제5장 통계량과 그분포 237

5.1모집단과 표본 237

5.1.1 모집단과 개체 237

5.1.2 표본 239

5.2 표본자료의 정리와 표시 241

5.2.1 경험적분포함수 241

5.2.2 빈도수표와 잦음률표 242

5.2.3 표본자료의 도형표현 243

5.3 통계량과 그 분포 246

5.3.1 통계량과 표본화분포 246

5.3.2 표본평균과 그 표본분포 247

5.3.3 표본분산과 표준편차 250

5.3.4 표본모멘트와 그 함수 252

5.3.5 순서통계량과 그 분포 254

5.3.6 표본분위수와 표본중위수 258

5.3.7 5개수 요약과 상자선도 259

5.4 3대표본분포 265

5.4.1 분포 266



5.4.2 분포 268



5.4.3 분포 270



5.5 충분통계량 275

5.5.1 충분성의 개념 275

5.5.2 인자분해정리 278

제6장 파라메터추정 283

6.1 점추정의 개념과 불편성 283

6.1.1 점추정과 불편성 283

6.1.2 유효성 286

6.2 모멘트추정과 일치성 288

6.2.1 치환원리와 모멘트추정 288

6.2.2 확률함수가 기지일 때 미지파라메터의 모멘트추정 289

6.2.3 일치성 290

6.3 최대우도추정과 EM알고리듬 293

6.3.1 최대우도추정 293

6.3.2 EM알고리듬 296

6.3.3 점근정규성 298

6.4 최소분산불편추정 302

6.4.1 평균두제곱오차 303

6.4.2 평등최소분산불편추정 304

6.4.3 충분성준칙 306

6.4.4 크라메르-라오부등식 307

6.5 베이스추정 311

6.5.1 통계적추정의 기초 311

6.5.2 베이스공식의 밀도함수형식 312

6.5.3 베이스추정 313

6.5.4 공액사전분포 315

6.6 구간추정 317

6.6.1 구간추정의 개념 317

6.6.2 중심축량법 320

6.6.3 한개 정규모집단파라메터에 대한 믿음구간 321

6.6.4 큰표본의 믿음구간 324

6.6.5 표본량 확정 325

6.6.6 두 정규모집단에 대한 믿음구간 327

# 제5장 통계량과 그분포

앞의 4개 장은 확률론의 범주에 속한다. 이미 본바와 같이 우연량과 그 확률분포는 우연현상의 통계적성질을 전면적으로 묘사하고있다. 확률론의 많은 문제들에서 일반적으로 확률분포는 알려진다고 가정되며 모든 계산과 추리는 이 기지분포에 기초하여 진행된다. 그러나 실천문제들에서는 그렇지 않은 경우가 많다. 한가지 실례를 보자.

실례 5.0.1 어떤 회사에서 제품을 구입하려고 하는데 매 제품은 합격품 또는 불합격품이지만 이 제품무지에 대하여서는 불합격률 가 존재하게 된다. 그러므로 만일 이 제품무지에서 우연적으로 한개를 뽑고 로 뽑아진 제품의 불합격품수를 표시한다면 는 2점분포 에 따르지만 그 파라메터 는 미지이다. 분명히 의 크기는 이 제품무지의 품질을 결정한다. 때문에 사람들은 에 대하여 다음과 같은 문제를 제기한다.

• 의 크기는 어떤가?

• 가 어떤 범위의 값을 취하는가?

• 가 설정된 요구조건(실례로 )을 만족한다고 볼수 있는가?

실례 5.0.1에서 취급하는 문제는 통계학의 범주에 속한다. 통계학은 응용성이 매우 강한 학문으로서 300여년의 력사를 가지고 있다. 칼 피어슨(K.Pearson, 1857-1936)과 피셔(R.A.Fisher, 1890-1962)의 연구로부터만 보아도 약 200여년의 력사를 가지고있으며 사회경제적효과성을 가져왔다.

일반적으로 통계학은 우연적인 영향을 받는 자료를 어떻게 효과적으로 수집하고 분석하는가를 연구하는 학문으로 알려져 있다. 다년간의 연구와 발전을 거쳐 통계학은 여러 학문에 침투하게 되였다. 많은 자료와 관련된 실천문제를 통계학의 방법으로 분석하고 해결할수 있으며 또 해결하여야 한다. 통계학의 발전과 완성과정에 따라 연구내용은 매우 풍부해졌으며 표본조사, 실험설계, 회귀분석, 다변량통계분석, 시계렬분석, 비파라메터통계, 베이스방법 등 여러 분야를 이루고있다.

이제 통계학의 가장 기본적인 개념인 모집단과 표본에 대한 학습으로부터 시작하여 통계학의 내용들을 소개하겠다.

## 5.1모집단과 표본

### 5.1.1 모집단과 개체

하나의 통계문제에서 우리는 연구대상의 전체를 **모집단**이라고하고 그것을 구성하는 매개 성원을 **개체**라고 부른다. 대부분의 실천문제들에서 모집단속의 개체는 실제 존재하는 사람이나 사물이다. 례컨대 어떤 대학 학생들의 키상태를 연구할 때 그 대학 전체 학생들은 문제의 모집단을 이루며 매 학생은 곧 하나하나의 개체이다. 사실 매 학생은 성별, 나이, 키, 몸무게, 민족, 출생지 등과 같은 많은 특징을 가지고있다. 그러나 이 문제에서 관심하는것은 오직 이 학교 학생들의 키가 어떠한것이고 다른 특징들은 고려하지 않는다. 이렇게 매 학생(개체)이 가지고있는 수량지표값인 키가 곧 개체이고 모든 키 전체를 총체라고 본다. 그러면 문제의 배경을 고려하지 않을 때 모집단은 바로 수들의 모임으로서 그가운데는 큰것도, 작은것도 있고 어떤것은 나타나는 기회가 크고 어떤것은 작다. 이렇게 보면 어떤 확률분포로 모집단을 묘사하고 요약하는것이 적합하다. 이런 의미에서 **모집단은 곧 하나의 분포**이고 수량지표는 이 분포에 따르는 우연량이다. 앞으로 '모집단으로부터 발취한다'는 표현과 '어떤 분포로부터 발취한다'는것은 같은 의미로 쓰겠다.

실례5.1.1 테프의 품질지표는 그 테프에 있는 흠집의 개수이다. 테프마다 흠집의 수가 대응되는데 모든 테프들의 흠집수는 하나의 모집단을 이룬다. 이 모집단에서 대부분은0(흠집이 없는 합격품)이고 1, 2, 3 등도 있지만 8개이상인 경우는 드물다. 연구에 의하면 테프의 흠집수 는 뽜쏭분포 에 따르지만 분포의 파라메터 는 알려져있지 않다. 분명히 의 크기가 제품의 품질을 결정하며 생산자의 경제적리득에 직접 영향을 준다.

이 실례에서 모집단분포의 형태는 뽜쏭분포로 규정되여있지만 아직 미지파라메터 를 포함하기때문에 모집단분포를 확정하려면 를 결정해야 한다. 이것이 바로 통계학이 해결하려는것이다.

실례5.1.2 일반적인 측정문제를 생각하자. 측정인원이 물리적량 를 반복측정할 때 가능한 모든 측정결과는 의 실수이므로 모집단은 에서 값을 취하는 우연량 이다. 그렇다면 이 모집단의 분포에 대해 무엇을 알수 있는가?

한가지 확정할수 있는것은 측정결과 는 물리적량 와 측정오차 의 합계 즉 이라는것이다. 여기서 는 확실하지만 알려지지 않은 량(파라메터라고 부른다)이고 은 우연량이다. 그러므로 모집단분포에 관한 가정은 주로 의 분포에 관한 가정이다. 다음과 같은 가정들이 어떤 경우에는 합리적이다.

① 중심극한정리에 의하여 가장 보편적인것은 이라고 가정하는것이다. 이때 측정값의 모집단은 하나의 정규분포 즉 이다. 이 모집단에는 두개의 미지수 가 있는데 이것들을 추론하는것은 통계학에서 연구하여야 할 문제이다.

② 가령 오차가 정규분포를 따른다는것뿐 아니라 분포의 분산(일반적으로 측정체계 자체의 정밀도에 의해 결정될수 있다.)도 안다면 라고 가정할수 있다. 여기서 은 기지의 상수이다. 이렇게 모집단은 여전히 정규분포족 이지만 모집단에 한개의 미지수 가 있고 그것을 추론하는것은 통계학이 연구하여야 할 문제이다.

③ 만일 오차가 정규분포에 따른다는 근거가 없지만 그 분포가 0대칭이라면 모집단분포는 분포의 류형은 모르지만 일정한 제한을 가지는 분포로 된다. 이때 일반적으로 유한개의 파라메터들에 의하여 설명할수 없으며 이것을 비파라메터분포라고 부른다. 이 부분에 대한것은 7장에서 취급하게 된다.

어떤 문제들에서 우리는 매 연구대상에 대하여 두개 또는 그 이상의 지표들을 관측해야 할수 있으며 이때 다차원우연벡토르들과 그의 동시분포로 모집단을 설명할수 있다. 이러한 모집단을 **다차원모집단**라고 부른다. 례를 들어 학생의 나이, 키, 몸무게 등 세가지 지표를 료해하기 위해서는 3차원우연벡토르를 리용하여 모집단을 설명할수 있다. 이것은 3차원모집단로서 다차원분석의 연구대상이다.

