 정확히 조사되므로 이 비표본오차는 적고 반면에 표본오차가 존재하게 될 것이다.

표본을 뽑기 용이하도록 모집단의 구성단위를 기록한 리스트, 모집단의 소재를 알 수 있는 지도 등을 표본추출틀이라고 하고, 이 틀을 구성하고 있는 구성단위를 추출단위라고 한다. 표본을 모집단으로 부터 뽑는 방법에는 여러가지가 있으나 널리 쓰이는 방법으로는 난수표를 이용하여 完全無作為로 표본을 뽑는 任意抽出法, 모집단을 여러 층으로 나누어서 각 층마다 무작위로 뽑는 層別抽出法, 모집단리스트로부터 등간격을 사용하여 표본을 뽑는 系統抽出法 등이 있다. 이외에도 多段抽出法, 集落抽出法 등이 있으나 자세한 설명은 생략하기로 한다.

표본에 대하여 평균, 표준편차 등은 표본집단의 성질을 나타내는 특성치이며 이러한 특성치를 統計量이라고 한다. 이에 비하여 모집단의 성질을 나타내는 평균, 표준편차 등은 母數라고 한다.

통계량은 표본추출의 시행에서 얻어지는 일종의 확률변수이며 어떤 확률분포를 갖는다. 그 반면에 모수는 모집단의 성질을 설명해

주는 고유한 상수이다.

모집단 구성단위의 크기는 일반적으로 N으로 표시하고 표본구성단위의 수는 n으로 나타낸다. 통계량과 모수를 비교하고 그들이 각각 어떻게 계산되는가를 검토하여 보자. 표본에 추출될 n개의 구성단위를 $X_1, X_2, \cdots\cdots, X_n$이라 하고 모집단의 구성단위를 또한 $X_1, X_2, \cdots\cdots, X_N$이라 쓰자. 그러면 <표 4.1>과 같은 표를 만들 수 있다. 이 표에서 표본비율 $\bar{P}$를 구할 때 쓰이는 X는 어떤 주어진 특성을 가진 구성단위의 수(예로서 불량품의 개수 나이가 70세 이상의 사람수)를 나타내며 $\bar{P}$는 n개 중에서 X개라는 비율을 말한다. 반면에 모집단비율 P는 모집단 전체의 수 N에서 어떤 주어진 특성을 가진 구성단위의 수 X개의 비율을 말한다.

<표 4.1> 통계량과 모수의 비교

	통 계 량	모 수
평 균	$\bar{X} = \sum_{i=1}^{n} X_i / n$	$\mu = \sum_{i=1}^{N} X_i / N$
분 산	$S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / n-1$	$\sigma^2 = \sum_{i=1}^{N} (X_i - \mu)^2 / N$
표준편차	$S = \sqrt{S^2}$	$\sigma = \sqrt{\sigma^2}$
비 율	$\bar{P} = \frac{X}{n}$	$P = \frac{X}{N}$

모집단의 구성단위 중에서 확률의 개념에 근거하여 뽑은 변량이 표본단위이므로 표본단위 들로부터 얻어지는 통계량도 또한 확률변수이다. 그러므로 통계량은 도수분포로 정리하여 고찰할 수 있으며 통계량의 분포를 표본분포라고 한다. 먼저 표본평균 $\bar{X}$에 대하여 관찰하여 보자.

4.2 標本平均의 分布

표본평균의 분포이론을 설명하기 위하여 먼저 매우 간단한 다음과 같은 문제를 생각하여 보자. 어떤 회사에 5명의 종업원이 있고 그들의 연령은 <표 4.2>와 같다.

<표 4.2> 5명의 종업원으로 이루어진 모집단

종 업 원	연 령
A	27
B	39
C	30
D	36
E	42

이 5명이 하나의 유한 모집단을 형성하고 있다고 가정하고 이것으로부터 3명의 종업원을 임의 추출하여 이 표본에 뽑힌 종업원의 평균연령을 $\bar{X}$로 놓고 이 표본평균의 분포를 알아보자.

5명중에서 3명을 뽑는 조합의 수는 모두 <표 4.3>에서 보듯이 10가지이며 각 조합이 임의표본을 하나 선택할 때에 여기에 뽑힐 확률은 0.1이다.

<표 4.3> 가능한 표본조합과 표본평균

조 합	표본의 구성단위	관 찰 치	$\bar{X}$	$P(\bar{X})$
1	A,B,C	27,39,30	32	0.1
2	A,B,D	27,39,36	34	0.1
3	A,B,E	27,39,42	36	0.1
4	A,C,D	27,30,36	31	0.1
5	A,C,E	27,30,42	33	0.1
6	A,D,E	27,36,42	35	0.1
7	B,C,D	39,30,36	35	0.1
8	B,C,E	39,30,42	37	0.1
9	B,D,E	39,36,42	39	0.1
10	C,D,E	30,36,42	36	0.1

<표 4.3>의 오른쪽 두칸으로부터 표본평균 $\bar{X}$가 취할 수 있는 값과 그의 확률을 알 수 있으며 이를 정리하면 <표 4.4>를 얻을 수 있는데 이와같은 표를 표본평균 $\bar{X}$의 표본분포라고 한다.

<표 4.2>의 모집단으로부터 평균과 분산을 각각 구하면 다음

<표 4.4>　　$\bar{X}$의 표본분포

표본평균 $\bar{X}$	$P(\bar{X})$
31	0.1
32	0.1
33	0.1
34	0.1
35	0.2
36	0.2
37	0.1
39	0.1

과 같다.

$$\mu = \frac{\Sigma X}{N} = \frac{1}{5}(27+39+30+36+42) = 34.8$$

$$\sigma^2 = \frac{\Sigma (X-\mu)^2}{N} = \frac{1}{5}[(27-34.8)^2 + \cdots\cdots + (42-34.8)^2]$$

$$= \frac{1}{5}(154.8)$$

$$= 30.96$$

다음 <표 4.4>에 있는 $\bar{X}$의 표본분포로 부터 평균과 분산을 구하여 보자,

$\bar{X}$의 평균 $\mu_{\bar{X}}$는 식 (3.9)에 있는 확률변수 X의 기대치를 구하는 것과 같은 요령으로 구하면

$$E(\bar{X}) = \mu_{\bar{x}} = \Sigma \bar{x} p(\bar{x}) \qquad (4.1)$$

$$= (31)(0.1) + (32)(0.1) + \cdots\cdots + (39)(0.1) = 34.8$$

이 되는 것을 알 수 있고 $\bar{X}$의 분산 $\sigma^2_{\bar{x}}$는 식 (3.11)과 같은 방법으로 구하여

$$\sigma^2_{\bar{x}} = \Sigma (x - \mu_{\bar{x}})^2 \; p(\bar{x}) \qquad (4.2)$$

$$= (31-34.8)^2(0.1) + (32-34.8)^2(0.1) + \cdots\cdots +$$

$$= + (39-34.8)^2(0.1)$$

$$= 5.29$$

을 얻는다.

표본평균 $\bar{X}$의 분포이론을 고찰하여 보자. 정규분포 $N(\mu, \sigma^2)$의 모집단에서 임의추출한 n개의 표본을 $X_1, X_2, \cdots\cdots, X_n$이라 하고 그 표본평균은 <표 4.1>에 있는 바와 같이

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} X_1 + \frac{1}{n} X_2 + \cdots\cdots + \frac{1}{n} X_n$$

이라 하자. 그러면 $\bar{X}$의 기대치는

$$E(\bar{X}) = \mu_x = \frac{1}{n} E(X_1) + \frac{1}{n} E(X_2) + \cdots\cdots + \frac{1}{n} E(X_n)$$

$$= \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots\cdots + \frac{1}{n}\mu = \mu \qquad (4.3)$$

가 되며, $\overline{X}$의 분산은 $X_1, X_2, \cdots\cdots, Xn$이 서로 독립이므로

$$\sigma^2_{\bar{x}} = (\frac{1}{n})^2 \sigma^2_{x_1} + (\frac{1}{n})^2 \sigma^2_{x_2} + \cdots\cdots + (\frac{1}{n})^2 \sigma^2_{x_n}$$

$$= (\frac{1}{n})^2 \sigma^2 + (\frac{1}{n})^2 \sigma^2 + \cdots\cdots + (\frac{1}{n})^2 \sigma^2$$

$$= (\frac{1}{n})^2 (n\sigma^2) = \frac{\sigma^2}{n} \qquad (4.4)$$

이다.

위에서 $\overline{X}$의 기대치와 분산을 구할때 밑에 설명된 <정리 4.1>을 일부 사용하였다. 증명없이 정리를 소개한다.

<정리 4.1> X, Y, Z를 확률변수라 하고 $\sigma^2_x, \sigma^2_y, \sigma^2_z$을 그들의 분산이라 하며 C, k를 상수라고 할 때 다음이 성립한다.

(가) 만약 $Y = kX$이면 $E(Y) = kE(X)$이고 $\sigma^2_y = k^2 \sigma^2_x$이다.

(나) 만약 $Y = kX + C$이면 $E(Y) = kE(X) + C$이고 $\sigma^2_y = k^2 \sigma^2_x$이다.

(다) 만약 $Z = X + Y$이고 X와 Y가 서로 독립이면 $E(Z) = E(X) + E(Y)$이고 $\sigma^2_z = \sigma^2_x + \sigma^2_y$이다.

(라) 만약 $Z = kX \pm CY$이고 X와 Y가 서로 독립이면 $E(Z) = kE(X) \pm CE(Y)$이고 $\sigma^2_z = k^2 \sigma^2_x + C^2 \sigma^2_y$이다.

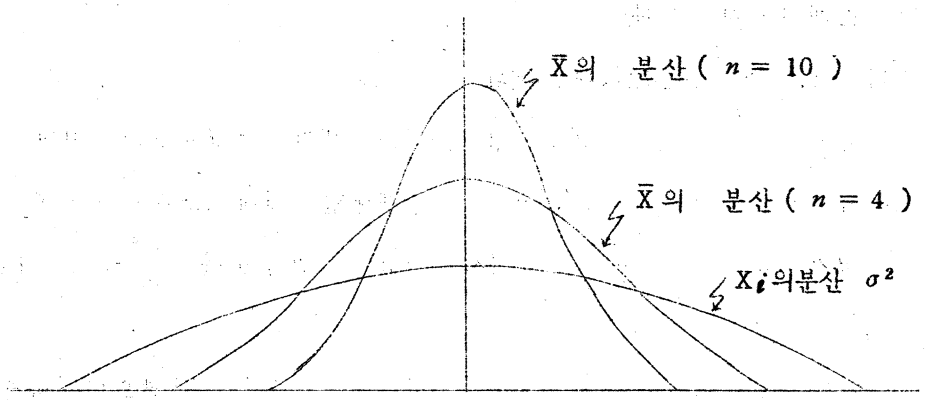
위의 식 (4.3)과 (4.4)에서 얻은 $\overline{X}$의 평균과 분산에 대한 결과를 종합하여 요약하면 다음과 같은 정리를 얻는다.

<정리 4.2> 평균치 μ, 분산 σ^2의 정규분포를 하는 모집단에서 추출한 n개의 임의표본 $X_1, X_2, \cdots\cdots, Xn$에서 얻어지는 표본평균 $\bar{X}$는 평균치 μ, 분산 $\frac{\sigma^2}{n}$인 정규분포를 한다.

위의 정리에 의하면 표본평균으로부터 모집단의 평균을 추출하고져 할 때에는 $\bar{X}$의 분산이 $\sqrt{n}$의 역수에 비례하므로 n이 커짐에 따라서 추정의 정밀도가 <그림 4.1>에 도시한 바와 같이 향상 된다는 것을 알 수 있다. 또한 표본평균의 분포가 $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ 임을 알 때 다음 형태의 변형은 표준정규분포를 함을 <정리 4.1>에 의해 알 수 있을 것이다. 즉

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma}{n}}} \qquad (4.5)$$

은 N(0, 1)이다. 위의 식 (4.5)는 많은 응용력을 가지고 있다. 간단한 예를 들어서 설명하여 보겠다.



<그림 4.1> Xi와 $\bar{X}$의 분포

<예제 4.1> 어느 고등학교 학생들의 키가 대략 평균치 155 ㎝, 표준편차 10 ㎝를 가진 정규분포를 따른다고 하자. 만약 4명의 이 고등학교 학생을 임의로 추출하였을 경우 이 학생들의 평균치가 150 ㎝에서 160 ㎝사이에 있을 확률을 구하라.

<풀 이> 우리가 구하려는 확률은 $P(150 \leq \bar{X} \leq 160)$ 이며 $\bar{X}$의 평균치 $\mu_{\bar{x}} = \mu = 155$, 표준편차가 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{4}} = 5$ 이므로 식 (3.24)와 식 <4.5>를 이용하여

$$P(150 \leq \bar{X} \leq 160) = P\left(\frac{150-155}{5} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{160-155}{5}\right)$$

$$= P(-1 \leq Z \leq 1) = \phi(1) + \phi(1)$$

$$= 2(0.3413) = 0.6826$$

이라는 확률을 얻는다.

다음으로 표본평균 $\bar{X}$의 분포를 다루는데 있어서 매우 중요한 정리를 소개하고자 한다.

<정리 4.3> 中心極限의 定理

평균치 μ, 분산 σ^2을 가진 정규분포가 아닌 모집단에서 임의표본 $X_1, X_2, \cdots\cdots X_n$을 추출할 때에 표본평균 $\bar{X}$는 표본의 크기 n이 증가됨에 따라서 평균치 μ, 분산 $\frac{\sigma^2}{n}$을 가지는 정규분포에 수렴한다.

이 정리는 통계학의 발전에 매우 중요한 공헌을 하였으며 특수 수리통계학 및 통계적 방법론의 확립에 중추적인 역할을 하였다.

이 정리에 의하면 정규분포가 아닌 모든 모집단에 적용되므로 개개의 관찰점 X가 이항분포이든 지수분포이든 관계가 없다. 예를 들어 X가 이항분포를 한다 하자. 앞장에서 배운 바와 같이 $E(X) = np$, $\sigma^2 = npq$ 이므로 n의 수가 커지면 표본평균 $\overline{X}$는 $N(np, \frac{npq}{n}) = N(np, pq)$에 수렴한다는 말이다.

여기에서 한가지 유의하여 둘 사항은 "모집단이 정규분포를 한다."라고 말하면 이 모집단은 무한모집단이다. 왜냐하면 유한모집단은 이론적으로 정확히 정규분포를 할 수 없기 때문이다. 따라서 <정리 4.2>와 같은 것은 무한모집단에 대한 정리이다.

<정리 4.3>은 무한 또는 유한모집단에 모두 적용되는 얘기이며 n이 증가함에 따라서 이론적으로 $\overline{X}$가 정규분포(무한모집단)에 접근한다는 얘기이다.

표본평균 $\overline{X}$와 관련하여 마지막으로 다음의 정리가 의미가 있다. 이는 중심극한의 정리에 근거하여 이루어졌다.

<정리 4.4> 평균치 μ, 분산 σ^2인 어떤 모집단에서 임의표본 $X_1, X_2, \cdots\cdots, Xn$을 추출하여 표본평균 $\overline{X}$를 구하면 다음이 성립한다.