1차원모집단과 2차원모집단은 기본적인것으로서 이것들을 잘 리해하면 더 높은 차원의 모집단을 연구하는데 도움이 된다. 그래서 여기서는 1차원모집단만을 취급하며 일부 내용은 2차원모집단도 취급한다.

모집단에는 또한 유한모집단과 무한모집단이 있는데 여기서는 무한모집단을 기본연구대상으로 한다.

### 5.1.2 표본

모집단의 분포를 알기 위하여 모집단으로부터 우연적으로 개의 개체를 발취하고 그 지표값을 이라고 표시하자. 그러면 을 모집단의 **표본**, 을 **표본용량**(**표본량**)이라고 부른다.

우선 표본은 2중성을 가진다. 한편으로 표본은 모집단로부터 임의로 발취되는것으로서 그전에 그 수값을 예측할수 없다. 따라서 표본은 우연량이라고 볼수 있으며 으로 표시할수 있다. 다른 한편 표본은 발취된 다음 관측을 거쳐 확정적인 관측값을 가지기때문에 한조의 수값이라고 볼수 있다. 그러므로 이때는 로 표시하는것이 타당하다. 그러나 여기서는 편리상 표본이든 그의 관측값이든 로 표시하겠다. 그러므로 독자들이 리해상 구분하여야 한다.

실례 5.1.3 맥주공장에서 생산하는 병맥주의 함량은 640ml로 규정되여있는데 우연성때문에 사실상 모든 맥주의 함량을 640ml로 하는것은 불가능하다. 이제 어떤 맥주공장에서 생산하는 맥주가운데서 우연적으로 10병을 발취하고 그 함량을 측정하여 다음과 같은 결과를 얻었다(단위:ml).

641, 635, 640, 637, 642, 638, 645, 643, 639, 640

이것은 용량이 10인 표본의 관측값으로서 대응하는 모집단은 그 공장에서 생산하는 병맥주의 함량이다.

실례5.1.4(그룹표본) 어떤 공장에서 생산하는 전자소자의 수명을 조사한다. 이 공장에서 생산하였거나 생산하게 될 이 종류의 모든 소자의 수명은 모집단(무한모집단이라고 생각할수 있다)이고 그속에서 100개를 선택하여 그 수명을 측정한다. 일정한 원인으로 하여 시시각각 실험을 관측할수 없으며 오직 정기(례컨대 24h)적으로 관측한다. 그러면 매 소자에 대하여 그의 수명이 어떤 범위에 속하는가만을 관측할수 있는데 이것을 표 5.1.1에 보여주었다.

표 5.1.1 100개 소자의 수명자료

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 수명범위 | 소자개수 | 수명범위 | 소자개수 | 수명범위 | 소자개수 |
| (0, 24] | 4 | (192, 216] | 6 | (384, 408] | 4 |
| (24, 48] | 8 | (216, 240] | 3 | (408, 432] | 4 |
| (48, 72] | 6 | (240, 264] | 3 | (432, 456] | 1 |
| (72, 96] | 5 | (264, 288] | 5 | (456, 480] | 2 |
| (96, 120] | 3 | (288, 312] | 5 | (480, 504] | 2 |
| (120, 144] | 4 | (312, 336] | 3 | (504, 528] | 3 |
| (144, 168] | 5 | (336, 360] | 5 | (528, 552] | 1 |
| (168, 192] | 4 | (360, 384] | 1 | >552 | 13 |

표 5.1.1의 표본관측값은 구체적인 값이 없고 범위만 있다. 이런 표본을 **그룹표본**이라고 하는데 한가지 **불완전표본**이다. 반대로 실례 5.1.3에서 10개의 맥주함량은 **완전표본**이라고 부른다. 그룹표본은 완전표본에 비하여 언제나 정보의 손실이 있는데 이것은 그룹표본의 부족점이다. 더 많은 정보를 얻기 위해서는 완전표본을 얻어야 하지만 이 실례에서와 같이 부득이한 경우 그룹표본을 리용할수 있다. 그러나 실천에서 표본량가 특별히 클 때(례하면 일 경우) 흔히 완전표본대신 그룹표본을 리용하는데 이때 표본에 대한 그룹화를 진행하고 정리할 필요가 있다. 그룹표본은 표본을 간단명료하게 표시할수 있으며 사람들로 하여금 모집단을 더 잘 인식할수 있게 하다는 우점을 가지고 있다.

모집단으로부터 표본을 발취하는것도 몇가지 방법이 있다. 표본에 의하여 모집단에 대한 믿음직한 추론을 내리기 위해서는 표본이 모집단을 잘 대표하기 바란다. 따라서 발취방법에 대한 몇가지 요구가 제기되는데 많이 리용되는 단순우연발취방법에 대해서는 다음과 같은 두가지 요구가 나선다.

• 표본이 **대표성**을 만족하여야 한다. 즉 모집단의 매 개체가 표본으로 선택될 동등한 동등한 기회를 가져야 이것은 매개 표본 가 모집단 와 같은 분포를 가진다는것을 의미한다.

• 표본이 **독립성**을 만족하여야 한다. 즉 표본에서 매개의 값이 다른것의 값에 영향을 주지 말아야 한다. 이것은 이 서로 독립이라는것을 의미한다.

단순우연발취방법으로 얻은 표본을 **단순우연표본**이라고 한다. 여기서 특별히 지적하지 않는 한 모든 표본은 단순우연표본이다. 그러므로 표본 는 서로 독립이고 동일분포를 가지는 우연량이라고 볼수 있다. 이것을 i.i.d표본이라고도 부르는데 그 공동분포는 바로 모집단분포이다.

모집단 가 분포함수 를 가지고 을 그 모집단에서 발취한 용량이 인 표본이라고 하면 표본의 **동시분포함수**는 이다.

무한모집단에 대하여서는 대표성과 독립성이 쉽게 만족되는데 중요한것은 의식적이든 무의식적이든 인위적인 간섭을 제거하는것이다. 유한모집단에서는 모집단의 개체수가 많을 때 특히 표본량에 비하여 매우 많을 때 독립성도 기본상 만족된다.

실례 5.1.5 개의 제품이 있는데 표본검사를 진행하여 불합격품률 를 알려고 한다. 이때 개를 발취하여 불합격품인가를 하나씩 검사한다. 만일 합격품을 0, 불합격품을 1로 표시한다면 모집단은 다음과 같은 2점분포이다.

.

하나하나씩 발취된 표본을 로 표시하고 반환발취를 진행한다면 는 독립동일분포에 따른다. 그러나 비반환발취(실제적으로 진행하는 발취는 보통 이런 형식이다)를 진행한다면 두번째로 발취하는 불합격품의 확률은 처음으로 발취한것이 합격품인가 불합격품인가에 의존한다. 두번째로 발취한것이 불합격품일 확률은 만일 처음으로 발취한것이 불합격품이면 이고 합격품이면 이다.

분명히 이렇게 얻은 표본은 단순우연표본이 아니다. 그러나 이 매우 클 때 우의 두가지 경우에 확률이 모두 와 근사하다는것을 알수 있다. 따라서 이 크고 이 크지 않을 때(경험적규칙은 ) 이 표본은 근사적으로 단순우연표본으로 볼수 있다.

련습문제 5.1

1. TV방송국이 어떤 TV프로(실례로 매일 저녁 9시~9시 반까지의 체육프로)의 시청률정형을 료해하려고 한다.

① 이 연구의 모집단은 무엇인가?

② 이 연구의 표본은 무엇인가?

2. 어떤 도시에서 성인남자의 흡연률을 조사하려고 50명의 전문가를 특별히 초청하여 우연조사를 하게 하였다. 매 전문가마다 100명의 성인남자를 조사할것을 요구한다면 이 조사에서 모집단과 표본은 각각 무엇이며 모집단은 어떤 분포로 묘사하면 좋겠는가?

3. 한 공장에서 어떤 제품을 대량생산하는데 그 불합격품률 는 미지이며 개 제품을 한통으로 포장한다. 제품의 품질을 검사하기 위하여 임의로 개 통을 선택하고 그중의 불합격품수를 조사한다. 이때 모집단과 표본은 각각 무엇이며 표본분포를 지적하시오.

4. 양어장에 몇마리의 물고기가 있는가를 평가하기 위하여 다음과 같은 방안을 제기하였다. 물고기를 한 그물 잡아올린 다음 그속의 마리 모두에 물에 씻기지 않는 빨간 색칠을 한 다음 다시 놓아준다. 하루 지나서 다시 물고기를 한 그물 걷어올리고 보니 마리가 있는데 빨간 색칠을 한 물고기가 마리 있었다. 양어장에 물고기가 대략 몇마리 있는가를 추정할수 있는가? 모집단과 표본은 무엇인가?

5. 어떤 공장에서 생산하는 축전기의 수명은 지수분포에 따른다. 그 평균수명을 알기 위하여 개 제품을 선택하여 실제 수명을 측정한다. 이때 무엇이 모집단이고 표본인가? 표본분포를 지적하시오.

## 5.2 표본자료의 정리와 표시

### 5.2.1 경험적분포함수

이 분포함수가 인 모집단으로부터 발취한 표본이라고 하자. 이제 표본관측값들을 작은것으로부터 큰것으로 정렬하여 으로 표시하면 이것을 **순서표본**이라고 부른다. 순서표본을 리용하여 다음과 같은 함수를 정의한다.



이때 는 비감소오른쪽련속함수이며 을 만족한다. 이로부터 는 하나의 분포함수이며 이것을 그 표본의 **경험적분포함수**라고 부른다.

실례 5.2.1 어느한 식료품공장에서 통졸임음료를 생산하는데 생산선에서 임의로 5통을 선택하여 순중량(단위:g)을 잰 결과는 351, 347, 355, 344, 351이다. 이것은 용량이 5인 표본으로서 정렬하면 다음과 같은 순서표본을 얻을수 있다.

.

그 경험적분포함수(그림 5.2.1에 표시)는 다음과 같다.



매 고정된 에 대하여 는 표본가운데서 사건 가 발생하는 잦음률이다. 이 고정일 때 는 표본의 함수로서 하나의 우연량이다.

임의로 주어진 실수 에 대하여 를 정의하면 경험적분포함수의 정의로부터 이다.



1.0

0.8

0.6

0.4

0.2

*O*

344

347

351

355



그림 5.2.1 경험적분포함수

가 독립동일분포하는 우연량이고 그 공동분포가 라는것을 생각하면 베르느이의 큰수법칙으로부터 이 클 때 는 로 확률수렴한다. 더 의의있는 결과도 있는데 바로 글리벤코의 정리이다. 아래에서 증명을 하지 않고 소개만 하겠다.

정리 5.2.1[글리벤코(Glivenko)정리] 이 분포함수가 인 모집단으로부터 발취한 표본, 는 그의 경험적분포함수라고 하자. 이때 이면 이다.

정리 5.2.1는 이 매우 클 때 경험적분포함수는 모집단분포함수 에 대한 좋은 근사라는것을보여준다. 고전통계학에서 모든 통계적추론이 표본을 근거로 하는 리유도 여기에 있다.

### 5.2.2 빈도수표와 잦음률표

표본자료의 정리는 통계연구의 기초이며 자료를 정리하는 가장 일반적인 방법의 하나는 그 빈도수표 또는 잦음률표를 작성하는것이다. 그것을 다음의 실례로부터 소개한다.

실례 5.2.2 어느한 공장 로동자들의 어떤 제품의 생산능력을 연구하기 위하여 우연적으로 20명 로동자들의 어느 하루의 제품생산량을 조사하였다. 자료는 다음과 같다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 160 | 196 | 164 | 148 | 170 |
| 175 | 178 | 166 | 181 | 162 |
| 161 | 168 | 166 | 162 | 172 |
| 156 | 170 | 157 | 162 | 154 |

이제 20개 자료(표본)를 정리하는데 구체적인 절차는 다음과 같다.

① 표본을 그룹화한다. 일반적으로 그룹의 수는 5~20개이며 용량이 작은 표본은 보통 5~6개 그룹으로, 용량이 100정도인 표본은 7~10개로, 용량이 200정도인 표본은 9~13개, 용량이 300이상인 표본은 12~20개 그룹으로 나눌수 있다. 그 목적은 충분한 수의 그룹을 리용하여 자료의 변이를 나타낼수 있게 하는것이다. 이 례에는 20개의 자료만 있는데 이것을 5개 그룹 즉 로 나눈다.

② 매 그룹간격을 확정한다. 매 그룹의 구간길이는 같을수도 있고 같지 않을수도 있다. 실천에서는 흔히 길이가 같은 구간을 선택하여 비교하는데 편리하도록 한다. 이때 매 그룹의 구간길이를 **그룹간격**이라고 하며 그 근사공식은 이다. 여기서 는 각각 표본의 최대, 최소 관측값이다.

이 실례에서 최대관측값은 196, 최소관측값은 148이므로 그룹간격은 근사적으로 이다.

편리상 그룹간격을 10으로 한다.

③ 매 그룹의 제한을 확정한다. 매 그룹구간의 끝점은 로서다음과 같은 그룹구간을 형성한다.

.

여기서 은 최소관측값보다 약간 작고 는 최대관측값보다 약간 크다. 이 실례에서는 , 를 취할수 있다. 따라서 구간범위는

(147,157], (157,167], (167,177], (177,187], (187,197].

일반적으로 매 그룹을 (그룹상한+그룹하한)/2로 정의되는 그룹중심값으로 대표할수 있다.

④ 표본자료가 매 구간에 속하는 개수(빈도수)를 통계하고 그 잦음률표를 만든다. 이 실례의 잦음률을 표 5.2.1에 주었다. 이 표로부터 많은 정보를 알수 있다. 례를 들면 40% 로동자들의 생산량이 157~167이라는것, 생산량이 167개보다 적은 사람이 12명으로서 60%를 차지하며177개이상인 사람은 3명으로서 15%를 차지한다는것 등을 들수 있다.

표 5.2.1 실례 5.2.2의 빈도수, 잦음률표

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 그룹번호 | 그룹구간 | 그룹중심값 | 빈도수 | 잦음률 | 루적잦음률(%) |
| 1 | (147,157] | 152 | 4 | 0.20 | 20 |
| 2 | (157,167] | 162 | 8 | 0.40 | 60 |
| 3 | (167,177] | 172 | 5 | 0.25 | 85 |
| 4 | (177,187] | 182 | 2 | 0.10 | 95 |
| 5 | (187,197] | 192 | 1 | 0.05 | 100 |
| 합계 |  |  | 20 | 1 |  |

### 5.2.3 표본자료의 도형표현

앞에서 도표형식을 소개하였는데 많은 경우에 보다 직관적인 도형으로 표현할수도 있다.

1. 히스토그람

빈도수분포에서 가장 흔히 리용하는 도형표현은 **히스토그람**인데 그룹간격이 같은 경우에 너비가 같은 길쭉한 직4각형으로 표시되며 직4각형의 높이가 빈도수의 크기를 나타낸다. 도형에서 가로자리표는 관심하는 변수의 값구간을, 세로자리표는 빈도수를 표시하여 빈도수히스토그람을 얻을수 있다. 그림 5.2.2는 실례 5.2.2의 빈도수히스토그람을 그린것이다. 세로축을 잦음률로 바꾸면 잦음률히스토그람이 얻어진다.

모든 직4각형들의 면적의 합이 1이 되도록 하기 위하여 세로축을 잦음률/그룹간격으로 할수 있는데 이렇게 얻은 히스토그람을 단위잦음률히스토그람 또는 간단히 잦음률히스토그람이라고 한다. 이 세 종류의 히스토그람은 세로축의 눈금선택만 차이날뿐 히스토그람 그 자체는 아무런 변화도 없다.

5

4

3

2

1

6

7

8

그림 5.2.2 실례 5.2.2의 빈도수히스토그람

*O*

147

157

167

177



187

197

빈도수

2. 줄기잎도형(stem-and-leaf display)

히스토그람외에 흔히 리용하는 다른 한가지 방법은 줄기잎도형이다. 다음의 실례로 설명한다.

실례 5.2.3 어떤 회사에서 응모자들에 대하여 능력평가를 진행한다. 150점을 기준으로 한 응모자 50명의 시험성적을 정렬한것은 다음과 같다.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 64 | 67 | 70 | 72 | 74 | 76 | 76 | 79 | 80 | 81 |
| 82 | 82 | 83 | 85 | 86 | 88 | 91 | 91 | 92 | 93 |
| 93 | 93 | 95 | 96 | 96 | 97 | 97 | 99 | 100 | 100 |
| 102 | 104 | 106 | 106 | 107 | 108 | 108 | 112 | 112 | 114 |
| 116 | 118 | 119 | 119 | 122 | 123 | 125 | 126 | 128 | 133 |

이 자료들로 줄기잎도형을 만드는데 매 수값을 두 부분으로 나누어 앞부분(백의 자리, 열의 자리)을 줄기로, 뒤부분(하나의 자리)을 잎으로 한다.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 수값 |  | 분리 |  | 줄기 | 잎 |
| 112 | → | 11 | 2 | → | 11 | 2 |

그리고 세로줄을 하나 그리고 왼쪽에 줄기를, 오른쪽에 잎을 써넣으면 줄기잎도형이 얻어진다. 응모자들의 시험성적에 대한것을 그림 5.2.3에 주었다.

줄기잎도형의 생김새는 돌려놓은 히스토그람과 비슷하지만 구체적인 수값이 더 표시되였다. 그러므로 자료의 구체적인 값범위를 한눈에 알아볼수 있고 자료전부의 정보를 보존할수 있다.

두 그룹의 표본을 비교할 때 서로 등을 맞댄 줄기잎도형을 그려보는것은 간단하고 직관적이며 효과적인 방법이다.

실례5.2.4 아래의 자료는 한 공장의 두 직장에서 각각 40명 로동자들의 일생산량을 보여준것이다(표5.2.2). 이 자료를 비교하기 위해 서로 등을 맞댄 줄기잎도형을 그려보았다(그림 5.2.4).