(가) $\mu_{\bar{x}} = \mu$

(나) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ (무한모집단의 경우)

$= \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}}$ (유한모집단의 경우, N은 모집단의 크기)

(다) $\frac{X - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\overline{X} - \mu}{\sigma_{\bar{x}}}$는 $N(0, 1)$에 수렴한다.

(가)는 자명하며 (나)를 설명하기 위하여 <표 4.2,3,4>에 있는 문제를 보자. 식 (4.2)를 사용하여 $\bar{X}$의 표준편차를

$$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}^2} = \sqrt{5.29} = 2.3$$

을 얻을 수 있는데 이 $\sigma_{\bar{x}}$는 직접 <정리 4.4>를 사용하면 모집단이 유한(N = 5)이고 표본의 크기가 $n = 3$이며 $\sigma = \sqrt{30.96} = 5.564$이므로 $\sigma_{\bar{x}}$는

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{5-3}{5-1}} \cdot \frac{5.564}{\sqrt{3}} \fallingdotseq 2.3$$

임을 알 수 있다. (다)항의 설명은 $\bar{X}$가 $N(\mu, \sigma_{\bar{x}}^2)$에 수렴하므로 식 (4.5)와 같은 변수변환을 하면 $\bar{X}$가 $N(0, 1)$에 수렴함을 알 수 있다.

4.3 標本比率의 分布

모집단이 두가지의 특성으로 나누어져 있다고 하자. 예를들어 생산되는 제품이 합격품이거나 불합격품일 경우 또는 사람들의 여론이 찬성이거나 반대이거나 하는 경우 등이 될 것이다. 이때에 어느 한가지 특성에 관심이 있고 이 특성을 가진 구성단위들이 모집단에서 어느정도의 비율을 차지하는가를 알고자 할 때가 많다.

이와같은 비율을 모집단비율이라 하여 P로 나타낸다. 표본의 크기가 n인 단순확률표본을 추출하였을때 n개 중에서 X개가 관심있는 특성을 가졌다고 하자. 이 경우 표본비율 $\bar{P}$를 다음과 같

이 정의한다.

$$\bar{P} = \frac{X}{n}$$

이 정의는 <표 4.1>에서 이미 언급이 되었고 여기에서는 $\bar{P}$가 갖는 표본분포를 다루어 보자. 모집단이 유한인 경우에 $\bar{P}$의 분포는 다음과 같다.

<정리 4.5> 크기가 N인 유한모집단에서 크기가 n인 단순확률표본을 뽑을때 표본비율 $\bar{P}$의 표본분포는 초기하분포를 따르며 확률밀도함수는 다음과 같다.

$$P(\bar{P}) = P\left(\frac{X}{n}\right) = \frac{\binom{NP}{X}\binom{N(1-P)}{n-X}}{\binom{N}{n}}, \quad \begin{matrix} X \leq NP \\ X \leq n \end{matrix}$$

위의 초기하확률분포는 이미 3장의 5절에서 상세히 설명하였으므로 반복을 피하겠다. 만약 모집단이 무한이라면 다음의 정리가 성립한다.

<정리 4.6> 무한모집단으로부터 크기 n인 단순확률표본을 뽑을때 표본비율 $\bar{P}$의 분포는 이항분포를 따르며 확률함수밀도는 다음과 같다.

$$P(\bar{P}) = P\left(\frac{X}{n}\right) = \binom{n}{X} P^x (1-P)^{n-x}, \quad X = 0, 1, 2, \cdots\cdots, n$$

위의 이항분포도 또한 3장에서 자세히 논의하였으므로 여기서는 설명을 생략한다. 이처럼 표본비율 $\bar{P}$는 확률변수로서 유한모집단인 경우는 초기하분포를 하고 무한모집단인 경우는 이항분포를 한다.

이 $\bar{P}$의 평균과 분산을 구하여 보면 다음 정리에 있는 결과를 얻을 수 있다.

<정리 4.7> 표본비율 $\bar{P}$의 기대치와 표준편차는 다음과 같다.

$$E(\bar{P}) = \mu_{\bar{P}} = P$$

$$\sigma_{\bar{P}} = \sqrt{\frac{N-n}{N-1}} \cdot \sqrt{\frac{P(1-P)}{n}}, \text{ (유한모집단일 때)}$$

$$= \sqrt{\frac{P(1-P)}{n}} \text{ (무한모집단일 때)}$$

위에서 반드시 $\bar{P}$의 평균은 유한이거나 무한 모집단이거나 관계없이 모집단비율 P가 되며 $\bar{P}$의 표준편차는 약간의 차이가 있다.

그 차이가 유한모집단의 경우가 $\sqrt{\frac{N-n}{N-1}}$배 만큼 적어지는데 이 계수를 유한수정계수라고 부른다. 이 계수는 N이 매우 커지면 1에 가까워져서 $\sigma_{\bar{P}}$를 구하는데 있어서 유한모집단과 무한모집단간에 차이가 없어진다.

아직까지 표본비율 $\bar{P}$의 확률분포와 그의 평균, 표준편차를 알아보았는데 $\bar{P}$도 하나의 확률변수이기 때문에 중심극한의 정리가 적용된다. $\bar{P}$에서 $\mu_{\bar{P}}$를 빼고 $\sigma_{\bar{P}}$로 나누면 이 변환된 확률변수는 중심극한의 정리에 의하여 표본의 크기 n이 증가함에 따라 표준

정규분포에 수렴할 것이다. 즉

$$\frac{\bar{P}-\mu_{\bar{p}}}{\sigma_{\bar{P}}} \sim N(0,1) \qquad (4.6)$$

여기에서 $\mu_{\bar{p}}$와 $\sigma_{\bar{p}}$는 <정리 4.7>에 있는 바와 같다. 지금까지 논의한 표본비율에 관하여 예제를 풀어보기로 하자.

<예제 4.2> 공장에서 A라는 공정에서 생산되는 제품은 불량품이 0.1이라고 한다. 생산되는 제품중에서 크기 $n=10$인 표본을 단순확률추출하여 표본의 불량률 $\bar{P}$를 관찰한다. 다음 질문에 답하라.

(가) 불량률의 표본비율 $\bar{P}$의 평균과 분산을 구하라.

(나) 임의로 추출된 하나의 표본의 $\bar{P}$가 0.2보다 클 확률을 구하라.

<풀이> 공정에서는 제품을 연속적으로 생산하고 있다고 가정하고 있으므로 제품의 집단은 무한모집단이며 따라서 <정리 4.7>에 의하여

$$\mu_{\bar{p}} = P = 0.1$$

$$\sigma_{\bar{p}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{(0.1)(0.9)}{10}} = 0.0949$$

이 된다. 두번째 질문은 다음의 절차에 의하여 구하면 된다.

$$P(\bar{p}>0.2) = P\left(\frac{\bar{P}-\mu_{\bar{p}}}{\sigma_{\bar{p}}} > \frac{0.2-\mu_{\bar{p}}}{\sigma_{\bar{p}}}\right.$$

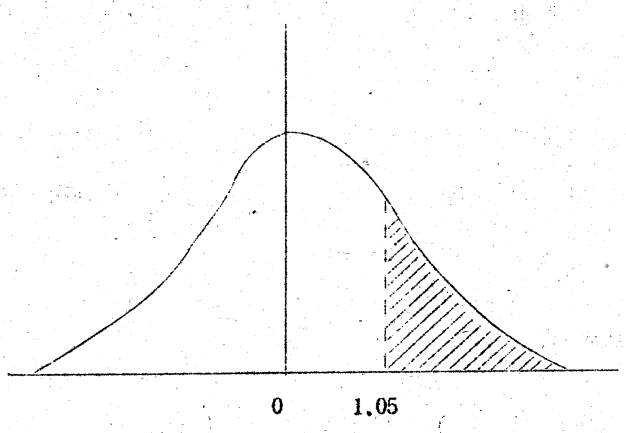
이라 쓸 수 있고 $\frac{\bar{P}-\mu_{\bar{P}}}{\sigma_{\bar{P}}}$는 식 (4.6)에 의하여 대략적으로 N (0 . 1)이므로 이를 표준정규확률변수 Z로 대체하고 $\mu_{\bar{P}}=0.1$ $\sigma_{\bar{P}}=0.0949$로 하면

$$P(\bar{P}>0.2)=P\left(Z>\frac{0.2-0.1}{0.0949}\right)=P(1.05)$$

이 된다. P (Z > 1.05)의 확률은 <그림 4.2 >에서 빗금친 부분의 면적이 된다. 이 면적은 Z의 값이 0에서 +∞까지의 면적에서 Z의 값이 0에서 1.05까지의 면적을 빼면 얻을 수 있다. 즉

$$P(Z>1.05)=\phi(\infty)-\phi(1.05)$$

이다.



<그림 4.2 > P (Z > 1.05 >

그런데 $\phi(\infty)$는 <그림 4.2 >에서 전면적의 반이므로 0.5이고 $\phi(1.05)$는 대부분의 통계학책에 실려있는 표준정규분포표로부터 0.3531임을 알 수 있다. 따라서

$$P(Z>1.05)=0.5-0.3531=0.1469$$

이 된다.

4.4 其他 統計量의 分布

지금까지 표본평균 $\bar{X}$와 표본비율 $\bar{P}$의 분포를 주로 설명하였는데 이외에도 통계적 방법에 많이 쓰이는 통계량들에 대하여 그들의 분포를 알아보자.

4.4.1 標本分散의 分布

먼저 표본분산 S^2을 생각하여 보자. <표 4.1>에 있는 바와 같이 표본의 분산 S^2은 표본이 $X_1, X_2, \cdots, Xn$일 때

$$S^2 = \sum_{i=1}^{n} (Xi - \bar{X})^2 / (n-1)$$

에 의하여 얻어진다. 매 표본마다 n개의 관찰치 $(X_1, X_2, \cdots, Xn)$이 다르므로 S^2의 값이 달라지겠는데 그러면 S^2은 어떤 분포를 따를 것인가? S^2은 음의 값을 취할 수 없기 때문에 좌우대칭인 정규분포는 아니며 x^2분포처럼 변수가 양의 영역에서 어떤 확률밀도함수를 가질 것이다.

통계학의 이론에 의하면 표본분산 S^2에 $(n-1)$을 곱하고 모집단의 분산인 σ^2으로 나누어

$$x^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Xi-\bar{X})^2}{\sigma^2} \qquad (4.7)$$

라는 통계량을 구하면 이 통계량은 자유도 $\phi = n-1$ 인 x^2 분포를 하게 된다. 이 x^2 통계량의 평균과 분산은 3장의 5절에서 본 바와 같이

$$E(x^2) = n-1$$
$$\sigma_{x^2} = 2(n-1) \qquad (4.8)$$

이 된다. 따라서 식 (4.8)과 (4.7)을 관련시켜 표본분산 S^2 의 평균과 분산을 구해보면

$$E(S^2) = \sigma^2$$
$$\sigma^2_{S^2} = \frac{2\sigma^4}{n-1} \qquad (4.9)$$

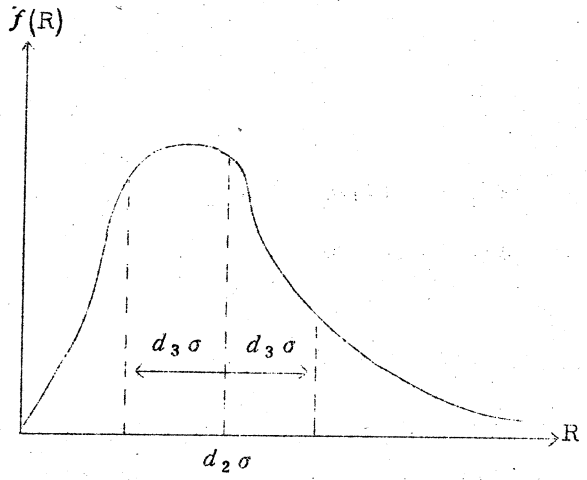
이 된다.

4.4.2 標本範囲의 分布

먼저 R도 표본분산 S^2 처럼 음의 값을 취할 수 없으며 <그림 4.3>에서 표시한 바와 같이 한쪽으로 기운 모양의 분포를 가진다. 범위의 평균치와 분산을 구하면 다음과 같다.

$$E(R) = d_2\sigma$$
$$\sigma^2_R = d_3^2\,\sigma^2 \qquad (4.10)$$

여기서 상수 d_2와 d_3는 표본의 크기 n에 따라 달라지는 상수로서 $n = 2 \sim 5$에 대하여 <표 4.5>에 표시하여 놓았다.



<그림 4.3> 범위(R)의 분포

<표 4.5> d_2, d_3의 값

n	d_2	d_3
2	1.128	0.853
3	1.693	0.888
4	2.059	0.880
5	2.326	0.864

4.4.3 t分布와 F分布

평균치 μ, 분산 σ^2의 정규분포를 하는 모집단으로부터 임의추출한 크기 n의 표본의 평균치 $\bar{X}$는 $N(\mu, \frac{\sigma^2}{n})$의 분포를 함을 앞에서 알아봤다. 또 $\bar{X}$를 표준화하여($\bar{X}$에서 $\mu_{\bar{x}}$를 빼고 $\sigma_{\bar{x}}$로 나누면) Z로 놓으면 이 변수는 N(0, 1)인 분포

를 따른다는 것도 앞에서 말한 바이다. 즉

$$Z = \frac{\overline{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

는 표준정규분포를 갖는다. 여기서 σ 대신에 표본에서 얻어진 표준편차 S를 대입한 것을 t로 놓으면

$$t = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \qquad (4.11)$$

이 되는데 이것은 정규분포를 따르지 않고 자유도 $\phi = n - 1$을 갖는 t 분포를 한다. 이 t분포에 대해서는 앞의 3장에서 논의한 바와 같이 자유도에 따라서 형태가 달라지며 정규분포와 같이 좌우 대칭이다.

분산이 같은 두개의 정규모집단으로부터 취한 크기 n_1과 n_2의 두개의 표본에서 구한 표본분산을 각각 S_1^2, S_2^2로 놓고 그 비례

$$F = \frac{S_1^2}{S_2^2} \qquad (4.12)$$

를 F라 놓으면 이 확률변수 F는 자유도 $\phi_1 = n_1 - 1$, $\phi_2 = n_2 - 1$인 F분포를 하며 ϕ_1과 ϕ_2의 조합에 따라서 분포의 형태가 달라진다. 이 F분포의 평균치와 분산에 대해서는 3장에서 언급한 바와 같다.

△ 연습문제 △

(4.1) 다섯명의 종업원으로 이루어진 회사(하나의 모집단)가 있다. 이 다섯명의 종업원이 일년중에 결석한 수는 다음과 같다.

종 업 원	결석한날수
A	12
B	5
C	0
D	3
E	0

만약 크기 $n=2$인 단순확률표본을 추출하는 경우를 가정하여 다음의 질문에 답하라.

(가) 이 모집단의 평균과 표준편차를 구하라.

(나) 표본평균 $\bar{X}$의 표본분포를 구하라.

(다) 위의 (나)에서 얻어진 표본분포로부터 $\bar{X}$의 평균과 표준편차를 구하라.

(라) 위의 (나)에서 얻은 표본분포에 의하면 $\bar{X}$가 5보다 클 확률은 얼마냐?