그림 5.2.4에서 줄기는 중간에 있으며 왼쪽은 직장 1의 자료를, 오른쪽은 직장 2의 자료를 표시하였다. 줄기잎도형으로부터 직장 1의 로동자들의 생산량은 웃쪽에 편중되여있고 직장 2의 로동자들의 생산량은 대부분 중간에 위치하여있다는것, 직장 2의 평균생산량은 직장 1보다 높다는것, 직장 2의 매 로동자들의 생산량은 비교적 집중되여있는 반면에 직장 1의 로동자들의 생산량은 상대적으로 분산되여있다는것을 알수 있다.

6

7

8

9

10

11

12

13

4 7

0 2 4 6 6 9

0 1 2 2 3 5 6 8

1 1 2 3 3 3 5 6 6 7 7 9

0 0 2 4 6 6 7 8 8

2 2 4 6 8 9 9

2 3 5 6 8

3

그림 5.2.3 시험성적에 대한 줄기잎도형

5

6

7

8

9

10

6

6 7 7 8 8

2 2 4 5 5 5 5 6 6 6 6 8 8 9

0 1 1 3 3 3 4 4 4 6 6 7 7 8

2 2 3 5 8

7

그림 5.2.4 두 직장 생산량에 대한 등을 맞댄 줄기잎도형

6 2 0

8 7 7 7 5 5 5 4 2 1 1

8 7 7 6 6 4 4 2 1

8 7 6 6 5 3 2

7 3 3 2 1 0

5 3 0 0

직장 1

직장 2

표 5.2.2 어느한 공장 두 직장의 40명 로동자들의 일생산량

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 직장 1 | | | | | | 직장 2 | | | | | |
| 50 | 52 | 56 | 61 | 61 | 62 | 56 | 66 | 67 | 67 | 68 | 68 |
| 64 | 65 | 65 | 65 | 67 | 67 | 72 | 72 | 74 | 75 | 75 | 75 |
| 67 | 68 | 71 | 72 | 74 | 74 | 75 | 76 | 76 | 76 | 76 | 78 |
| 76 | 76 | 77 | 77 | 78 | 82 | 78 | 79 | 80 | 81 | 81 | 83 |
| 83 | 85 | 86 | 86 | 87 | 88 | 83 | 83 | 84 | 84 | 84 | 86 |
| 90 | 91 | 92 | 93 | 93 | 97 | 86 | 87 | 87 | 88 | 92 | 92 |
| 100 | 100 | 103 | 105 |  |  | 93 | 95 | 98 | 107 |  |  |

련습문제 5.2

1. 아래에 어느한 공장에서 표본조사를 통하여 얻은 로동자 10명의 한주일간 제품생산량을 주었다.

149, 156, 160, 138, 149, 153, 153, 153, 169, 156, 156.

이 자료를 가지고 경험적분포함수를 만들고 도형으로 표현해보시오.

2. 다음의 표는 정리하여 얻은 그룹표본을 보여준다.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 그룹번호 | 1 | 2 | 3 | 4 | 5 |
| 그룹구간 | (38,48] | (48,58] | (58,68] | (68,78] | (78,88] |
| 빈도수 | 3 | 4 | 8 | 3 | 2 |

이 그룹표본의 경험적분포함수를 구하시오.

3. 어느한 지역에서 어떤 전공지표의 졸업생 30명의 2018년 생활비자료는 다음과 같다.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 9090 | 10860 | 1200 | 9990 | 13200 | 10910 |
| 10710 | 10810 | 10300 | 13360 | 9670 | 15720 |
| 8250 | 9140 | 9920 | 12320 | 9500 | 7750 |
| 12030 | 10250 | 10960 | 8080 | 12240 | 10440 |
| 8710 | 11640 | 9710 | 9500 | 8660 | 7380 |

① 이 자료의 잦음률분포(6개 조로 나눌것)를 구성하시오.

② 히스토그람을 그려보시오.

4. 어느한 회사에서 250명 종업원들의 출근소요시간(단위:분)을 조사하였는데 다음과 같은 불완전잦음률분포표가 얻어졌다.

|  |  |
| --- | --- |
| 출근소요시간 | 잦음률 |
| 0~10 | 0.10 |
| 10~20 | 0.24 |
| 20~30 |  |
| 30~40 | 0.18 |
| 40~50 | 0.14 |

① 잦음률분포표를 완전하게 보충해보시오.

② 출근소요시간이 반시간이하인 종업원은 몇명인가?

5. 간행물 40종의 월발행량(단위:100부)은 다음과 같다.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 5954 | 5022 | 14667 | 6582 | 6870 | 1840 | 2662 | 4508 |
| 1208 | 3852 | 618 | 3008 | 1268 | 1978 | 7963 | 2048 |
| 3077 | 993 | 353 | 14263 | 1714 | 11127 | 6926 | 2047 |
| 714 | 5923 | 6006 | 14267 | 1697 | 13876 | 4001 | 2280 |
| 1223 | 12579 | 13588 | 7315 | 4538 | 13304 | 1615 | 8612 |

① 그룹간격을 1700(100권)으로 하여 이 자료의 잦음률분포표를 만드시오.

② 히스토그람을 그려보시오.

6. 아래의 자료에 대하여 줄기잎도형을 구성하시오.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 472 | 425 | 447 | 377 | 341 | 369 | 412 | 399 |
| 400 | 382 | 366 | 425 | 399 | 398 | 423 | 384 |
| 418 | 392 | 372 | 418 | 374 | 385 | 439 | 408 |
| 429 | 428 | 430 | 413 | 405 | 381 | 403 | 479 |
| 381 | 443 | 441 | 433 | 339 | 379 | 386 | 387 |

7. 아래의 자료에 대하여 줄기잎도형을 구성하시오.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 40.6 | 39.6 | 37.8 | 36.2 | 38.8 |
| 38.6 | 39.6 | 40.0 | 34.7 | 41.7 |
| 38.9 | 37.9 | 37.0 | 35.1 | 36.7 |
| 37.1 | 37.7 | 39.2 | 36.9 | 38.3 |

## 5.3 통계량과 그 분포

### 5.3.1 통계량과 표본화분포

표본은 모집단으로부터 얻어지기때문에 표본속에는 모집단의 여러 방면의 정보가 들어있지만 이런 정보는 비교적 분산되여 때로는 무질서해보인다. 표본속에 분산되여있는 모집단에 관한 정보를 집중하여 모집단의 여러가지 특징을 반영하기 위해서는 표본에 대한 처리를 진행하여야 한다. 표와 도표는 일종의 처리형식으로서 사람들로 하여금 모집단에 대한 초보적인 인식을 얻도록 한다. 사람들이 표본으로부터 모집단의 여러가지 파라메터들에 대하여 알려고 할 때 더 효과적인 처리방법은 표본의 함수를 구성하는것인데 이때 서로 다른 표본함수는 모집단의 각이한 특징을 반영할수 있다.

정의 5.3.1 가 어떤 모집단에서 발취한 표본이라고 하자. 표본함수 에 미지파라메터가 포함되여 있지 않으면 를 **통계량**(statistic)이라고 하고 통계량의 분포를 **표본분포**(sampling distribution)라고 한다.

이 정의에 따르면 만일 가 표본이면 과 5.2.1에서 취급한 도 모두 통계량이다. 그러나 이 미지일 때 와 같은것들은 통계량이 아니다. 강조하여둘것은 통계량은 미지파라메터에 의존하지는 않지만 그의 분포는 미지파라메터에 의존한다는것이다.

아래에서와 5.4에서 흔히 보는 통계량과 그 표본분포를 소개한다.

### 5.3.2 표본평균과 그 표본분포

정의 5.3.2 가 어떤 모집단에서 발취한 표본이라고 하자. 이때 그의 산수평균값을 표본평균이라고 하며 일반적으로 로 표시한다. 즉

.

그룹표본의 경우에 표본평균의 근사공식은 다음과 같다.



여기서 는 그룹수이고 는 째 그룹의 중심값, 는 째 그룹의 빈도수이다.

실례 5.3.1 어느한 단위에서 수집한 20명 청년들의 한달간 문화생활에 지출한 비용에 대한 자료는 다음과 같다.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 790 | 840 | 840 | 880 | 920 | 930 | 940 | 970 | 980 | 990 |
| 1000 | 1010 | 1010 | 1020 | 1020 | 1080 | 1100 | 1130 | 1180 | 1250 |

20명의 평균비용은 이다.

이 20개의 자료를 그룹화하면 아래와 같은 빈도수, 잦음률표를 얻을수 있다.

표 5.3.1 실례 5.3.1의 빈도수, 잦음률분포표

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 그룹번호 | 그룹구간 | 그룹중심값 | 빈도수 | 잦음률(%) |
| 1 | (770,870] | 820 | 3 | 15 |
| 2 | (870,970] | 920 | 5 | 25 |
| 3 | (970,1070] | 1020 | 7 | 35 |
| 4 | (1070,1170] | 1120 | 3 | 15 |
| 5 | (1170,1270] | 1220 | 2 | 10 |
| 합계 |  |  | 20 | 100 |

표 5.3.1의 그룹표본에 대하여 식 (5.3.2)을 리용하여 계산하면

.

이렇게 두가지 계산결과가 서로 다르다. 그것은 식 (5.3.2)이 실제 표본관측자료를 리용하지 않고 근사결과만을 주기때문이다.

표본평균에 대하여 다음과 같은 몇가지 성질이 있다.

성질 5.3.1 표본자료와 표본평균사이의 차를 편차라고 부른다면 표본의 모든 편차들의 합은 0 즉 이다.

평균값의 계산공식으로부터 볼 때 이 계산은 모든 자료를 리용하며 매 자료도 식에서 동등한 지위에 있다. 표본중심 와 자료들사이의 편차는 정수 또는 부수가 될수 있으며 이것은 서로 상쇄되여 표본의 모든 편차들의 합은 0이 되여야 한다.

성질 5.3.2 표본관측값과 평균값의 편차의 두제곱합이 제일 작다. 다시 말하여 모양의 함수들가운데서 이 제일 작다. 여기서 는 임의로 주어진 상수이다.

증명. 임의로 주어진 상수 에 대하여



아래에서 표본평균값의 분포를 보자.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 11 | 8 |  | 표본 1 | 표본 2 | 표본 3 | 표본 4 |
| 12 | 13 |
| 8 | 9 |  | 11  11  9  10  8 | 8  13  10  11  9 | 13  11  11  10  9 | 12  9  10  10  11 |
| 11 | 10 |  |
| 9 | 11 |  |
| 10 | 8 |  |
| 10 | 12 |  |
| 11 | 9 |  |
| 8 | 11 | 표본평균 | 9.8 | 10.2 | 10.8 | 10.4 |
| 10 | 13 |

그림 5.3.1 4개 표본의 표본평균

실례 5.3.2 이제 20개의 수로 이루어진 모집단이 있다고 하고 그것으로부터 동시에 용량이 5인표본을 발취한다고 하자. 그림 5.3.1은 첫째 표본의 발취과정을 보여주고있는데 왼쪽은 모집단, 오른쪽은 모집단로부터 우연발취한 표본이다. 표본을 기록한 다음 다시 반환하고 두번째 표본을 발취한다. 여기서는 총 4개의 표본을 발취하며 매 표본에는 5개의 관측값이 있다.

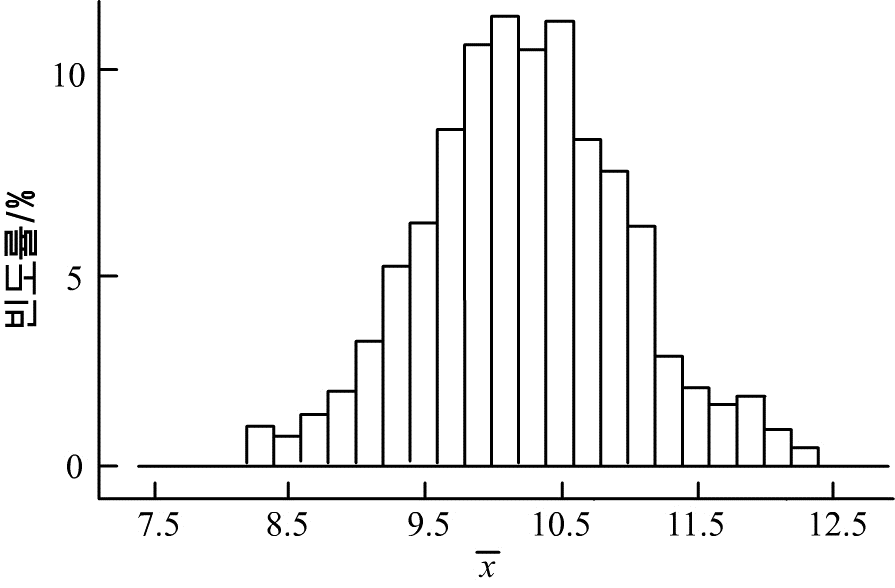


그림 5.3.2 500개 표본의 평균을 나타내는 히스토그람

매 표본의 표본평균을 계산하였는데 표본발취의 우연성으로 하여 표본평균값들도 서로 다르다.

만일 비슷하게 표본 5, 표본 6, …을 발취하고 매번 표본평균 를 계산한다면 그것들사이의 차이는 표본발취의 우연성으로 하여 초래된것이다. 이렇게 무제한하게 계속 발취해나간다면 대량의 값들을 얻을수 있는데 그림 5.3.2는 이렇게 얻은 500개 값으로 만들어진 히스토그람으로서 의 표본분포를 반영한다. 그 형태가 정규분포와 매우 비슷한데 그것은 우연적인것이 아니라 다음과 같은 정리에 의하여 담보되는 사실이다.

정리 5.3.1 는 어떤 모집단으로부터 얻어진 표본이고 는 표본평균이다. 이때

① 모집단분포가 이면 의 **정확한 분포**는 이다.

② 만일 모집단분포가 미지이거나 정규분포가 아니고 이 존재한다면 이 비교적 클 때 의 **점근분포**는 이다. 이것을 일반적으로 로 표시한다. 여기서 점근분포는 이 비교적 클 때의 근사분포이다.

증명. ① 합성적공식을 리용하면 이라는것을 알수 있고 이로부터 이다.

② 중심극한정리로부터 인데 이것은 이 비교적 클 때 의 점근분포가 이라는것을 의미한다. 따라서 증명된다.

실례 5.3.3 그림 5.3.3은 세개의 서로 다른 모집단에 대한 표본평균값의 분포밀도함수를 보여준다. 세가지 모집단은 각각 ① 평등분포, ② 거꿀3각분포, ③ 지수분포이다. 표본량이 증가함에 따라 표본평균값 의 표본분포는 점차 정규분포에 접근하며 그것들의 평균값은 변하지 않고 분산은 원래의 로 줄어든다. 표본량이 30일 때 세개의 표본분포가 모두 정규분포로 근사한다는것을 알수 있다. 아래에서 그에 대하여 구체적으로 설명한다.

①의 모집단분포는 평등분포 이고 그 모집단분포의 평균과 분산은 각각 3, 4/3이다. 만일 그 모집단에서 용량이 30인 표본을 발취할 경우 그 표본평균의 점근분포는 다음과 같다.

.

②의 모집단확률밀도함수는 이다. 이것은 거꿀3각분포로서 그 평균과 분산이 각각 3, 2이다. 만일 그 모집단에서 용량이 30인 표본을 발취할 경우 그 표본평균의 점근분포는 다음과 같다.

.

③의 모집단분포는 지수분포 이고 그 평균과 분산은 모두 1이다. 만일 그 모집단에서 용량이 30인 표본을 발취할 경우 그 표본평균 의 점근분포는 다음과 같다.

.

이 세개의 모집단은 모두 정규분포가 아니지만 표본평균값의 분포는 모두 정규분포에 근사하며 차이나는것은 평균과 표준편차뿐이다. 그림 5.3.3에 보여준 곡선들은 이것들사이의 공통점과 차이점을 보여준다.

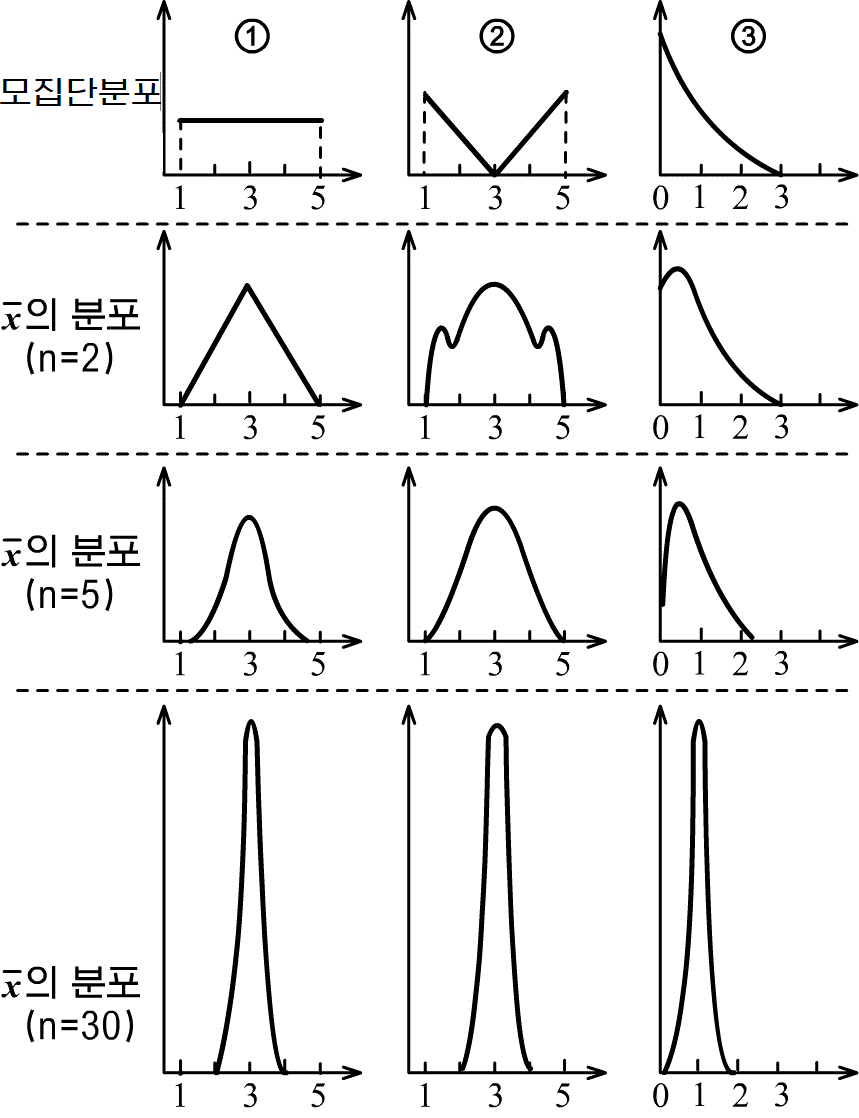


그림 5.3.3 서로 다른 모집단들의 표본평균분포

### 5.3.3 표본분산과 표준편차

정의 5.3.3 는 어떤 모집단으로부터 얻어진 표본이다. 그것과 표본평균 사이의 편차의 두제곱합의 평균



을 **표본분산**이라고 하고 그 두제곱뿌리 을 **표본표준편차**라고 한다. 표본표준편차는 표본평균과 측정단위가 같기때문에 표본분산에 비해 더 실용적이다. 이 크지 않을 때는 보통



을 표본분산 (**불편분산**이라고도 부른다. 6장에서 설명한다)으로 리용한다. 또한 그의 두제곱뿌리 을 표준편차라고 한다. 실천에서 은 보다 더 널리 리용된다. 앞으로 표준편차를 언급할 때 일반적으로 을 가리킨다.