(4.2) 한강에서 모래를 채취하여 트럭으로 모래를 운반하는데 한트럭당 실려있는 모래양이 평균 3톤이고 표준편차가 0.5톤인 정규분포를 한다고 한다.

(가) 모래를 운반하는 트럭 9대를 임의 추출할 경우 평균 $\bar{X}$가

갖는 분포는 무엇이냐?

(나) 위의 (가)의 경우에 $\bar{X}$가 2.5톤 미만이 될 확률은 무엇이냐?

(4.3) 어느 가면파티에 4명의 여자와 4명의 남자가 참석하였다. 이 파티에서 임의로 3명을 뽑을 경우 이 표본속에 뽑히는 여자의 표본비율 $\bar{P}$의 분포를 구하라. $n=3$이므로 $\bar{P}$가 취할 수 있는 값은 $0, \frac{1}{3}, \frac{2}{3}, 1$뿐이다. 이 분포에서 $\bar{P}$의 평균과 표준편차를 구하라. 이 문제의 모집단은 $N=8$인 유한모집단임에 유의하시요.

(4.4) 한국의 장년층의 남자들이 담배를 피우는 비율이 70%라고 알려져 있다. 크기가 $n=100$인 단순확률표본을 구하기 위하여 장년층의 남자중 100명을 임의 추출하였다. 이 표본에 뽑힌 사람 중에서 담배를 피우는 사람의 비율 $\bar{P}$를 구해 보았다.

(가) 이 표본비율 $\bar{P}$의 분산은 무엇이냐?

(나) 이 표본비율 $\bar{P}$는 대략적으로 정규분포를 따르겠는가?

(다) 이 표본비율 $\bar{P}$가 0.75보다 클 확률은 무엇이냐?

(라) 표본의 크기 n이 정해져 있지 않은 단계에서 $\bar{P}$의 표준편차 $\sigma_{\bar{P}}$를 0.02정도로 하고 싶다면 표본의 크기 n을 얼마로 정해 주어야 할 것이냐?

V章 : 標 本 調 査

통계적 개념의 기초가 되는 것은 모집단과 표본이다.

모집단에 대해서 어떤 지식을 얻고자 할 때 모집단 전체에 대한 센서스 또는 完全調査를 통해서 우리가 원하는 정보를 얻을 수 있다. 그러나 실제로 현실에서는 비용, 시간, 자격을 갖춘 조사원의 부족 등의 문제점이 있어 우리는 모집단에서 표본을 추출하여 모집단에 대한 統計的 推論이나 의사결정을 하게 된다.

이 장에서는 여러가지 표본조사 방법을 소개하고 추출된 표본을 분석하는 방법을 알아 보고자 한다.

표본조사에서의 추론법은 확률표본에 기초를 둔 추론법과는 다름에 주의해야 한다. 즉 표본조사에서는 대부분 유한모집단을 대상으로 하여 非復元抽出法을 많이 사용하므로 엄격한 의미에서 확률표본에서 요구되는 독립성이 만족되지 않는 것이 특징이다.

표본조사에서 우선적으로 생각해야 할 문제는 다음과 같다.

(가) 비용을 가능한 적게 하면서도 모집단의 성질을 올바르게 반영하는 추출법은 무엇인가?

(나) 표본에 있는 정보를 유용하고 명백하게 설명하는 방법은 무엇인가?

(다) 표본의 정보를 기초로 하여 모집단에 대한 의사결정과 추론을 어떻게 하는가?

(라) 여기서 얻은 추론이나 결론의 신뢰성은 어떠한가?

5.1 母集団의 設定과 標本抽出

어떤 문제에 대하여 표본조사를 통해서 정보를 얻고자 할 때, 우선 대상(모집단)을 명확하게 설정해야 하며 그 대상 전체에서 표본이 추출되어야 한다. 이러한 추론의 대상이 되는 모집단을 対象母集団이라 한다.

모집단의 설정이 쉬운 작업같으나 실제로는 여러가지 어려운 점이 있다. 예를들어 대학생의 의식구조에 관한 여론조사를 할 때, 초급대학이나 전문대학의 포함여부를 결정해야 하며 대학원생이나 청강생의 여부도 결정해야 한다. 더구나 특정지역의 특정대학에서 표본을 추출한다면, 거기서 얻어지는 결론은 우리나라 전체 대학생의 의식구조와 많은 차이가 있을 수 있다.

표본추출을 통한 통계적 방법을 사용하기 위해서는 채취과정에 반드시 任意性이 들어가야 한다. 따라서 모집단의 구성요소인 각 抽出単位 또는 추출단위의 집합이 표본가운데 포함될 확률이 얼마인가를 알 수 있어야 한다.

이러한 조건을 만족시키는 추출법을 확률추출이라 하며 확률추출에 의해서 얻어진 표본을 確率標本이라 한다. 확률표본이 아닌 것을 비확률표본이라 하며 비확률표본의 경우에는 추정량의 분산을 모르는 단점이 있다.

확률추출에서 가장 기본이 되는 추출법으로 단순확률추출과 層化抽出이 있다. 이외에 系統抽出, 集落抽出 등이 있다.

5.2 偏倚의 原因과 乱數表의 사용

확률추출에서 추정량의 기대값과 추정하는 모수와의 차이를 偏倚라 하며 이런 편기없는 추정량을 不偏推定量이라 한다.

편기의 원인은 매우 다양하다. 예를 들어 추정량 자체가 편기된 추정량이거나 잘못 선택된 측량도구를 사용한다거나 설문조사에서 질문방법이 잘못되었거나, 부적당한 면담방법을 사용했을 때는 편기된 추정량을 얻게되며 따라서 결과의 오차도 커진다.

그러나 가장 중요한 편기의 원인은 표본설계의 잘못이다. 이에 대한 전형적인 예로서 1936년 미국 대통령선거에 대한 Literary Digest사의 여론 실패담이 있다. Literary Digest사는 전화번호부와 자동차 등록대장에서 추출한 200만명 이상의 유권자를 대상으로 여론조사를 실시한 결과 Landon후보가 Roosevelt 후보에 압도적으로 승리하리라는 것이었으나 투표결과는 Roosevelt 후보가 압도적으로 승리했다.

1936년대에 미국에서 전화나 자동차를 소유하는 것은 상류층이었으며 Landon은 상류층에 인기가 있었고 Roosevelt는 서민층에 인기가 있었기 때문에 Literary Digest社의 표본설계는 근본적으로 잘못된 것이었다.

설문조사 등에서 비응답자가 많을 때도 심각한 편기가 생길 수 있다. 이 경우엔 비응답자에 대한 표본을 다시 추출하여 직접 면담등을 통해서 이들에 대한 추정값을 얻어 전체표본을 수정해

출 필요가 있다.

대상모집단이 결정되고 표본추출법이 선정되었다 하더라도 N 개의 단위로 구성된 모집단에서 n 개의 표본을 임의로 추출하는 것이 기술적으로 쉬운 일이 아니다. 1 부터 N 까지의 카드를 만들어 잘 섞은 다음 이중 임의로 n 개를 추출하는 방법등 여러가지 방법이 있을 수 있으나 난수표를 사용하는 것이 좀 더 체계적이고 편리한 방법이다.

예를들어 1 부터 75 까지의 숫자중에서 5 개의 숫자를 임의로 취하고 싶다고 하자. 난수표로부터 다음과 같은 두자리 숫자를 얻었다.

09,81,67,56,19,67,88,99,35,……

이 가운데 75 가 넘는 숫자와 중복되는 숫자를 버리고 차례로 다음과 같은 다섯개의 숫자를 얻을 수 있다.

9,67,56,19,35

또한 많은 난수를 사용할 경우에는 확장된 난수표를 사용하거나 전자계산기를 이용하여 擬似亂數를 생성해서 사용하기도 한다.

5.3 單純確率抽出

N 개의 추출단위가 있는 모집단에서 n 개의 표본을 비복원추출할 때 $\binom{N}{n}$ 개의 모든 가능한 같은 확률로서 나타나는 확률추출을 단순확률추출이라 한다.

일반적으로 복원추출보다 비복원추출이 더 바람직한데 그 이유를 간단한 예로 알아보자.

지금 관심의 대상이 되는 모집단이 4개의 가구(추출단위)로 이루어져 있으며 각 가구의 식구수(특정치)가 아래와 같이 이루어져 있다고 생각하자.

$$\chi_1^* = 5 \ , \ \chi_2^* = 3 \ , \ \chi_3^* = 1 \ , \ \chi_4^* = 2$$

이 모집단의 가구당 식구수에 관한 조사를 하기 위해 두가구를 임의로 추출할 경우 복원추출과 비복원추출의 각각에 대한 모든 가능한 특성치 값들은 다음과 같다. (<표 5.1>)

<표 5.1>

복 원 추 출	비 복 원 추 출
{5,5} {3,3} {1,1} {2.2}	{5,3} {3,1} {1,2}
{5,3} {3,1} {1,2}	{5,1} {3,2}
{5,1} {3,2}	{5,2}
{5,2}	

<표 5.1>에서 $\bar{X} = (X_1 + X_2) / 2$가 모평균인 (5 + 3 + 1 + 2) / 4 = 2.75에 얼마만큼 가까이 분포되어 있나를 두가지 경우에 대해 각각 알아보자.

복원추출의 경우 : $\bar{X} = (X_1 + X_2)/2$의 분포

$\bar{x}$의값	1	1.5	2	2.5	3	3.5	4	4.5
확 률	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

$$E(\bar{X}) = 1 \times \frac{1}{16} + 1.5 \times \frac{2}{16} + \cdots\cdots + 5 \times \frac{1}{16} = 2.75$$

$$E(\bar{X}^2) = 1^2 \times \frac{1}{16} + 1.5^2 \times \frac{2}{16} + \cdots\cdots + 5^2 \times \frac{1}{16} = 8.656$$

$$Var(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2 = 8.656 - (2.75)^2 = 1.094$$

비복원추출의 경우 : $\bar{X} = (X_1 + X_2)/2$의 분포

$\bar{x}$의값	1.5	2	2.5	3	3.5	4
확 률	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$E(\bar{X}) = 1.5 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots\cdots + 4 \times \frac{1}{6} = 2.75$$

$$E(\bar{X})^2 = 1.5^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + \cdots\cdots + 4^2 \times \frac{1}{6} = 8,292$$

$$Var(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2 = 8.292 - (2.75)^2 = 0.729$$

위의 예에서 보는 바와 같이 두가지 경우에 기대값은 2.75로 같으나 분산은 비복원추출의 경우가 더 작다. 이것은 일반적으로 어느 모집단에서나 성립되며 따라서 복원추출보다는 비복원추출이 더 바람직한 추출법이 된다.

일반적으로 모집단에서 N개의 추출단위를 $u_1, u_2, \cdots\cdots, u_N$이라 하고 이들의 특성치를 $\chi_1^*, \chi_2^*, \cdots\cdots \chi_N^*$라 하자. 그러면 이들 특성치의 평균과 분산은 다음과 같다.

$$\text{모평균} : \mu = \frac{\sum_{i=1}^{N} \chi_i^*}{N}$$

$$\text{모분산} : \sigma^2 = \frac{\sum_{i=1}^{N} (\chi_i^* - \mu)^2}{N-1} \qquad (5.1)$$

엄격한 의미에서 모분산의 정의에 따라 분모가 N이어야 되나 편의상 (N−1)을 사용하기로 한다.

표본조사에서 알고자 하는 값은 모평균 μ이며 이에 대한 추론은 표본평균 $\bar{X}$에 기초를 둔다. 또한 실제로 모분산을 모를 경우 모분산 σ^2은 표본분산 S^2으로 추정한다.

単純確率標本 : $X_1, X_2, \cdots\cdots, X_n$

標本平均 : $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

標本分散 : $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \qquad (5.2)$

위와같이 표본평균과 표본분산을 정의하면 $\bar{X}$와 S^2은 각각 μ와 σ^2에 대한 불편추정량이 되며 $\bar{X}$의 분산은 다음과 같다.

$$\mathrm{Var}(\bar{X}) = \sigma^2 \left(\frac{1}{n} - \frac{1}{N}\right) = \frac{\sigma^2}{n}(1-f), \quad \text{단} \quad f = \frac{n}{N} \qquad (5.3)$$

여기서 $f = n/N$은 抽出率이라 하며 $(1-f)$를 有限母集団修正項이라 한다.

여기서 간단히 단순확률추출에서 $\bar{X}$와 S^2의 성질을 알아보자.

$$E(\bar{X}) = \mu$$

$$E(S^2) = \sigma^2$$

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}(1-f), \text{단} \quad f = \frac{n}{N} \qquad (5.4)$$

$$\text{추정된} \quad \mathrm{sd}(\bar{X}) = \frac{S}{\sqrt{n}}\sqrt{1-f}$$

위에서 보듯이 $\bar{X}$의 표준편차는 $1/\sqrt{n}$에 비례하여 감소하므로 모집단의 크기에 관계없이 수천개의 표본으로도 정확한 추정을 할 수 있음을 유의하자. 또한 μ에 관한 95 % 신뢰구간은 대략

$$\bar{x} \pm 2\frac{S}{\sqrt{n}}\sqrt{1-f}, \quad \text{단} \quad f = \frac{n}{N} \qquad (5.5)$$

이 되며 이 측정값은 n와 $N-n$이 클 때는 매우 정확하다.

<예제 5.1> 어느 국민학교에서 학생들의 1일 영양섭취에 대한 조사를 하기 위하여 표본조사를 실시했다. 이 국민학교의 재학

생 3600명 가운데 100명을 단순확률추출로서 선정했다. 이 학생들의 부모의 협조를 얻어 각자의 식사내용을 일주일 동안 기록하게 하고 이를 기초로 하여 각 개인의 1일 평균 영양섭취량을 칼로리로 환산하여 다음과 같은 결과를 얻었다고 하자.

$\bar{x} = 752$

$S = 138$

그러면 95% 근사신뢰구간은

$$\bar{x} \pm 2\frac{S}{\sqrt{n}}\sqrt{1-f} = 752 \pm 2 \cdot \frac{138}{\sqrt{100}}\sqrt{1-\frac{100}{3600}}$$

$$= 752 \pm 27.2$$

또는

(724.8,779.2)

이 된다. 여기서 수정항 $(1-f)$를 생략하면 신뢰구간은 (724.4 779.6)이 된다.

실제로 모집단의 크기인 N에 비하여 n이 아주 작을 경우에는 $1-f \fallingdotseq 1$이므로 수정항 $(1-f)$를 생략해도 좋다.

5.4 比率推定을 위한 標本調査

실업율이나 특정정당에 대한 지지율, 납품된 제품의 불량율 등과 같이 有限母集団에서 어느 속성에 대한 비율을 알고자 할 때가 많다. 이러한 경우 속성을 갖는 추출단위에 1을 부여하고

나머지 단위에 0을 부여했을 때 X를 n개의 표본 가운데 그 속성을 갖는 단위의 갯수로 정의하면 X는 超幾何分布를 따른다.

	모 집 단	표 본
단위의 갯수	N	n
속성을 갖는 단위의 갯수	M	X
속성의 비율	$P = \frac{M}{N}$	$\hat{P} = \frac{X}{n}$

표본비율 $\hat{P} = X/n$의 기대값과 분산은 각각

$$E(\hat{P}) = P$$

$$Var(\hat{P}) = \frac{P(1-p)}{n}\left(\frac{N-n}{N-1}\right) \qquad (5.6)$$

이며 $Var(\hat{P})$에 대한 근사값으로는 P대신 X/n를 대입해서 사용하기도 하나 불편추정량인

$$\hat{P}(1-\hat{P})\frac{(N-n)}{N(n-1)} \qquad (5.7)$$

을 많이 사용한다. 이를 이용하여 속성비율 P와 속성단위갯수 M에 대한 95 % 신뢰구간을 구하면 다음과 같다.

$$\hat{P} = X/n$$

P의 95% 근사신뢰구간

$$\hat{P} \pm 2\sqrt{\hat{P}(1-\hat{P})\frac{N-n}{N(n-1)}} \qquad (5.8)$$

$$\hat{M} = N\frac{X}{n} = N\hat{P}$$

M의 95% 근사신뢰구간

$$N\hat{P} \pm 2\sqrt{N\hat{P}(1-\hat{P})\left(\frac{N-n}{n-1}\right)} \qquad (5.9)$$

위의 공식에서 N이 크고 $\frac{n}{N}$이 작을 때는 $\hat{P}=\frac{X}{n}$ 의 분산은 $P(1-P)/n$에 근사하므로 $\frac{1}{n}$에 비례하여 감소한다. 즉 앞에서 언급했듯이 모집단의 크기에 관계없이 수천가지 표본으로도 정확한 추정을 할 수 있다. 또 통계오차에는 표본오차 외에도 여러가지 비표본오차가 있고 표본의 크기가 증가함에 따라 비표본오차가 커지는 경우가 많다. 즉 적당한 크기의 표본을 정하여 비표본오차가 유발될 수 있는 조사항목에 대해서는 우수한 조사원으로 하여금 철저히 관리하게 함으로써 비표본오차를 줄이는 것이 모집단 전체를 조사하거나 또 지나치게 큰 표본을 선택하는 것보다 효과적일 때가 많다.

<예제 5.2> 어느 큰 회사에 1024개의 전구가 납품되었다. 이 전구의 불량율을 검사하기 위해서 60개를 임의로 추출하여 불량품의 갯수를 조사한 결과 12개의 불량품이 발견되었다면 불량율 P의 점추정과 구간추정은 다음과 같다.

우선 P의 점추정 값은

$$\hat{P} = \frac{X}{n} = \frac{12}{60} = 0.20$$

$\hat{P}$의 분산에 대한 추정값은

$$\hat{P}(1-\hat{P})\frac{(N-n)}{N(n-1)} = (0.2)(0.8)\frac{(1024-60)}{1024(60-1)} = 0.0026$$

이 된다. 따라서 95%의 오차한계는 $\pm 2\sqrt{0.0026} = \pm 0.10$이며 95% 근사신뢰구간은 (0.10,0.30)이다.

5.5 層化確率抽出

모집단을 동질적인 몇개의 집단 또는 층으로 나눈 다음 각 층에서 어떤 규칙에 의하여 표본을 추출하는 것이 층화확률추출이다. 모집단을 동질적인 층으로 나눌 수 있을때 층화추출법을 사용하면 추정량의 분산을 줄일 수 있는 장점이 있다.

층을 나누는 기본 원리는 층간의 변이성을 크게하고 층내부에서의 변이성을 작게 하는 것이다. 따라서 각 층에서는 상대적으로 작은 표본을 갖고도 정확한 추정을 할 수 있게 된다.

예를 들어 한 도시의 월 생계비에 대해서 조사하고 싶다고 하자. 만약 도시의 동쪽에 고소득층이 많이 살고 서쪽에 저소득층이 많이 산다고 하면 전도시에 대한 단순확률추출보다는 도시를 동서의 두 층으로 나누어 층화확률추출을 하는 것이 지역별 비교

도 되고 전체적인 추정도 더 정확히 할 수 있다.

지금 N개의 추출단위로 된 모집단이 $N_1, N_2, \cdots\cdots N_h$ 개의 단위로 구성된 층으로 나누어져 각 층의 평균과 분산을 각각 μ_i 와 σ_i^2 이라 하자. 또한 각 층에서 크기가 각각 $n_1, n_2, \cdots\cdots n_h$ 인 표본을 단순확률추출하기로 하고 i 번째 층에서 j 번째 관측치를 X로 표시하자.

모집단의 구성과 통계량은 <표 5.2>와 같다.

<표 5.2> 모집단의 구성과 통계량

		층(strata)			
		1	2	……	h
모집단:	크 기	N_1	N_2	……	Nh
	평 균	μ_1	μ_2	……	μh
	분 산	σ_1^2	σ_2^2	……	σ_h^2
표 본:	크 기	n_1	n_2	……	nh
	표본평균	$\bar{X}_1$	$\bar{X}_2$	……	$\bar{X}h$
	표본분산	s_1^2	s_2^2	……	sh^2

$$N = \sum_{i=1}^{h} N_i, \quad \text{母平均} \quad \mu = \frac{1}{N} \sum_{i=1}^{h} N_i \mu_i$$

$$\bar{X}i = \frac{1}{ni} \sum_{j=1}^{ni} X_{ij}, \quad Si^2 = \frac{1}{ni-1} \sum_{j=1}^{ni} (X_{ij} - \bar{X}_i)^2$$

각 층에서 표본평균을 $\bar{X}i$ 라 하면 $E(\bar{X}) = \mu i$ 이며 분산은

$$Var(X_i) = \frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

가 된다. 전체 모집단의 평균은 각 층의 평균에 대한 가중평균으로서

$$\mu = \frac{n_1}{N}\mu_1 + \frac{n_2}{N}\mu^2 + \cdots\cdots + \frac{nh}{N}\mu h \qquad (5.10)$$

이 되므로 μ에 대한 불편추정량은

$$\bar{X}st = \frac{N_1}{N}\bar{X}_1 + \frac{N_2}{N}\bar{X}_2, \cdots\cdots + \frac{Nh}{N}\bar{X}h$$

로 주어진다. 각 층에서 추출된 표본은 서로 독립이므로 $\bar{X}st$의 분산은 다음과 같다.

$$Var(\bar{X}st) = \sum_{i=1}^{h} Var\left(\frac{Ni}{N}\bar{X}_i\right) = \frac{1}{N^2}\sum_{i=1}^{h} N_i^2\, Var\,(\bar{X}i)$$

$$= \frac{1}{N^2}\sum_{i=1}^{h} Ni^2\,\frac{\sigma^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

$$= \frac{1}{N^2}\sum_{i=1}^{h} N_i(N_i - n_i)\frac{\sigma_i^2}{n_i} \qquad (5.11)$$

로 주어진다. 또 μ에 대한 95% 근사신뢰구간은

$$\bar{X}st \pm \frac{2}{N}\sqrt{\sum_{i=1}^{h} N_i(N_i - n_i)\,\frac{S_i^2}{n_i}} \qquad (5.12)$$

이 된다.

실제로 층화추출법이 단순추출법보다 좋은 이유를 다음 예에서 생각해 보자.

<예제 5.3> 어느 과수원에 64그루의 사과나무가 있는데, 그 중 40그루는 완전히 자란 나무이고 24그루는 덜 자란 나무이다. 지난해에 각 나무당 수확량을 기록해 둔 결과가 다음과 같다. (<표 5.3>, 단위는 상자이다.)

<표 5.3>

층 1 (성숙한 나무)	8	7	5	6	6	10	7	6
	5	4	4	7	6	6	3	8
	4	7	8	10	8	4	6	6
	6	4	6	4	8	7	9	8
	6	3	9	9	7	8	11	9
층 2 (미성숙한 나무)	5	3	4	3	5	2	4	3
	5	3	3	4	5	4	3	4
	4	5	4	3	3	4	3	5

이와같은 모집단에서 모평균 μ를 모른다는 가정하에 단순확률추출과 층화추출을 한다고 하자. 64그루 전체에 대한 평균과 분산은 각각

$\mu = 5.5625$

$\sigma^2 = 4.66$

이다. 또한 층 1과 층 2의 각각에 대한 평균과 분산은

$\mu_1 = 6.625$

$\sigma_1^2 = 3.986$

$\mu_2 = 3.792$

$\sigma_2^2 = 0.781$

이다. 지금 전체에서 8개를 단순확률추출하는 경우와 층 1에서 6개 그리고 층 2에서 2개를 층화확률추출하는 경우의 각각에 대한 표본평균의 분산은 다음과 같다.

(가) 단순확률추출

$$Var(\bar{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$$

$$= \frac{4.66}{8}\left(1 - \frac{8}{64}\right) = 0.51$$

(나) 층화확률추출

$$Var(\bar{X}) = \frac{1}{N^2}\left[N_1(N_1 - n_1)\frac{\sigma_1^2}{n_1} + N_2(N_2 - n_2)\frac{\sigma_2^2}{n_2}\right]$$

$$= \frac{1}{64^2}\left[40(40-6)\frac{3.98}{6} + 24(24-2)\frac{0.781}{2}\right]$$

$$= 0.27$$

위에서 보는 바와 같이 표본의 크기 $n=8$에 대한 표본평균의 분산은 층화추출의 경우가 더 작게 됨을 알 수 있다.

5.6 標本크기의 割当

층표본크기 n은 비용, 시간 등의 제약에 의해 결정되나 추정량의 분산을 작게 하도록 n개를 각 층에 할당하는 것이 바람직할 것이다. 우선 각 층의 크기에 비례해서 각 층의 표본의 크기를 할당하는 比例割当을 생각할 수 있다. 즉 i번째 층의 표본의 크기 n_i는 다음과 같이 주어진다.

$$n_i = n\left(\frac{N_i}{N}\right), \quad i = 1, 2, \cdots\cdots, h \tag{5.13}$$

이 방법에 의하면 각 층이 층의 크기에 따라 가중되는 장점이 있으며 실제로 많이 사용되고 있다.

다음은 추정량 $\bar{X}st$의 분산을 최소로 하는 할당법을 알아보자. 추정량의 분산을 최소로 하기 위해서는 분산이 큰 층에 상대적으로 더 많은 표본크기를 할당하는 것이 바람직하다. 즉 i번째 층의 분산이 σi^2이면 i번째 층의 표본크기를 $N\sigma i$에 비례하도록 할당하는 방법을 最適割当이라 한다. 최적할당을 할때 i번째 층의 표본의 크기 ni는

$$ni = n\frac{N_i\ \sigma_i}{\sum_{j=1}^{h} N_j\ \sigma_j} \tag{5.14}$$

이 된다.

<예제 5.4> 인구가 각각 $N_1 = 4$ 만, $N_2 = 2$ 만, $N_3 = 3$ 만인 세 도시에서 $n = 400$ 명의 표본을 층화추출하려고 한다. 다음에 대한 각 층별 표본크기를 결정하여라.

(가) 비례할당

(나) 표준편차가 $\sigma_1 = 20, \sigma_2 = 12, \sigma_3 = 14$ 라는 가정하에 최적할당

<풀 이>

(가) 비례할당

$$n_1 = n\left(\frac{N_1}{N}\right) = 400\left(\frac{4}{9}\right) = 178$$

$$n_2 = n\left(\frac{N_2}{N}\right) = 400\left(\frac{2}{9}\right) = 89$$

$$n_3 = n\left(\frac{N_3}{N}\right) = 400\left(\frac{3}{9}\right) = 133$$

(나) 최적할당

$$\Sigma n_i \sigma_i = 800,000 + 240,000 + 420,000$$

$$= 1,460,000$$

$$n_1 = n\frac{N_1 \sigma_1}{\Sigma N_i \sigma_i} = 400\left(\frac{800}{1460}\right) = 219$$

$$n_2 = n\frac{N_2 \sigma_2}{\Sigma N_i \sigma_i} = 400\left(\frac{240}{1460}\right) = 66$$

$$n_3 = n\frac{N_3\ \sigma_3}{\Sigma N_i\ \sigma_i} = 400\left(\frac{420}{1460}\right) = 115$$

5.7 比率推定을 위한 層化抽出

모집단에서 어느 속성에 대한 비율을 추정하기 위한 층화추출법을 소개하기 위하여 앞에서와 같은 기호를 사용하기로 하고 각 층에서 속성의 비율을 $P_1, P_2, \cdots\cdots, Ph$ 라 하자. 표본크기의 할당은 비례할당인 $n_i = n(N_i/N)$ 을 사용할 수도 있고 $\hat{P}st$ 의 분산을 최소로 하는 최적할당인

$$ni = n\frac{N_i\sqrt{P_i(1-P_i)}}{\sum_{j=1}^{h} N_j\sqrt{P_j(1-P_j)}} \qquad (5.15)$$

를 사용할 수도 있다. 물론 최적할당을 사용하기 위해선 P_i 의 근사값에 대한 사전정보가 있어야 한다. <표 5.4>와 같이 층으로 나누어 층화추출에서의 비율에 관한 추론을 알아보자.

<표 5.4 >　　층(Strata)

	1	2 ------- h	
크　기	N_1	N_2 ..... Nh	$N = \sum_{i=1}^{h} N_i$
모집단: 비　율	P_1	P_2 ..... Ph	$P = \sum_{i=1}^{h} \frac{N_i P_i}{N}$
크　기	n_1	n_2 .....nh	$n = \sum_{i=1}^{h} n_i$
표 본:관 측 치	X_1	X_2 ..... Xh	
표본비율	$\hat{P}_1 = \frac{X_1}{n_1}$	$\hat{P}_2 = \frac{X_2}{n_2}$ ... $\hat{P}h = \frac{Xh}{nh}$	

$$\hat{P}st = \frac{N_1}{N}\hat{P}_1 + \frac{N_2}{N}\hat{P}_2 + \cdots\cdots + \frac{Nh}{N}\hat{P}h = \frac{1}{N}\sum_{i=1}^{h} N_i \hat{P}_i$$

95 %　근사오차한계 :

$$\pm \frac{2}{N}\sqrt{\sum_{i=1}^{h} \frac{N_i(N_i - n_i)}{(n_i - 1)}\hat{P}_i(1 - \hat{P}_i)} \qquad (5.16)$$

<예제 5.5 > 어느 대학에서 학부 졸업생 가운데 졸업후 취직을 원하는 재학생의 비율을 추정하고자 200 명을 임의로 추출하여 이들에게 설문지를 우송했다. 이 가운데 160 명이 회신을 했고 160 명 가운데 108 명이 취업을 원했다고 하자. 무응답자 40 명을 하

나의 층으로 간주하고 이 가운데서 4명을 임의로 추출하여 면담을 해 본 결과 1명만이 취업을 원했을때, 취업을 원하는 학생의 비율을 추정해 보자.

$$\hat{P}st = \frac{N_1}{N}\hat{P}_1 + \frac{N_2}{N}\hat{P}_2 = 0.8\left(\frac{108}{160}\right) + 0.2\left(\frac{1}{4}\right) = 0.59$$

이며 $(1 - n_1/N_1)$과 $(1 - n_2/N_2)$를 1로 간주했을때 95% 근사오차한계는

$$\pm 2\left\{\frac{(0.8)^2\left(\frac{108}{160}\right)\left(\frac{52}{160}\right)}{159} + \frac{(0.2)^2\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)}{3}\right\}^{\frac{1}{2}} = \pm 0.116$$

5.8 其他 標本抽出法

앞에서 소개한 단순확률추출법과 층화추출법 이외에 많이 사용되는 표본추출법으로는 계통추출법과 집락추출법이 있다.