표본분산( 또는 )은 표본의 흩어짐정도를 측정하기 위해 널리 리용되는 통계량으로서 그 정의에서 은 표본량, 를 편차의 두제곱합, 을 편차두제곱합의 자유도라고 한다. 그 의미는 가 결정되면 개의 편차 가운데서 개의 편차만이 자유롭게 변할수 있으며 번째는 이기때문에 마음대로 값을 취할수 없다는것이다.

표본편차의 두제곱합에는 다음과 같은 3가지 표시식이 있다.



이것들은 모두 표본분산을 계산하는데 리용될수 있다.

그룹표본인 경우 표본분산의 근사계산공식은 다음과 같다.

.

여기서 와 는 각각 번째 구간의 그룹중심값과 빈도수이며 는 식 (5.3.2)으로 주어지는 표본평균의 근사값이다.

실례 5.3.4 실례 5.3.1의 표본에 대하여 을 이미 알고있다. 그 표본분산과 표본표준편차는각각 다음과 같다.



표 5.3.1의 그룹표본에 대해서는 표 5.3.2(표본평균은 그룹표본에 의해 계산된다)에서와 같이 계산할수 있다.

표 5.3.2 그룹표본분산에 대한 계산표

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 그룹중심값 | 빈도수 |  |  |  |
| 820 | 3 | 2460 | -180 | 97200 |
| 920 | 5 | 4600 | -80 | 32000 |
| 1020 | 7 | 7140 | 20 | 2800 |
| 1120 | 3 | 3360 | 120 | 43200 |
| 1220 | 2 | 2440 | 220 | 96800 |
| 합계 | 20 | 20000 |  | 272000 |

그러므로 이다.

다음 정리는 표본평균의 수학적기대값과 분산, 표본분산의 수학적기대값이 모집단의 분포형태에의존하지 않는다는것을 보여준다.

정리 5.3.2 모집단 가 2차모멘트를 가진다고 하자. 즉 . 그리고 가 그 모집단으로부터 발취한 표본이고 , 이 각각 표본평균, 표본분산이라고 하면

,

.

이 정리는 표본평균의 기대값이 모집단평균과 같으며 표본평균의 분산은 모집단분산의 이라근것을 의미한다.

증명. 이므로 식 (5.3.7)이 성립된다. 이제 식 (5.3.8)을 증명한다. 이고 , 이므로 .

량변을 각각 로 나누면 식 (5.3.8)을 얻는다.

### 5.3.4 표본모멘트와 그 함수

표본평균과 분산의 일반화는 표본모멘트이며 이것은 널리 리용되는 통계량이다.

정의 5.3.4 는 표본, 는 정의 옹근수라고 하면 통계량



를 표본의 차원점모멘트라고 한다. 이때 표본의 1차원점모멘트는 표본평균이다. 통계량



를 차중심모멘트라고 한다. 이때 2차중심모멘트는 표본분산이다.

모집단이 분포중심에 관하여 대칭일 때 와 를 리용하여 표본특징을 묘사하는것은 대표적이지만 대칭이 아닌 경우에는 와 만으로는 부족하다. 이것을 위해서는 분포형태를 묘사하는 통계량이 필요하다(2.7.5, 2.7.6을 참고). 여기서는 표본의 비대칭도와 뾰족도를 소개하는데 이것들은 모두 표본중심모멘트의 함수이다.

정의 5.3.5 가 표본이면 통계량



을 **표본비대칭도**라고 한다.

표본비대칭도 는 표본자료의 대칭성에 대한 편기정도와 편기방향을 반영한다. 분명히 자료가 완전히 대칭이면 이며 비대칭이면 이다. 여기에서 을 으로 나누는것은 단위의 영향을 제거하기 위해서이다. 는 무본위량으로서 자료분포의 편기방향과 정도를 잘 묘사한다. 이면 표본이 대칭(그림 5.3.4(a) 참고)이라는것이며 가 0보다 훨씬 크면 표본의 오른쪽꼬리가 길다(그림 5.3.4(b) 참고)는것 다시 말하여 표본가운데 몇개의 비교적 큰수가 있어서 그것으로 모집단분포가 오른쪽으로 편기된것을 반영한다는것이다. 또한 가 0보다 훨씬 작으면 분포의 왼쪽꼬리가 길다(그림 5.3.4(c) 참고)는것이며 이것은 표본들가운데 몇개의 특별히 작은 수가 있어서 그것으로 모집단분포가 왼쪽으로 편기된것을 반영한다.

4

7

8

9

12

7

8

9

7

8

9

12

15

1

4

(a) 표본 (4, 7, 8, 9, 12)의 비대칭도

(b) 표본 (7, 8, 9, 12, 15)의 비대칭도

(c) 표본 (1, 4, 7, 8, 9)의 비대칭도

그림 5.3.4 표본비대칭도 의 실례(표본량 )

정의 5.3.6 가 표본이면 통계량



을 **표본뾰족도**라고 한다.

표본뾰족도는 모집단밀도곡선에서 봉우리주위의 가파로운 정도와 꼬리의 굵기를 반영하는 통계량이다. 가 0보다 훨씬 클 때 밀도곡선은 그 봉우리주위에서 정규분포에서보다 가파르고 꼬리는 더 가늘며 그러므로 뾰족한 정점형이라고 한다. 가 0보다 훨씬 작을 때 분포밀도곡선은 봉우리부근에서 정규분포보다 평탄하고 꼬리가 더 굵으므로 평평한 정점형이라고 한다.

실례 5.3.5 표 5.3.3은 각각 50명씩인 두 학급의 영어과목시험점수이다.

표 5.3.3 두 학급의 영어성적

|  |  |  |  |
| --- | --- | --- | --- |
| 성적 | 그룹중심값 | 학급 1의  인원수 | 학급 2의 인원수 |
| 90~100 | 95 | 5 | 4 |
| 80~89 | 85 | 10 | 14 |
| 70~79 | 75 | 22 | 16 |
| 60~69 | 65 | 11 | 14 |
| 50~59 | 55 | 1 | 2 |
| 40~49 | 45 | 1 | 0 |

이제 두 학급의 평균성적, 표준편차, 표본비대칭도와 표본뾰족도를 계산한다. 표 5.3.4와 표 5.3.5는 각각 학급 1과 2에 대한 계산과정을 보여준다.

표 5.3.4 학급 1의 성적에 대한 계산과정

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 95 | 5 | 475 | 1843.20 | 35389.440 | 679477.2480 |
| 85 | 10 | 850 | 846.40 | 7786.880 | 71639.2960 |
| 75 | 22 | 1650 | 14.08 | -11.264 | 9.0112 |
| 65 | 11 | 715 | 1283.04 | -13856.832 | 149653.7856 |
| 55 | 1 | 55 | 432.64 | -8998.912 | 187177.3696 |
| 45 | 1 | 45 | 948.64 | -29218.112 | 899917.8496 |
| 합계 | 50 | 3790 | 5368 | -8908.8 | 1987874.56 |

표 5.3.5 학급 2의 성적에 대한 계산과정

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 95 | 4 | 380 | 1474.56 | 28311.552 | 543581.7984 |
| 85 | 14 | 1190 | 1184.96 | 10901.632 | 100295.0144 |
| 75 | 16 | 1200 | 10.24 | -8.192 | 6.5536 |
| 65 | 14 | 910 | 1632.96 | -17635.968 | 190468.4544 |
| 55 | 2 | 110 | 865.28 | -17997.824 | 374354.7392 |
| 합계 | 50 | 3790 | 5168 | 3571.2 | 1208706.56 |

두 학급의 평균성적, 표준편차, 비대칭도와 뾰족도는 다음과 같다.



이로부터 두 학급의 평균성적이 같고 표준편차도 거의 같으며 표본비대칭도가 각각 -0.16, 0.068로서 두 학급의 성적이 모두 기본적으로 대칭이라는것을 알수 있다. 그러나 표본뾰족도는 두 학급에서의 차이가 명백하다. 학급 2의 성적분포는 비교적 평평한 반면에 학급 1은 어느 정도 봉우리부분이 두드러졌다.