계통추출법은 표본을 시간적으로나 공간적으로 일정한 간격을 두고 취하는 추출법으로 방법이 간편하고 모집단 전체에서 표본이 골고루 추출되는 장점이 있다. 그러나 모집단의 순서 가운데 예기치 못했던 주기성이 있다면 계통추출에 의한 결과는 상당히 편기된 결과가 될 수 있음에 유의하여야 한다.

집락추출은 우선 모집단을 집락이라는 부분집단으로 나눈다음 임의로 몇개의 부분집단을 추출하고 추출된 각 부분집단에서 확률추

출을 하는 방법이다. 층화추출과는 반대로 집락추출에서는 각 부분집단들 사이에 변이성이 적어야만 선택된 부분집단에서 모집단의 특성을 올바르게 추출해 낼 수 있다. 이 방법은 비용절감을 주로 목적으로 한다. 또한 추출된 집단 내에서 다시 집락추출을 적용시키는 多段抽出法도 많이 사용된다.

지금까지 여러가지 표본조사의 방법을 소개했는데 실제로 비용, 시간, 효율성 등 여러가지를 고려해서 이런 방법들을 복잡한 형태로 사용하는 것이 보통이다.

△ 연습문제 △

(5.1) 다음 각 경우에 대하여 대상모집단과 추출단위를 정의하고 적당한 표본설계를 하여라.

(가) 어느 공장에서 생산되는 제품의 품질을 관리하기 위하여 수시로 표본조사를 하려고 한다.

(나) 어느 TV방송국에서 제작한 특집 프로그램에 대한 시청율을 조사하고 싶다.

(대) 국회의원 선거에서 어느 특정 선거구의 후보자들에 대한 지지도를 조사하고 싶다.

(5.2) 확률변수 X가 다음과 같은 확률분포를 갖고 있다.

x	2	4	6	8
$f(x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

(가) 이 모집단의 평균과 분산을 구하여라.

(나) 복원추출을 통하여 2개의 표본을 임의로 추출할 때 $\bar{X}$의 분포함수, 평균, 분산을 구하라.

(다) 비복원추출의 경우에는 어떠한가?

(5.3) 4000가구가 있는 한 도시에서 가구당 평균 식구수를 조사하기 위하여 360가구를 단순확률추출하여 조사한 결과 평균 식구수가 3.45명, 표준편차가 1.07이었다. 이 도시의 가구당 평균 식구수에 대한 95% 신뢰구간을 구하여라.

(5.4) 32개의 철선의 인장강도가 다음과 같이 주어졌다.

25, 20, 35, 21, 22, 22, 24, 25

30, 28, 24, 20, 20, 25, 20, 19

25, 23, 20, 24, 28, 24, 24, 22

28, 26, 23, 25, 22, 27, 25, 23

(가) 난수표를 사용하여 10개의 표본을 단순확률추출하고, 이 표본에서 모집단의 평균과 분산을 추정하여라.

(나) 모집단의 분산은 $\sigma^2 = 13.97$이다. 크기 10인 단순확률표본에서 $\bar{X}$의 분산은 얼마인가? 크기가 20이면 어떤가?

(5.5) 한 소도시의 가구당 월 생계비를 조사하기 위하여 도시를 생활수준에 따라 4개의 층으로 나눈다음 층화추출을 통하여 다음 결과를 얻었다.

층(strata)

	Ⅰ	Ⅱ	Ⅲ	Ⅳ
층의 크기	3750	3272	1387	2475
표본 크기	50	45	30	30
표본 평균	12.6	14.5	18.6	13.8
표본 분산	2.8	2.9	4.8	3.2

가구당 평균 생계비에 대한 추정값을 구하고 95% 신뢰구간을 구하여라.

(5.6) 한 대학에서 새로 제정하는 교칙에 대한 지지율을 조사하기 위하여 표본조사를 실시했다. 각 학년과 대학원생을 층으로 취급하여 다음 결과를 얻었다.

층(strata)

	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
층의크기	3500	3000	2800	2500	1800
표본크기	50	50	40	40	30
찬성한수	42	45	28	26	14

전교생의 지지율에 대한 추정값을 구하고, 95% 신뢰구간을 구하여라.

Ⅵ章 : 指　　數

지수란 일종의 통계비율로 통일한 통계계열의 2숫자의 크기를 시간적 차이에 따라 비교하는 비율을 말한다. 이것은 시계열의 변동을 상대적으로 표시하기 위하여 계산되는 경우가 많다. 이 때 분모로 하는 시점을 기준 또는 기준시라 한다.

지수의 기준은 일정한 시점에 이것을 고정시키는 것이 보통이다. 이와 같이 어떤 일정한 시점이 기준으로 사용되는 경우를 고정기준이라고 하며 연차지수에서 1개년의 값 또는 수개년의 평균값, 월차지수에서 1개월의 값 또는 12개월의 평균값을 고정기준으로 사용할 때를 각각 고정단초기준, 고정광초기준이라고 한다. 여기에 대하여 계열의 직전값을 기준으로 각기의 지수를 계산할 때 그 기준을 연쇄기준이라고 한다. 연쇄기준으로 계산한 지수를 연환지수라고 하고, 최초의 기준시의 지수와 비교하려면 기준시를 옮기면서 순차로 연승해야 한다. 연환지수를 연승하여 계산한 것을 연쇄지수라고 한다. 연쇄기준으로 계산되는 지수는 품목이나 가중치가 변경될 경우에 편리하지만 기준이 매번 옮겨지므로 계열 전체의 일괄적 비교를 할 수 없으므로 일반적으로 고정기준지수를 많이 사용한다.

지수는 개별지수와 종합지수로 나누어 볼 수 있다. 개별지수란 개개의 시계열에 관한 변동을 상대적으로 계산한 지수이고 종합지수란 몇가지의 시계열변동을 종합적으로 계산한 지수이다. 지수라고

하면 종합지수를 지칭하는 것으로 직접 측정할 수 없는 양의 변동을 간접적으로 표시한 숫자이다.

지수는 처음에 물가수준의 변동을 측정하기 위하여 고안된 것이나 이어서 생산량, 거래량 등의 측정과 보다 추상적인 경제활동의 변동을 파악하는 방법으로 쓰이게 되었다. 따라서 물가지수는 지수의 가장 대표적인 것이 되며 이론적으로도 깊이 연구되고 있으므로 이에 관하여 조사하여 보기로 한다.

6.1 物価指數

6.1.1 物価指數의 分類

물가지수란 시장에서 거래되는 상품이나 용역의 가격변동을 표시하는 것으로서 여러가지 종류가 있으나 도매물가지수, 소비자물가지수 혹은 생계비지수 및 소매물가지수로 나눌 수 있다.

도매물가지수란 도매거래에서의 물가수준의 변동을 측정하는 지수를 말하며, 경제활동의 변동을 민감하게 반영한다고 볼 수 있다.

이에 대하여 소비자물가지수는 소비자생활에서의 물가수준의 변동을 나타내는 지수이며, 화폐가치의 변동측정이란 의미에서 화폐구매력의 변동을 나타내는 소비자물가지수가 보다 중요한 것이라 할 수 있다. 또 소매물가지수란 소매시장에서 거래되는 상품가격의 종합지수를 말하며 소비자물가지수와 다른 것은 써비스가격을 제외하고 있다는 점이다.

6.1.2 物価指數의 算式

(가) 총합지수계산

물가지수는 시장에서 거래되는 상품 및 써비스의 물가변동을 종합적으로 나타내는 것이므로 상품과 써비스의 가격계열을 사용하여 계산한다.

총합지수의 계산은 평균과 비율의 두가지 계산과정으로 처리되는데 그 계산의 순서에 따라 물가지수의 계산방법은 두가지로 분류된다. 평균을 먼저 계산한 다음 비율을 계산하는 것을 총계법 또는 총합법이라고 하고, 비율을 먼저 계산한 다음 평균을 계산하는 것을 비율평균법이라고 한다.

총계법은 상품 및 써비스의 가격을 합하여 비교하는 산식이다. 각 상품의 기준시 가격을 P_0, 비교시 가격을 P_t, 상품품목수를 N이라고 하면 총계법은 다음 식과 같다.

$$\text{총계지수} = \frac{\Sigma P_t / N}{\Sigma P_0 / N} \times 100 = \frac{\Sigma P_t}{\Sigma P_0} \times 100 \qquad (6.1)$$

이 식은 결국 상품가격의 합계를 비교하는 것이어서 이러한 계산법을 총계법이라고 한다. 100을 곱한 것은 기준시의 지수를 100으로 나타낼 경우이므로 모든 지수 산식에서 공통승수이다.

비율 평균법은 개개의 상품에 대하여 개별가격지수를 계산하고 이 개별지수를 전 상품에 대하여 평균하는 방법이다. 사용되는 평균은 산술평균이 일반적이지만 기하평균을 권장했던 시기도 있었으

며, 계산식은 다음과 같다.

$$\text{산술평균지수} \quad \frac{1}{N}\Sigma\frac{Pt}{Po} \tag{6.2}$$

$$\text{기하평균지수} \quad n\sqrt{\pi\frac{Pt}{Po}} \tag{6.3}$$

지수를 계산하는 기본적인 계산식은 위의 세가지이지만 계산식에 따라 값이 달라진다. 총계법은 계산식에서 알 수 있는 바와 같이 상품단위의 수에 의해 값이 달라지는 불완전한 결점이 있는데 반하여 비율법은 그러한 불완전성은 없지만 어떤 평균치를 사용하느냐에 따라 값이 달라지며 평균의 성질에 의하여 산술평균으로 계산된 값이 기하평균에 의한 것보다 크게 된다.

<예제 6.1> 다음 자료에서 a) 종합지수식과 b) 비율지수식을 사용하여 물가지수를 계산해 보아라. (단위 : 10원)

품목 \ 년도	1975	1976
A	203	245
B	87	103
C	157	223

<풀 이>

a) $\frac{\Sigma Pt}{\Sigma Po} \times 100 = \frac{245 + 103 + 223}{203 + 87 + 157} \times 100 = 127.74$

b) i) 산술평균식

$$\frac{1}{N}\Sigma\frac{Pt}{Po} \times 100 = \frac{1}{3}\left(\frac{245}{203} + \frac{103}{87} + \frac{223}{157}\right) \times 100 = 127.04$$

ii) 기하평균식

$$\sqrt[n]{\pi\frac{Pt}{Po}} \times 100 = \sqrt[3]{\frac{245}{203} \times \frac{103}{87} \times \frac{223}{157}} \times 100 = 126.7$$

<예제 6.2> 1971년부터 1975년까지 가격의 연환비가 125, 118, 139, 145, 165이었다. a) 1970년 기준의 1972년 상대가격을 계산하라. b) 1971년 기준에 대한 연쇄지수를 구하라.

<풀 이> 각 연도의 비율을 보면 각각 $P_{70/71} = 1.25$, $P_{71/72} = 1.18$, $P_{72/73} = 1.39$, $P_{73/74} = 1.45$, $P_{74/75} = 1.65$이다.

a) $P_{70/72} = P_{70/71} \times P_{71/72} = 1.25 \times 1.18 = 1.475 = 147.5\%$

b) $P_{71/70} = 1/P_{70/71} = 1/1.25 = 80\%$

$P_{71/71} = 100\%$

$P_{71/72} = 118\%$

$P_{71/73} = P_{71/72} \times P_{72/73} = 1.18 \times 1.39 = 1.64 = 164\%$

$P_{71/74} = P_{71/72} \times P_{72/73} \times P_{73/74} = 1.18 \times 1.39 \times 1.45 = 237.8\%$

$P_{71/75} = P_{71/72} \times P_{72/73} \times P_{73/74} \times P_{74/75}$

$$= 1.18 \times 1.39 \times 1.45 \times 1.65 = 329.4\ \%$$

(나) 가중지수

거래되는 상품이나 써비스의 가격계열을 종합적으로 지수를 작성하는 것이지만 소비생활상 그 상품들의 중요도가 각각 다르다.

이러한 중요도에 관계없이 계산된 지수를 단순지수라고 하고 이에 반하여 품목의 중요도인 가중치를 반영한 지수를 가중지수라고 한다.

단순지수라 할지라도 품목의 중요성이 전혀 무시된 것은 아니다. 단순한 총계지수식은 상품의 가격지수를 기준시단위가격을 가중치로 가중시킨 가중산술평균으로 볼 수 있고 단순한 비율법의 지수에서도 품목수 또는 종류의 선택에서 가중치가 고려된 것이라고 할 수 있다. 이런 점에서 종합지수는 모두 어떤 의미의 가중지수라고 할 수 있으나 그 가중치는 보다 합리적으로 처리되어야 할 것이다. 단순총계지수의 경우 기준시의 단위가격이 반드시 상품의 중요도에 비례하는 것은 아니다.

각 품목의 가중치를 W라고 하면 총계지수와 비율평균지수에 대응하는 가중지수산식은 아래와 같다.

$$\text{가중 총계지수} \quad Pot = \frac{\Sigma W P_t}{\Sigma W P_o} \tag{6.4}$$

$$\text{가중산술평균지수} \quad Pot = \frac{\Sigma \left(W \cdot \frac{P_t}{P_o} \right)}{\Sigma W} \tag{6.5}$$

가중기하평균지수 $P_{ot} = \Sigma W \sqrt{\pi \left(\frac{P_t}{P_o}\right)^W}$ (6.6)

평균지수에서는 상품의 거래금액 또는 소비금액을 가중치로 사용하고 총계지수에서는 거래수량 또는 소비수량을 가중치로 사용한다.

이 수량을 q, 단위가격을 P라고 하면 거래 또는 소비금액은 pq로 나타낼 수 있고, 수량 역시 가격의 시점표시와 같이 기준시, 비교시의 수량을 각각 q_o, q_t라고 하면 기준시의 금액은 p_oq_o 비교시의 금액은 p_tq_t로 계산된다. (6.4)식의 가중치에 q_o 또는 q_t를 대입하면 다음의 2식으로 변형된다.

기준시수량가중총계식(라스파이레스산식) $P_{ot} = \frac{\Sigma p_t q_o}{\Sigma P_o q_o}$ (6.7)

비교시수량가중총계식(파아쉐산식) $P_{ot} = \frac{\Sigma p_t q_t}{\Sigma p_o q_t}$ (6.8)

(6.8)식을 변형하면

$$P_{ot} = \frac{\Sigma p_t q_t}{\Sigma \frac{p_o}{p_t}(p_t q_t)} \qquad (6.8')$$

이 된다. 또 (6.5)식의 가중치에 금액의 가중치 p_oq_o, p_tq_t를 사용하면

기준시 금액가중산술평균지수

$$P_{ot} = \frac{\Sigma p_o q_o \left(\frac{p_t}{p_o}\right)}{\Sigma p_o q_o} = \frac{\Sigma p_t q_o}{\Sigma p_o q_o} \qquad (6.9)$$

비교시 금액가중산술평균지수

$$P_{ot} = \frac{\Sigma\, p_t q_t \left(\frac{p_t}{p_o}\right)}{\Sigma\, p_t q_t} \qquad (6.10)$$

(6.7)식과 (6.9)식은 같은 결과가 된다. 이것은 기준시 수량을 가중시킨 총계지수와 기준시의 금액을 가중시킨 산술평균지수의 값은 같게 된다는 것이다. (6.7)식을 라스파이레스 산식 (6.8)식을 파아쉐 산식이라고 한다. 이 두식은 물가지수에서 이론적으로나 실제적으로 중요한 의미를 가지고 있으며 현재 대부분의 국가는 라스파이레스 산식에 따라 물가지수를 계산하고 있다.

그 이유는 조사작성상의 편의에 있다. 즉 그것은 $p_o q_o$를 거래금액으로서 직접 조사하여 (6.9)식을 이용하면 되기 때문이다.

또한 (6.8)식은 (6.8′)식으로 변형하여 계산하면 보다 쉽게 처리된다. 그러나 파아쉐 산식은 비교시 금액 $p_t q_t$를 시점의 진행과 함께 부단히 새로이 조사해 가지 않으면 안되기 때문에 물가지수 그 자체로서는 실용화되지 않고 있으며 다만 라스파이레스지수의 체크재료로서 연구적으로 이용되고 있다.

일반적으로 가중물가지수는 기준시의 자료를 가중치로 사용하느냐 비교시의 자료를 가중치로 사용하느냐에 따라 그 결과가 다르게 된다. 그래서 형식적으로 양경우의 결과를 같게 평가하여 그 평균을 이용하는 방법도 있다. 즉 라스파이레스산식과 파아쉐산식의 평균을 생각할 때 산술평균과 기하평균을 사용하면 다음 식이 얻어

진다.

수량가중총계산술평균지수

$$P_{ot} = \frac{1}{2}\left(\frac{\Sigma p_t q_o}{\Sigma p_o q_o} + \frac{\Sigma p_t q_t}{\Sigma p_o q_t}\right) \qquad (6.11)$$

수량가중총계기하평균지수

$$P_{ot} = \sqrt{\frac{\Sigma p_t q_o}{\Sigma p_o q_o} + \frac{\Sigma p_t q_t}{\Sigma p_o q_t}} \qquad (6.12)$$

(6.12)식을 핏셔산식 또는 핏셔의 이상산식이라고 한다. 위에서와 같이 가중산식을 평균하는 것과는 달리 기준시와 비교시의 가중치만을 평균하여 사용하는 방법도 있다. 총계지수에서 가중치의 평균을 사용하면 다음 식과 같이 된다.

기준비교양시점수량가중평균지수

$$P_{ot} = \frac{\Sigma p_t(q_o + q_t)}{\Sigma p_o(q_o + q_t)} \qquad (6.13)$$

이 식을 에지워드 보울레이 산식 또는 마샬 에지워드 산식이라고 한다. 물가지수에 관한 산식에는 이밖에도 여러가지가 있으나 지나치게 이론적이라고 할 수 있다.

지금까지 고찰한 물가지수를 원자론적 접근법이라고 하고, 경제이론의 한계치이론, 근사치이론, 탄력성이론 등으로서 물가지수를 측정하려는 함수론적 접근법으로 구별하는 이론도 있다.

<예제 7.3> 다음 자료에서 가중지수를 계산해 보자.

가격수량 / 구분	1971 년		1975 년	
	가 격	수 량	가 격	수 량
A	70	12	180	14
B	50	9	60	10
C	40	7	90	8

<풀 이>

a) $$\frac{\Sigma p_t q_o}{\Sigma p_o q_o} = \frac{180 \times 12 + 60 \times 9 + 90 \times 7}{70 \times 12 + 50 \times 9 + 40 \times 7} = 212.1\,\%$$

b) $$\frac{\Sigma p_t q_t}{\Sigma p_o q_t} = \frac{180 \times 14 + 60 \times 10 + 90 \times 8}{70 \times 14 + 50 \times 10 + 40 \times 8} = 213.3\,\%$$

c) $$\frac{\Sigma p_o q_o (\frac{p_t}{p_o})}{\Sigma p_o q_o} = \frac{70 \times 12 (\frac{180}{70}) + 50 \times 9 (\frac{60}{50}) + 40 \times 7 (\frac{90}{40})}{70 \times 12 + 50 \times 9 + 40 \times 7}$$

$$= 212.1\,\%$$

d) $$\frac{\Sigma p_t q_t (\frac{p_t}{p_o})}{\Sigma p_t q_t} = \frac{180 \times 14 (\frac{180}{70}) + 60 \times 10 (\frac{60}{50}) + 90 \times 8 (\frac{90}{40})}{180 \times 14 + 60 \times 10 + 90 \times 8}$$

$$= 229.7\,\%$$

e) $\frac{1}{2}\left(\frac{\Sigma p_t q_t}{\Sigma p_o q_o} + \frac{\Sigma p_t q_t}{\Sigma p_o q_t}\right) = \frac{1}{2}(212.1 + 213.3) = 212.7\,\%$

f) $\sqrt{\frac{\Sigma p_t q_t}{\Sigma p_o q_o} \times \frac{\Sigma p_t q_t}{\Sigma p_o q_t}} = \sqrt{212.1 \times 213.3} = 212.7\,\%$

g) $\frac{\Sigma p_t (q_o + q_t)}{\Sigma p_o (q_o + q_t)} = \frac{180\times(12+14)+60\times(9+10)+90(7+8)}{70\times(12+14)+50\times(9+10)+40\times(7+8)}$

$= 212.8\,\%$

6.1.3 라스파이레스 산식과 파아쉐 산식의 검토

물가변동은 경제생활의 여러 분야에 영향을 미치므로 물가변동을 어떻게 정확히 측정하여 표시하는 가는 계량경제학의 중요한 과제중의 하나로 되어있다. 가장 기본이 되는 소비생활과 물가수준관계를 볼 때 물가지수 그 자체는 소비생활에서 소비자가 동일한 크기의 만족을 얻기 위하여 지불하여야만 하는 화폐구매력의 변동을 측정하는 것과 마찬가지이다. 소비자가 일정한 크기의 효용 u를 얻기 위하여 지불하는 기준시와 비교시의 화폐액을 각각 $Mo(u)$, $Mt(u)$라고 하면 물가의 변동 $Pot(u)$는 는 다음과 같이 표시된다.

$$Pot(u) = \frac{Mt(u)}{Mo(u)} \qquad (6.14)$$

시점 0에서 일정크기의 효용 u_o에 대한 소비자의 화폐지출액은

$$M_o(u_o) = \Sigma P_o q_o$$

이다. 소비관습에 변동이 없다면 소비자는 비교시에 있어서도 기준시와 같은 수량 q_o으로 같은 만족을 누릴 것이므로 비교시에도 $\Sigma P_t q_o$만큼 지불하여 같은 물건을 같은 양만큼 사서 같은 효용을 얻을 것이다. 그러나 그동안 상품의 변동이나 가격변동이 있을 것이고, 소비자는 상품의 종류와 수량의 선택을 바꾸어 보다 적은 지출로 전과 같은 효용을 충족시키려고 할 것이므로 이 때에 실제로 지불하는 금액 $M_t(u_o)$는 $\Sigma p_t q_o$보다 적어진다. 그러므로 $M_t(u_o) < \Sigma p_t q_o$의 관계에서 다음 부등식이 성립한다.

$$P_{ot}(u_o) = \frac{M_t(u_o)}{M_o(u_o)} < \frac{\Sigma p_t q_o}{\Sigma p_o q_o} \qquad (6.15)$$

이 식의 우변은 라스파이레스 산식이며, 그 식은 실제보다 크게 나타내는 결과가 된다.

비교시의 소비자의 화폐지출은

$$M_t(u_t) = \Sigma p_t q_t$$

이고, 같은 효용을 기준시에도 누리기 위하여 지불하는 금액 $M_o(u_t)$에 관하여 $M_o(u_t) < \Sigma p_o q_t$가 되므로 파아쉐 산식에 관하여는 다음 관계가 성립한다.

$$P_{ot}(u_t) = \frac{M_t(u_t)}{M_o(u_t)} > \frac{\Sigma p_t q_t}{\Sigma p_o q_t} \qquad (6.16)$$

이는 파아쉐 산식이 실제보다 적게 나타낸다는 것을 의미한다.

만약 기준시의 생활수준 u_o 와 비교시의 생활수준 ut에 큰 변동이 없다고 하면 다음 관계가 성립되며

$$P_{o_t}(u_o) = P_{o_t}(ut)$$

또한 (6.15)식과 (6.16)식에 의하여

$$\frac{\Sigma p_t q_o}{\Sigma p_o q_o} > \frac{\Sigma p_t q_t}{\Sigma p_o q_t}$$

의 관계가 근사적으로 성립된다. 물가지수의 라스파이레스 산식은 파아쉐 산식보다 일반적으로 값이 크다는 것이며 이는 앞에서 언급한 바 있는 파아쉐 산식이 라스파이레스 산식의 책크재료가 된다는 근거가 되기도 한다. 즉 파아쉐 산식은 진정한 물가지수의 하한계를 표시하고 있어서 (6.15)식에 보인 바와 같이 진정한 물가수준의 상한계인 라스파이레스 산식과 비교해서 양자의 차가 현저하게 되면 라스파이레스 산식 그 자체의 기준시라든가 가중치의 개정이 필요하다는 것을 표시할 수 있다.

물론 실제의 지수에서는 소비자의 합리적 행동이 반드시 예상되지 않으므로 파아쉐지수 편이 도리어 라스파이레스지수보다 큰 값을 표시하는 수도 가끔 있지만, 이 경우도 포함해서 양자의 차의 정도 여하에 의해 라스파이레스지수 설계의 타당성을 검토하는 방식을 파아쉐 책크라고 한다.

6.2 數量指數

물가에 관한 지수의 계산으로 물가의 변동을 파악하는 것처럼 생산수량, 거래수량 등에 관한 지수를 물가지수와 같은 방법으로 측정하여 그 변동을 파악하는 지표로 한다. 그러나 상품은 개수, 길이, 중량, 용적 등 단위가 다양하기 때문에 총계법으로 쉽게 수량계열을 총합하지 못한다.

비율평균법에 의한 단순총합수량지수는 다음과 같다.

$$\text{단순수량지수} \quad \frac{1}{N}\Sigma\frac{q_t}{q_o} \tag{6.17}$$

가중수량지수는 가중물가지수 계산식에서 가격요소 p와 수량요소 q를 바꾸어 사용하면 된다. 라스파이레스 산식, 파아쉐 산식 및 핏셔의 이상 산식에 의한 가중수량지수식은 그 식의 p와 q대신 q와 p를 대입하면 다음과 같이 된다.

$$\text{기준시 가격가중수량지수} \quad Q_{ot}=\frac{\Sigma q_t p_o}{\Sigma q_o p_o} \tag{6.18}$$

$$\text{비교시 가격가중수량지수} \quad Q_{ot}=\frac{\Sigma q_t p_t}{\Sigma q_o p_t} \tag{6.19}$$

$$\text{가격가중총계기하평균수량지수} \quad Q_{ot}=\sqrt{\frac{\Sigma q_t p_o}{\Sigma q_o p_o}\times\frac{\Sigma q_t p_t}{\Sigma q_o p_t}} \tag{6.20}$$

6.3 賃金指數

임금에 대한 지수는 화폐단위 그대로의 명목임금의 변동을 측정하는 명목임금지수와 명목임금에 소비자 물가를 감안한 실질임금 지수의 두가지로 계산될 수 있다.

기준시의 임금 총액을 W_o, 비교시의 임금 총액을 W_t라고 하고, 이에 대응하는 노동자 수를 각각 ℓ_0, ℓ_t라고 하면 각 시점의 평균임금은

$$W_o / \ell_o, \quad W_t / \ell_t$$

이고, 명목임금지수 N_{ot}와 실질임금지수 R_{ot}는 다음과 같이 된다.

$$N_{ot} = \frac{W_t / \ell_t}{W_o / \ell_o}, \quad R_{ot} = \frac{\frac{W_t}{\ell_t} / \frac{W_o}{\ell_o}}{P_{ot}} \qquad (6.21)$$

위의 지수들은 노동자 구성의 변동은 반영하고 있지 않으므로 기준시에 비하여 비교시에 노동자 구성비에 있어서 남자의 높은 연령층 비율이 높아졌다면 남자의 평균임금이 여자의 그것보다 높아지게 될 것이다. 이와 같이 평균임금의 상승은 반드시 임금수준의 상승을 나타내는 것이라고는 볼 수 없으므로 임금지수를 근로자의 성별, 연령별, 학력별, 직종별 등으로 각각 측정하는 것이 보다 구체적인 지수가 될 것이다.

기준시 및 비교시의 명목임금을 각각 N_o, N_t, 소비자 물가를 P_o, P_t라고 하면 대응시점의 실질임금 R_o, R_t는 다음과 같다.

$$R_o = \frac{N_o}{P_o}, \quad R_t = \frac{N_t}{P_t}$$

급격한 물가상승은 실질임금의 저하를 가져오므로 이에 대한 생계비의 앙등폭을 임금에 반영시키기 위한 것이 활척임금이다. 활척임금은 기준시와 비교시의 실질임금수준을 유지시키는 제도이므로 $Ro = Rt$가 되도록 비교시의 임금의 Nt를 결정하는 문제이다. 즉

$$\frac{No}{Po} = \frac{N_t}{P_t}$$

$$Nt = \frac{No \cdot P_t}{Po} = No \cdot Po_t \qquad (6.22)$$

이므로 비교시의 명목임금은 기준시의 명목임금에 기준시에 대한 비교시의 소비자물가지수를 곱해 주어야 한다.

△ 연습문제 △

(6.1) 어떤 상품의 1973년 평균가격이 1972년에 비하여 20% 올랐고, 1971년에 비하여 20% 내렸으며, 1974년에 비하여 50% 상승한 값이다. 1) 1971년 기준, 2) 1972년 기준, 3) 1974년 기준의 지수를 각각 구하라.

(6.2) 다음은 서울의 소매물가자료이다.

가격수량 / 품목	1970년		1975년	
	가 격	수 량	가 격	수 량
쌀	632	13	2,026	36
보 리 쌀	371	21	871	24
밀 가 루	773	19	2,440	22

1) 단순총계지수(수량무시)

2) 단순비율평균지수

3) 라스파이레스지수

4) 파아쉐지수

5) 핏셔지수

6) 라스파이레스 및 파아쉐수량지수를 각각 구하라.

(6.3) 다음은 우리나라 제조업 종업원의 평균임금 수준이다.

연 도	1970	1971	1972	1973	1974	1975
임 금	1430	1661	1892	2233	3021	3838
물 가	100.0	113.5	126.8	130.8	162.6	203.7

1) 연쇄지수

2) 1971년 기준에 대한 1975년의 임금지수

3) 노동자수가 1971년 1,284, 1975년 2,205 (단위 : 생략)일 때 그 명목임금지수

4) 2)의 임금지수에서 실질임금지수를 구하라.