### 5.3.5 순서통계량과 그 분포

표본모멘트외에 다른 한가지 대표적인 통계량은 순서통계량인데 이것은 실천과 리론에서 모두 널리 리용되고있다.

1. 정의

정의 5.3.7 가 모집단 로부터 발취된 표본이라고 할 때 를 이 표본의 번째 순서통계량이라고 부른다. 그 값은 표본관측값을 작은것으로부터 큰것으로 배렬한 후 얻은 번째 관측값이다. 여기서 을 그 표본의 최소순서통계량, 을 그 표본의 최대순서통계량, 을 그 표본의 순서통계량이라고 한다.

아는것처럼 하나의 간단한 우연표본에서 들은 독립동일분포하지만 순서통계량 는 독립이 아니며 분포도 서로 다르다. 다음의 실례를 보자.

실례 5.3.6 모집단 의 분포를 0, 1, 2만을 취하는 리산평등분포로서 그 분포렬이 다음과 같다고 하자.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | 2 |
|  | 1/3 | 1/3 | 1/3 |

이제 용량이 3인 표본을 발취하는데 가능한 경우는 모두 가지로서 그것들을 표 5.3.6의 왼쪽에 표시하고 오른쪽은 해당 순서통계량의 관측값을 표시하였다.

표 5.3.6 실례 5.3.6의 표본값과 그 순서통계량의 값

%FontSize=12
%TeXFontSize=12
\documentclass{article}
\pagestyle{empty}
\begin{document}
\[
\begin{tabular}{ccc|ccc||ccc|ccc}
\hline$x_1$ & $x_2$ & $x_3$ & $x_{(1)}$ & $x_{(2)}$ & $x_{(3)}$ & $x_1$ & $x_2$ & $x_3$ & $x_{(1)}$ & $x_{(2)}$ & $x_{(3)}$ \\
\hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 0 & 0 & 1 & 2 \\
0 & 0 & 1 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 1 & 2 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 2 & 2 & 0 & 2 & 2 \\
1 & 0 & 0 & 0 & 0 & 1 & 2 & 0 & 2 & 0 & 2 & 2 \\
0 & 0 & 2 & 0 & 0 & 2 & 2 & 2 & 0 & 0 & 2 & 2 \\
0 & 2 & 0 & 0 & 0 & 2 & 1 & 1 & 2 & 1 & 1 & 2 \\
2 & 0 & 0 & 0 & 0 & 2 & 1 & 2 & 1 & 1 & 1 & 2 \\
0 & 1 & 1 & 0 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 2 \\
1 & 0 & 1 & 0 & 1 & 1 & 1 & 2 & 2 & 1 & 2 & 2 \\
1 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & 2 & 1 & 2 & 2 \\
0 & 1 & 2 & 0 & 1 & 2 & 2 & 2 & 1 & 1 & 2 & 2 \\
0 & 2 & 1 & 0 & 1 & 2 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 2 & 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
2 & 0 & 1 & 0 & 1 & 2 & & & & & & \\
\hline
\end{tabular}
\]
\end{document}

표본이 우의 매 조의 관측값을 취할 확률이 모두 1/27이므로 의 분포렬은 각각 다음과같다.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | 2 |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | 2 |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | 2 |
|  |  |  |  |

분명히 이 세 순서통계량의 분포가 서로 다르다는것을 알수 있다.

더 나아가서 두개의 순서통계량의 동시분포를 얻을수 있다. 례하면 과 의 동시분포렬은 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | | |
| 0 | 1 | 2 |
| 0 | 7/27 | 9/27 | 3/27 |
| 1 | 0 | 4/27 | 3/27 |
| 2 | 0 | 0 | 1/27 |

이고 로서 둘이 같지 않으므로 은 독립이 아니다.

다음으로 순서통계량의 표본분포에 대해 론의한다. 이것은 일반적으로 련속모집단에서 쓰이므로 모집단 의 분포가 련속인 경우에 대하여서만 서술한다.

2. 한개 순서통계량의 분포

정리 5.3.3모집단 의 밀도함수를 , 분포함수를 , 를 표본이라고 하면 번째순서통계량 의 밀도함수는 다음과 같다.



증명.임의의 실수 에 대하여 순서통계량 가 소구간 에 놓일 사건을 생각하자. 이것은 "표본량이 인 표본에서 에 한개의 관측값(2개이상의 관측값이 그 구간에 놓이는 확률은 의 고차무한소)이 놓이고 개 관측값은 보다 작거나 같고 의 관측값은 보다 크다"는것과 동등하다. 이것을 직관적으로 그림 5.3.5에서 보여주었다.









1

그림 5.3.5 가 취하는 값을 보여주는 그림

표본의 매개 성분이 보다 작거나 같을 확률은 , 구간 에 놓일 확률은 , 보다 클 확률은 이고 개의 성분들을 이러한 세개의 그룹으로 나누면 나누는 방법은 총 가지이다. 그리하여 를 의 분포함수라고 한다면 다항분포를 리용하여 라는것을 알수 있다. 량변을 로 나누고 일 때 이다. 여기서 의 비령구간과 모집단의 비령구간은 같다. 특히 일 때 최소순서통계량 과 최대순서통계량 의 밀도함수들을 얻을수 있다.





실례 5.3.7 모집단밀도함수가 일 때 이 모집단으로부터 용량이 5인 표본을 한개발취하고 을 계산하시오.

풀기. 우선 의 분포를 구해야 한다. 모집단밀도함수로부터 쉽게 다음과 같은 모집단분포함수를 구할수 있다.



식 (5.3.13)으로부터 의 밀도함수를 구한다.



그리하여



실례5.3.8 모집단분포를 , 를 표본이라고 할 때 째 순서통계량의 밀도함수는 이다.

이것은 바로 2장 5절에서 소개한 베타분포 이며 따라서 이다.

3. 여러개의 순서통계량과 그 함수의 분포

임의의 두개 순서통계량의 동시분포에 대하여 론의하여보자. 3개 또는 그 이상의 순서통계량에 대한 분포는 비슷하게 처리할수 있다.

정리 5.3.4 정리 5.3.3의 기호표시밑에서 순서통계량 의 동시분포밀도함수는 다음과 같다.



증명. 증가량  및 에 대하여 사건 은 "용량이 인 표본 가운데서 개의 관측값은 보다 작거나 같고 한개는 구간 에 놓이고 개는 구간 에 놓이며 한개는 구간 에, 나머지 개는 보다 크다"는것으로 설명할수 있다(그림 5.3.6 참고).

따라서 다항분포를 리용하면



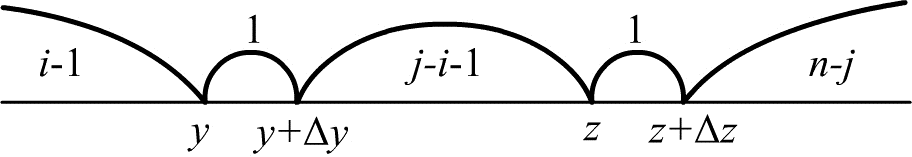


그림 5.3.6 와 가 취하는 값을 보여주는 그림

의 련속성을 고려하면 일 때 이고 따라서



이렇게 정리가 증명된다.

실천문제들에서 순서통계량들의 함수를 리용할수 있다. 례를 들어 **표본극차**라고 불리우는 는 널리 리용되는 통계량으로서 이 통계량의 분포를 유도하는것은 그리 어렵지 않다. 정리 5.3.4와 제3장에서 설명한 우연량함수의 분포를 구하는 방법만 리용하면 된다. 그러나 그 분포는 보통 적분으로 표시되며 몇가지 경우에만 초등함수로 표시할수 있다. 아래에 초등함수로 표시할수 있는 표본극차의 분포에 대한 실례를 주었다.

실례 5.3.9 모집단분포를 , 를 표본이라고 할 때 의 동시밀도함수는 이다. 라고 하면 로부터 를 유도할수 있다. 즉 의 동시분포밀도는 이다. 따라서 의 주변밀도함수는 다음과 같다.

.

이것은 바로 파라메터가 인 베타분포이다.

### 5.3.6 표본분위수와 표본중위수

표본중위수 는 대표적인 통계량의 하나이며 역시 순서통계량의 함수로서 다음과 같이 정의된다.



례를 들어 이면 이고 이면 이다.

더 일반적으로 표본의 분위수 는 다음과 같이 정의할수 있다(:옹근수).



례를 들어 이면 이고 이면 이다.

대부분의 모집단에 대하여 표본의 분위수에 대한 정확한 분포를 얻기는 쉽지 않다. 다행인것은 일 때 그것의 점근분포가 비교적 간단히 표시된다는것이다. 여기서 증명이 없이 다음과 같은 정리를 준다.

정리 5.3.5 모집단밀도함수를 , 를 그의 분위수라고 할 때 가 에서 련속이고 이면 일 때 표본의 분위수 에 대한 점근분포는 다음과 같다.

.

특히 일 때 표본의 중위수에 대하여



실례5.3.10 모집단이 밀도함수가 이고 분포함수가 인 코씨분포라고 하자.