Ⅶ章 : 回 帰 分 析

7.1 序 論

어떤 학문이든지 간에 그 연구활동에 있어서 관련된 변수들 간에 상호관련성을 찾으려고 할 때가 많다. 예를 들면 사람이 나이가 먹어갈수록 키가 커지는데 나이와 키 사이의 함수관계를 생각해 볼 수 있을 것이다. 다른 예로서 어떤 플라스틱제품의 견고도가 이 제품을 만드는 기계의 온도와 어떤 연관성을 가지고 있다고 하면 이들 변수간의 함수관계를 규명하는 것은 흥미있는 연구일 것이다.

위의 두가지 예에서 키와 견고도는 일반적으로 종속변수라 부르고 나이와 온도와 같은 종속변수에 영향을 주는 변수를 독립변수라고 한다.

우리 주위에는 독립변수와 종속변수사이의 함수관계를 알아내려고 하는 수없이 많은 문제들이 있으며 통계적 방법에서 널리 애용되고 있는 회귀분석은 이러한 함수관계를 알아내는데 매우 유익하게 쓰여지고 있다. 일반적으로 회귀분석은 한개의 독립변수와 한개의 종속변수 간의 관계분석에 제한되어 있지 않고 여러개의 변수들간의 함수관계를 규명하는데도 많이 쓰여지고 있으나 이 장에서는 제일 간단한 경우에 해당하는 한개의 독립 및 종속변수간의 선형 함수관계에 관하여 생각하기로 한다.

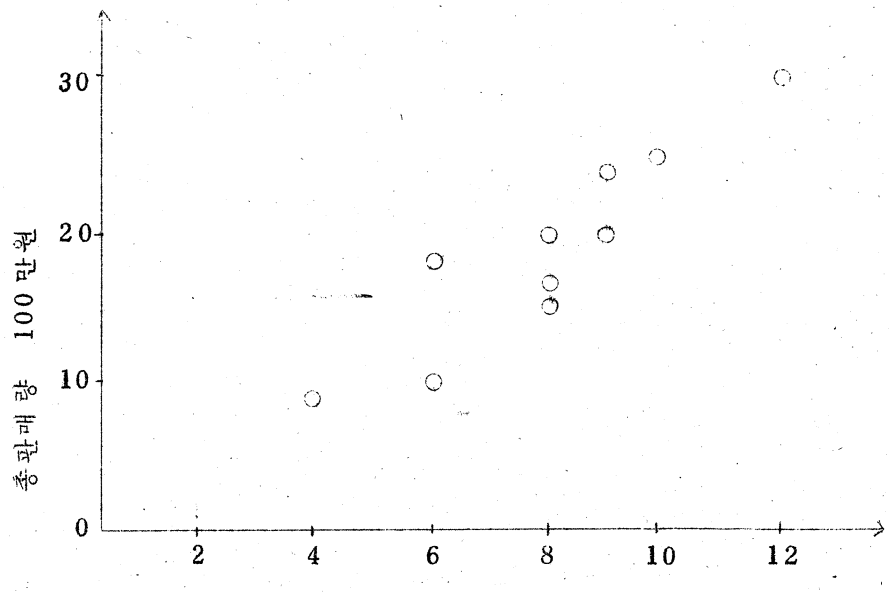
7.2 回帰分析의 基本概念

7.2.1 散布図

두 변수간의 함수관계를 연구하는 첫단계로서 먼저 하여야 할 것은 도표상에 관찰점을 그려보는 것이다. 이러한 도표를 산포도라고 하는데 많은 경우에 이 산포도로부터 두 변수간의 관계를 대략적으로 짐작할 수 있다.

<표 7.1> 표본의 광고료와 총판매액

상 점 번 호	광 고 료 (단위 : 10만원)	총 판 매 액 (단위 : 100만원)
1	4	9
2	8	20
3	9	22
4	8	15
5	8	17
6	12	30
7	6	18
8	10	25
9	6	10
10	9	20



<그림 7.1>　　광고료와 총판매량의 산포도

구체적인 예를 들어 보기로 하자. 어떤 특수한 종류의 상품을 팔고 있는 상점을 중심으로 광고가 판매량에 미치는 관계를 알아보기 위하여 10 개의 상점을 단순확률표본으로 추출하여 선택된 상점들의 년간 광고료와 총판매액을 알아보고 그 자료가 <표 7.1> 에 있는 바와 같다고 하자. 여기서는 광고료는 독립변수(앞으로 x 라고 표시)이고 판매액은 x의 증감에 따라 영향을 받으므로 종속변수(앞으로 y로 표시한다.)라고 한다.

<표 7.1>의 자료를 산포도로 그려보면 <그림 7.1>과 같다. <그림 7.1>로부터 우리는 x가 증가하면 y도 증가한다는 사실을 짐작할 수 있고 그 관계가 대략적으로 직선적인 것도 알 수

있다.

7.2.2 基本假定

한개의 독립변수를 가진 선형회귀모형을 적합시킬 때는 일반적으로 다음의 가정이 바탕을 이루고 있다.

1. 변수 x와 y사이에 존재하는 관련성은 선형함수 관계로 적절히 표현될 수 있다.

2. 주어진 x의 값에서 y의 관측치의 표준편차는 x의 값에 관계없이 일정하다.

3. 변수 x는 오차없이 관측할 수 있는 수학변수이며 확률변수는 아니다.

위의 가정들은 수식으로 표현하면 다음과 같다.

$$y = \alpha + \beta x + \varepsilon \tag{7.1}$$

여기서 α와 β는 母數이고 ε는 관측오차이다.

$$E(\varepsilon) = 0 \tag{7.2}$$

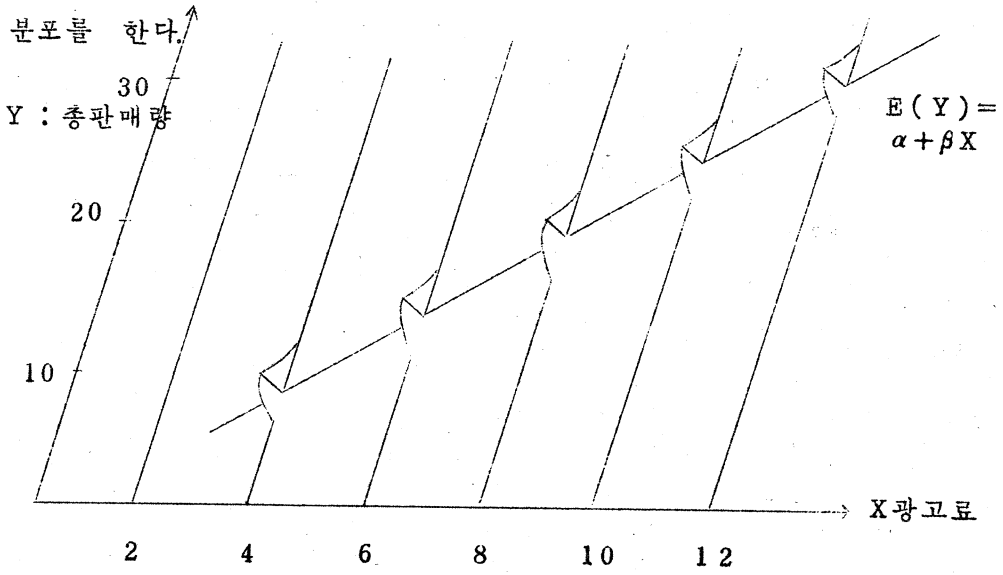
$$Var(\varepsilon) = \sigma^2$$

$$E(y) = \alpha + \beta x \tag{7.3}$$

$$Var(y) = \sigma^2$$

이에 첨가해서 y는 정규분포를 이룬다고 일반적으로 가정되는데 이러한 성질을 알기 쉽게 <표 7.1>의 광고료와 총판매액에 대한 x와 y의 좌표그림으로 그려보면 <그림 7.2>를 얻을 수

있다. 예를 들어 $x=6$에서의 y의 평균치는 $E(y)=\alpha+6\beta$이며 이는 <그림 7.2>에서 $x=6$과 $E(y)=\alpha+\beta x$가 만나는 점이며 이를 중심으로 y의 관측치가 표준편차 σ를 가지는 정규분포를 한다.



<그림 7.2> y의 확률분포

7.2.3 回帰分析의 目的

1. 정해진 어떤 독립변수의 값에 대하여 종속변수가 평균하여 어떤 추정치를 갖게 될 것인가에 대한 정보를 제공해 준다.

2. 회귀선을 사용하여 종속변수의 값을 추정할 때 이 추정에 관련된 오차의 크기에 대한 정보를 제공해 준다.

3. 두 변수사이에 존재하는 관련도를 측정할 수 있는 정보를

제공해 준다.

4. 여러가지 가설검정을 할 수 있다.

7.3 回帰線의 推定과 精度

7.3.1 回帰線의 推定

회귀분석의 첫번째 목적을 달성하기 위하여 x와 y의 관계를 설명하는 회귀선을 구해야 한다. 이 장에서는 선형회귀분석만을 다루므로 회귀선을 나타내는 방정식은 직선이 되며 어떤 주어진 y의 값에 대응하는 x에서의 y의 추정치는

$$\hat{y} = \alpha + \hat{\beta}x \qquad (7.4)$$

로 표시된다. 여기서 $\hat{\alpha}$는 식 (7.3)에서 α의 추정치이고 $\hat{\beta}$는 β의 추정치이다. α는 y좌표의 차단점이라고 불려지기도 하며 이는 $x=0$에서의 $\hat{y}$의 값이 된다. $\hat{\beta}$는 이 직선의 기울기이며 이 $\hat{\beta}$의 값은 x가 한 단위 증가할 때의 y의 증가량을 나타내 준다.

이제 $\hat{\alpha}$와 $\hat{\beta}$를 구하는 방법을 강구해 보자. 임의표본추출에 의하여 n개의 관찰점 (x_1, y_1), (x_2, y_2), ……(x_n, y_n)을 가지고 있다고 하자. 이 점들을 가장 잘 적합시킬 수 있는 직선 (7.4)을 구하기 위하여서는 적합도를 어떻게 측정할 것인가에 대한 표준을 정해야 할 것이다.

여러가지 표준들이 고려될 수 있으나 회귀분석방법에서 가장 널

리 적용되고 있는 최소자승법을 사용하기로 하자.

이 방법은 관찰점 y와 직선으로부터 얻어지는 $\hat{y}$의 편차의 자승을 모두 합친 것이 최소가 되어야 한다는데 원칙을 두고 있다.

다시 말해서

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (7.5)$$

이 최소치를 갖게 하는 (7.4)의 식이 최소자승법에 의하여 구하여지는 직선이다. 여기에서

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i \qquad (7.6)$$

이다.

이 방법의 특성은 n개의 관찰점으로부터 x와 y의 평균치를 $\bar{x}$, $\bar{y}$라고 하면 점 ($\bar{x}$, $\bar{y}$)는 최소자승법에 의하여 구하여지는 직선상에 위치하게 되고 y와 $\hat{y}$의 편차의 합이 0가 된다는 것이다. 즉

$$\sum_{i=1}^{n}(y_i - \hat{y})^2 = 0 \qquad (7.7)$$

이제 식 (7.5)를 최소화 시키는 식 (7.6)의 $\hat{\alpha}$와 $\hat{\beta}$를 구하는 방법을 생각해 보자. 우리가 최소화시키려는 것은

$$S = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \qquad (7.8)$$

이며 S를 $\hat{\alpha}$와 $\hat{\beta}$로 각각 편미분하여 0으로 놓으면 식(7.9)를 얻게 된다.

$$\frac{\partial S}{\partial \hat{\alpha}} = -2 \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$\frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \qquad (7.9)$$

식 (7.9)를 정리하면

$$\hat{\alpha} n + \hat{\beta} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\alpha} \sum_{i=1}^{n} x_i + \hat{\beta} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad (7.10)$$

이 되는데 이 식을 正規方程式이라고 부른다.

식 (7.10)을 $\hat{\alpha}$와 $\hat{\beta}$에 대하여 풀면

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \qquad (7.11)$$

이 된다.

앞에서 산포도를 그릴 때 사용된 예제에 대하여 최소자승법에 의하여 회귀선을 구하여 보자. <표 7.1>에 있는 자료로 부터 다음을 얻을 수 있다.

앞으로 Σ는 $\sum_{i=1}^{n}$을 의미한다.

$n = 10$

$\Sigma y_i = 9 + 20 + \cdots\cdots + 20 = 186$

$\bar{y} = 186/10 = 18.6$

$\Sigma x_i = 4 + 8 + \cdots\cdots 9 = 80$

$\bar{x} = 80/10 = 8$

$\Sigma x_i y_i = (4)(9) + (8)(20) + \cdots\cdots + (9)(20) = 1608$

$\Sigma x_i^2 = 4^2 + 8^2 + \cdots\cdots + 9^2 = 686$

$$\hat{\beta} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n}$$

$$= \frac{1608 - (80)(186)/10}{686 - (80)^2/10} = 2.609$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 18.6 - (2.609)(8) = -2.270$$

그러므로 적합된 회귀선은

$$\hat{y} = -2.270 + 2.609x \qquad (7.12)$$

이며 관찰된 x의 값에서 y의 평균치 $E(y)$에 대한 추정치인 $\hat{y}$를 구해보면 <표 7.2>를 얻을 수 있다. 잔차(또는 편차)인 $(y_i - \hat{y}_i)$도 <표 7.2>에 나타나 있다. 잔차의 합을 구해 보면 0이 되는데 이는 식 (7.7)에서 얘기한 최소자승법의 하나의 성질이다.

<표 7.2> $\hat{y}_i$와 $y_i - \hat{y}_i$

관찰점 번호	x_i	y_i	$\hat{y}_i = -2,270+2,609x_i$	$y_i-\hat{y}_i$
1	4	9	8.17	0.83
2	8	20	18.60	1.40
3	9	22	21.21	0.79
4	8	15	18.60	-3.60
5	8	17	18.60	-1.60
6	12	30	29.04	0.96
7	6	18	13.38	4.62
8	10	25	23.81	1.19
9	6	10	13.38	-3.38
10	9	20	21.21	-1.21

$$\Sigma(y_i - \hat{y}_i) = 0$$

7.3.2 回帰方程式의 精度

앞에서는 주어진 자료로부터 회귀선을 구하는 방법을 보았는데 회귀선만을 가지고는 관찰점들이 회귀선 주위에 어떻게 분포되어 있으며 회귀선이 이 점들을 어느 정도 잘 대변하고 있는가를 알 수 없다. 식 (7.11)을 이용하여 얻어지는 회귀방정식 (7.4)의 정도를 측정하는 여러가지 방법에 대하여 알아보기로

하자.

(가) 추정치의 표준오차

y의 관찰치의 값이 회귀선 상에 전부 있다고 한다면 이 회귀선을 사용하여 y의 추정치를 구했을 때 이 추정치에는 오차가 없을 것이다. 그러나 관찰점들이 회귀선으로부터 멀리 떨어져 있는 것이 많을 경우에는 회귀선으로부터 얻어진 y의 추정치와 실제 y의 관찰치와는 큰 차이가 있을 수 있다.

표본의 크기가 n인 표본으로부터 회귀선을 얻었을 때 n개의 관찰점들이 회귀선 주위에 어느정도 산재해 있는가를 알아내는 척도로서 널리 쓰이는 것이 추정치의 표준오차이고 이 오차는

$$S_{y\cdot x}=\sqrt{\frac{\Sigma\,(y_i-\hat{y}_i)^2}{n-2}} \qquad (7.13)$$

으로 구해진다.

<표 7.1>에 있는 표본자료에 대한 추정치의 표준오차를 구해보면 <표 7.2>를 이용하여

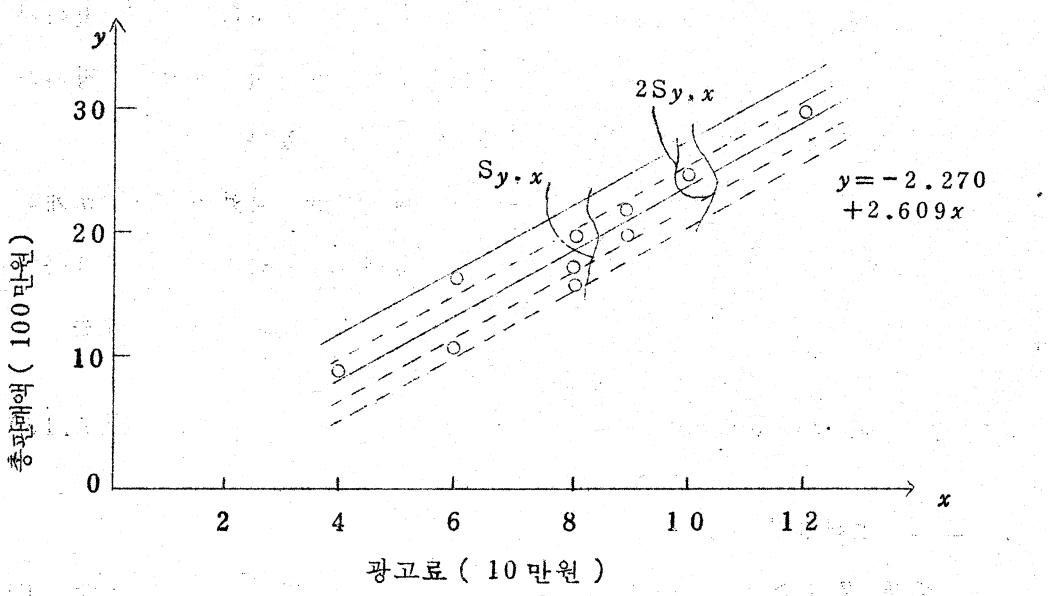
$$\Sigma(y_i-\hat{y}_i)^2=(0.83)^2+(1.40)^2+\cdots\cdots+(-1.21)^2=55.36$$

이므로

$$S_{y\cdot x}=\sqrt{\frac{\Sigma(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{55.36}{8}}=2.63(100\text{만원})$$

<그림 7.2>에서 보는 바와 같이 회귀선을 중심으로 y의 값들이 정규분포를 이룬다면 $S_{y\cdot x}$가 표준편차이므로 <그림 7.3>

에서 표시된 회귀선으로부터 거리가 2.63인 두개의 평행선 안에 약 60%의 y의 관찰치가 있게 되며 거리가 두개의 표준편차 ($2 \times 2.63 = 5.26$) 만큼 떨어져 있는 두개의 평행선 안에는 약 95.5%가 존재하게 된다.



<그림 7.3> 추정치의 표준편차

(나) 결정계수

앞 절에서 기술한 추정치의 표준편차는 회귀선으로부터 관찰점들이 어느 정도 멀리 떨어져 있는가에 대한 정보를 제공해 주며 두 변수 x와 y사이의 상관관계를 직접적으로 말해 주지는 않는다. 회귀분석의 목적중의 하나는 두 변수사이의 상관관계를 추정하는 것이며 하나의 측도로서 널리 쓰이는 결정계수에 대하여

논의하고자 한다.

회귀선을 중심으로 y값들의 변동은

$$\Sigma (y_i - \hat{y}_i)^2 \qquad (7.14)$$

에 의하여 표시하고 y값들이 그의 평균치를 중심으로 한 변동은

$$\Sigma (y_i - \bar{y})^2 \qquad (7.15)$$

로 나타내 진다. 결정계수는 이 변동들의 비례관계에서 얻어지며 보통 R^2으로 표시된다.

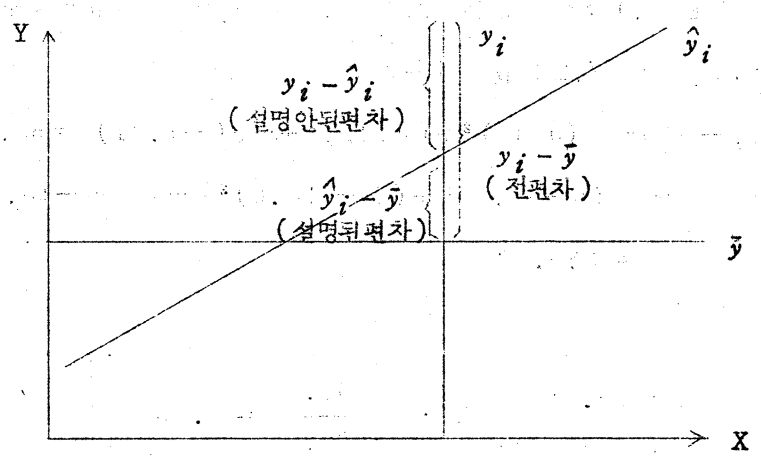
$$R^2 = 1 - \frac{\Sigma (y_i - \hat{y}_i)^2}{\Sigma (y_i - \bar{y})^2} \qquad (7.16)$$

R^2의 값을 해석하기 위하여 다음을 고려해 보자.

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \qquad (7.17)$$

전편차 = 설명안된편차 + 설명된 편차

이 관계는 <그림 7.4>를 보면 쉽게 이해할 수가 있다.



<그림 7.4> 전편차의 구분

위와같은 방법으로 전변동을 회귀선으로부터 설명안되는 변동과 설명되는 변동으로 구별할 수가 있다.

$$\Sigma (y_i - \bar{y})^2 = \Sigma (y_i - \hat{y}_i)^2 + \Sigma (\hat{y}_i - \bar{y})^2 \qquad (7.18)$$

(전변동) = (설명안된 변동) + (설명된 변동)

이제 다음의 비율을 고려해 보자.

$$\frac{\Sigma (y_i - \hat{y}_i)^2}{\Sigma (y_i - \bar{y})^2}$$

이 비율은 전변동 중에서 회귀방정식으로 설명이 안되는 변동에 대한 비율이며 따라서 식 (7.16)에 있는 R^2 은

$R^2 = 1 -$ (회귀방정식으로 설명이 안되는 변동의 비율)

= (회귀방정식으로 설명이 되는 변동의 비율)

이 된다. 따라서 회귀선이 100 %의 모든 변동을 설명할 수 있다면 $R^2 = 1$이 되며 만약 그 반대이면 $R^2 = 0$이 될 것이다.

예제로서 <표 7.1 >에 있는 표본자료에 대한 표본결정계수를 구하여 보자. <표 7.2 >를 이용하면

$$\Sigma (y_i - \hat{y}_i)^2 = (0.83)^2 + (1.40)^2 + \cdots + (-1.21)^2 = 55.36$$

$$\Sigma (y_i - \bar{y})^2 = (9 - 18.6)^2 + (20 - 18.6)^2 + \cdots + (20 - 18.6)^2$$

$$= 386.40$$

그러므로

$$R^2 = 1 - \frac{\Sigma (y_i - \hat{y}_i)^2}{\Sigma (y_i - \bar{y})^2} = 1 - \frac{55.36}{386.40} = 0.85$$

R^2의 값이 0.85라는 것은 전체변동 중에서 회귀선에 의하여 설명될 수 있는 부분이 85%라는 얘기이며 x와 y 사이의 상관관계가 높다는 말이 된다.

(대) 상관계수

상관계수는 단순히 R^2의 자승근이며 식 (7.4)에 있는 기울기 $\hat{\beta}$가 양이면 $R = \sqrt{R^2}$로 상관계수가 음의 값을 취할 수 없고 만약 기울기 $\hat{\beta}$가 음이면 $R = -\sqrt{R^2}$로 상관계수가 양의 값을 취할 수 없다. 한가지 특기할 것은 상관계수 R은 상관관계가 양인가 음인가를 말해 주지만 결정계수 R^2은 음의 값을 취할 수 없기 때문에 음의 상관관계를 R^2로 부터 탐지할 수 없다는 것이다. 이런 점에서 보면 상관계수의 용도가 좀 더 다양하다고 하겠다.

7.4 分散分析과 相関分析

7.4.1 分散分析

주어진 자료를 적합시키는데 있어서 직선회귀방정식이 유의한가 하는 것은 분산분석을 통하여 측정할 수 있다. 유의하다는 말은 주어진 자료를 설명하는데 있어서 회귀방정식이 의미있는 한몫을 한다는 뜻이며 어느 정도로 유의한가 하는 것은 분산분석표 <7.3>에서 마지막 난의 F의 값을 보면 알 수 있다. 회귀분석의 분산분석의 원천은 두가지이며 하나는 회귀이고 다른 것은 잔

차로 부터 오는 것이다.

<표 7.3> 분 산 분 석 표

원 천	자 승 합	자 유 도	자 승 평 균	F
회 귀	$\Sigma(\hat{y}_i-\bar{y})^2$	1	$\Sigma(\hat{y}_i-\bar{y}_i)^2$	$F=\frac{\Sigma(\hat{y}_i-\bar{y})^2}{\Sigma(y_i-\hat{y}_i)^2/n-2}$
잔 차	$\Sigma(y_i-\hat{y}_i)^2$	$n-2$	$\Sigma(y_i-\hat{y}_i)^2/n-2$	
전변동	$\Sigma(y_i-\bar{y})^2$	$n-1$	$\Sigma(y_i-\bar{y}_i)^2/n-1$	

분산분석에서 가장 중요한 것은 F의 값이며 이는 회귀의 자승평균과 잔차의 자승평균과의 비율이다. 이 비율이 크면 회귀의 자승합이 잔차의 그것보다 상대적으로 크다는 말이며 회귀방정식이 자료를 잘 설명해 주고 있다는 얘기이다. 이 F값이 크면 우리는 회귀방정식이 유의하다고 말한다. F의 값이 어느 정도 커야 되는 가에 대해서는 대부분의 통계학 책에 실려 있는 F-분포표를 찾아서 주어진 유의수준과 자유도($n-2$)를 이용하여 F의 기각치를 구하고 만약 분산분석표에서 구해진 F의 값이 F의 기각치보다 클 때에는 회귀방정식이 유의하고 그렇지 않으면 유의하지 않다.

<표 7.1>에 있는 예제에 대하여 분산분석표를 만들면 <표 7.4>와 같다. 자유도 (1, 8)과 유의수준 5%를 사용하여 F의 기각치를 구해보면 5.32인데 계산된 F의 값이 45.24이므로 우리가 사용해온 직선회귀방정식이 매우 유의함을 알 수 있다.

<표 7.4> 예제의 분산 분석표

원 천	자 승 합	자 유 도	자승평균	F
회 귀	313.04	1	313.04	45.24
잔 차	55.36	8	6.92	
전 변 동	368.40	9	40.93	

7.4.2 相関分析

지금까지 X와 Y의 함수관계가 식 (7.1)에 의하여 정의되고(즉 $y = \alpha + \beta x + \varepsilon$) 이 모형의 기본가정의 하나로서 Y는 확률변수이나 X는 수학변수로 가정한 후에 회귀분석의 여러 가지 측면을 생각해 왔다. 만약 X도 또한 확률변수이고 X와 Y가 어떤 二変量分布를 하고 있다면 X와 Y의 모집단의 상관계수는 다음에 의하여 정의한다.

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{7.19}$$

위에서 σ_X는 X의 표준편차로서 N이 모집단의 크기일 때 다음에 의하여 정의된다.

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}} \tag{7.20}$$

σ_Y는 Y의 표준편차로서 위 식에서 X를 Y로 대체하면 구해진다. 식 (7.19)에서 X와 Y의 共分散 Cov(X,Y)는 다음에 의하여 정의된다.

$$\mathrm{Cov}(X,Y)=\frac{1}{N}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) \qquad (7.21)$$

만약 X와 Y의 이변량분포가 연속적일 경우에는 식 (7.20)과 (7.21)에서 $\sum_{i=1}^{n}$ 표시 대신에 적분이 쓰여지게 될 것이다.

모집단으로부터 크기가 n인 확률표본을 뽑았을 때 n개의 자료점 $((x_1,,y_1),(x_2,y_2),\cdots\cdots,(x_n,y_n))$이 얻어졌다면 이 두 변수사이의 표본상관계수는

$$r=\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2}} \qquad (7.22)$$

으로 정의된다.

<표 7.1>에 있는 예제에 대하여 r의 값을 구해보자.

$$\Sigma(x_i-\bar{x})^2=(4-8)^2+(8-8)^2+\cdots+(9-8)^2=46$$

$$\Sigma(y_i-\bar{y})^2=(9-18.6)^2+(20-18.6)^2+\cdots(20-18.6)^2$$

$$=368.4$$

$$\Sigma(x_i-\bar{x})(y_i-\bar{y})=(4-8)(9-18.6)+(8-8)$$

$$(20-18.6)+\cdots+(9-8)(20-18.6)$$

$$=120$$

$$r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2}} = \frac{120}{\sqrt{(46)(368.4)}} = 0.92$$

표본상관계수 r의 제곱과 R^2은 일치한다. 또한 회귀방정식의 기울기 $\hat{\beta}$와 r과도 함수관계가 있다. 다음 식을 보자.

$$r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2}} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}$$

$$\cdot \frac{\sqrt{\Sigma(x_i-\bar{x})^2}}{\sqrt{\Sigma(y_i-\bar{y})^2}}$$

$$= \hat{\beta} \cdot \frac{S_x}{S_y} \qquad (7.23)$$

위에서 S_x는 표본치 x_i들의 표준편차이고 S_y는 y_i들의 표준편차이다.

$$S_x = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}$$

$$S_y = \sqrt{\frac{\Sigma(y_i-\bar{y})^2}{n-1}}$$

식 (7.23)에서 r과 $\hat{\beta}$가 선형의 관계가 있음을 알 수 있고 $\hat{\beta}$가 크면 r도 크고 $\hat{\beta}$가 음이면 r도 음이라는 것을 알 수 있다.

△ 연습문제 △

(7.1) 다음과 같은 $n=5$의 자료 $(xi, yi), i=1,2,3,4,5$가 있다.

x	1	2	3	4	5
y	2	3	5	7	8

(a) 산포도를 도표에 그려라.

(b) 직선을 적합시킬때 최소제곱추정값 $\hat{\alpha}$와 $\hat{\beta}$의 값을 구하여라.

(c) 추정된 직선을 산포도 위에 그려라.

(d) 잔차를 구하고 잔차의 합이 0이 됨을 보여라.

(e) 잔차의 제곱합을 구하라.

(f) 추정치의 표준오차 $Sy.x$를 구하라.

(g) 결정계수 R^2, 상관계수 R, 표본상관계수 r를 구하라.

(h) 직선회귀모형이 유의한가를 검정하고 분산분석표를 작성하라.