가 이 모집단의 중위수라는것을 쉽게 알수 있다. 즉 이다. 이 모집단으로부터 발취한 표본일 때 표본용량 이 비교적 크면 표본중위수 의 점근분포는 이다.

일반적으로 표본평균은 자료를 요약하는데 도움이 된다. 하지만 표본평균에는 부족점이 있다. 실례로 5개의 수 3, 5, 9, 10, 13가 있을 때 그 평균값은 (3+5+9+10+13)/5=8이다. 만일 실수(례하면 콤퓨터건반으로 입력할 때 3을 련이어 2번 누르는것과 같은 실수)하여 13을 133으로 잘못 입력하였다면 평균값은 (3+5+9+10+133)/5=32로 된다. 이것은 평균값이 고립값의 영향을 비교적 크게 받는다는것을 의미한다. 그러나 상대적으로 중위수는 고립값의 영향을 받지 않는다. 그러므로 자료에 고립값이 있을 때 평균값보다 중위수를 리용하는것이 좋다. 통계에서는 중위수의 이런 간섭에 대한 저항성을 **로바스트성**이 있다고 한다.

### 5.3.7 5개수 요약과 상자선도

순서통계량을 리용하는것들중의 하나는 5개수 요약(five-number summary)과 상자선도(boxplot)이다. 순서표본을 얻으면 최소관측값 , 최대관측값 , 중위수 , 첫 4분위수 , 세번째 4분위수 와 같은 5개의 값을 쉽게 계산할수 있다. 5개수 요약이라는것은 바로 이 5개의 수 를 리용하여 한조의 자료에 대하여 대략적으로 설명하는것을 의미한다.

실례 5.3.11 표 5.3.7은 어떤 회사 160명 판매원들의 월판매량자료에 대한 순서표본이다. 이 자료묶음에 대하여 를 계산할수 있다.

표 5.3.7 어떤 회사 160명 판매원의 월판매량자료에 대한 순서표본

%FontSize=12
%TeXFontSize=12
\documentclass{article}
\pagestyle{empty}
\begin{document}
\[
\begin{array}{rrrrrrrrrr}
\hline 45 & 74 & 76 & 80 & 87 & 91 & 92 & 93 & 95 & 96 \\
98 & 99 & 104 & 106 & 111 & 113 & 117 & 120 & 122 & 122 \\
124 & 126 & 127 & 127 & 129 & 129 & 130 & 131 & 131 & 133 \\
134 & 134 & 135 & 136 & 137 & 137 & 139 & 141 & 141 & 143 \\
145 & 148 & 149 & 149 & 149 & 150 & 150 & 153 & 153 & 153 \\
153 & 154 & 157 & 160 & 160 & 162 & 163 & 163 & 165 & 165 \\
167 & 167 & 168 & 170 & 171 & 172 & 173 & 174 & 175 & 175 \\
176 & 178 & 178 & 178 & 179 & 179 & 179 & 180 & 181 & 181 \\
181 & 182 & 182 & 185 & 185 & 186 & 186 & 187 & 188 & 188 \\
188 & 189 & 189 & 191 & 191 & 191 & 192 & 192 & 194 & 194 \\
194 & 194 & 195 & 196 & 197 & 197 & 198 & 198 & 198 & 199 \\
200 & 201 & 202 & 204 & 204 & 205 & 205 & 206 & 207 & 210 \\
214 & 214 & 215 & 215 & 216 & 217 & 218 & 219 & 219 & 221 \\
221 & 221 & 221 & 221 & 222 & 223 & 223 & 224 & 227 & 227 \\
228 & 229 & 232 & 234 & 234 & 238 & 240 & 242 & 242 & 242 \\
244 & 246 & 253 & 253 & 255 & 258 & 282 & 290 & 314 & 319 \\
\hline
\end{array}
\]
\end{document}

5개수 요약을 도형으로 표시한것을 **상자선도**라고 하는데 그것은 상자와 선분으로 구성된다. 그림 5.3.7은 실례 5.3.11의 표본자료에 대한 상자선도이며 그 작도법은 다음과 같다.

① 상자 하나를 그린다. 상자량쪽에 첫 4분위수와 세번째 4분위수를 표시하며 그 중위수위치에세로줄을 하나 상자 안쪽에 그린다. 이 상자는 표본자료의 50%를 포함하고있다.

② 상자의 좌우 량쪽에 각각 한줄씩 수평선을 그리고 각각 최소값, 최대값까지 연장한다. 매 선분은 표본자료의 25%를 포함한다.

그림 5.3.7월판매량자료의 상자선도

45

144

181

212

319

상자선도는 표본자료분포의 형태를 대체적으로 판단하는데 리용할수 있다. 그림 5.3.8은 세 종류의 상자선도를 보여주는데 각각 왼쪽편기된 분포, 대칭분포, 오른쪽편기된 분포에 대응한다.

그림 5.3.8세 종류의 일반적인 상자선도와 그에 대응하는 분포형태

왼쪽편기

대칭

오른쪽편기

만일 몇조의 자료를 비교해야 한다면 종이 한장에 동시에 그 자료들의 상자선도를 그릴수 있다. 그림 5.3.9는 어느한 공장에서 20일동안 생산한 어떤 제품의 직경자료에 근거하여그린 상자선도이다. 그림으로부터 18일째의 제품에 이상이 생겼다는것을 명백하게 알아볼수 있다.

40

30

50

그림 5.3.9 20일동안 생산된 제품의 직경을 나타내는 상자선도

1

2

3

4

5

6

7

8

9

10

12

14

16

18

20

련습문제 5.3

1. 책에서 우연적으로 10 페지를 검사해본 결과 페지마다 다음과 같은 개수의 오유가 발견되였다.

4, 5, 6, 0, 3, 1, 4, 2, 1, 4.

표본평균, 표본분산, 표본표준편차를 계산하시오.

2. 임의의 상수 에 대하여 다음과 같은 사실이 성립한다는것을 증명하시오.

.

3. 와 는 두조의 표본관측값으로서 라는 관계가 있을 때 표본평균 와 사이의 관계, 표본분산 과 사이의 관계를 구하시오.

4. 라고 할 때 , 라는것을 증명하시오.

5. 동일한 모집단에서 용량이 각각 인 두 표본을 추출하였다. 표본평균값은 각각 이고 표본분산은 각각 이다. 두 표본을 합쳐놓고 구한 평균과 분산을 각각 이라고 한다면 , 이라는것을 증명하시오.

6. 용량이 인 표본 의 표본평균값을 , 표본표준편차를 , 표본극차를 , 표본중위수를 라고 한다면 표본 의 매 관측값에 대하여 변환 을 적용하여 얻은 표본 에 대한 표본평균, 표본편차, 극차와 중위수를 구하시오.

7. 용량이 2인 표본 의 분산은 이라는것을 증명하시오.

8. 표본 를 로부터 발취하였다고 할 때 를 구하시오.

9. 모집단2차모멘트가 존재하고 이 표본일 때 와 ()의 상관곁수가 이라는것을 증명하시오.

10. 를 하나의 표본, 이 표본분산일 때 이라는것을증명하시오.

11. 모집단4차중심모멘트 가 존재할 때 표본분산 에 대하여 다음과 같은 식이 성립한다는것을 증명하시오.

.

여기서 은 모집단 의 분산을 나타낸다.

12. 모집단 의 3차모멘트가 존재하고 이 그 모집단으로부터 발취한 단순우연표본, 를 표본평균, 을 표본분산이라고 할 때 이라는것을 증명하시오. 여기서 이다.

13. 가 동일한 정규모집단 으로부터 독립적으로 발취한 용량이 같은 두 표본의 표본평균값이라고 하자. 이때 표본량 이 얼마여야 두 표본평균의 차가 보다 클 확률이 0.01을 넘지 않겠는가?

14. 체븨쉐브부등식을 리용하여 같은 동전을 몇번 던져야 앞면이 우로 향하는 잦음률이 (0.4,0.6) 사이에 들어갈 확률이 적어도 0.9이겠는가를 구하시오. 또 어떻게 해야 이 회수를 더 정확하게 계산할수 있겠는가, 그리고 그 회수는 얼마인가를 구하시오.

15. 지수모집단 로부터 40개의 표본을 발취하였다. 의 점근분포를 구하시오.

16. 는 평등분포 로부터 발취한 표본이다. 의 점근분포를 구하시오.

17. 는 2점분포 로부터 발취한 표본이다. 의 점근분포를 구하시오.

18. 는 정규모집단 로부터 발취한 표본이다. 의 표분편차를 구하시오.

19. 꼬리절단평균값도 역시 표본자료를 반영하는 대표적인 특징량이다. 그것은 자료의 량끝값들을 버리고 나머지값들을 리용하여 표본평균을 계산한다. 계산공식은 이다. 여기서 은 꼬리절단곁수, 은 순서표본이다.

이제 어떤 대학에서 16명의 대학생들을 조사하여 그들의 평상시 학습정형을 료해하였는데 다음의 자료는 매 대학생들이 매주 TV를 열람하는데 소모한 시간(단위:시간)이다.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 15 | 14 | 12 | 9 | 20 | 4 | 17 | 26 | 15 | 18 | 6 | 10 | 16 | 15 | 5 | 8 |

을 취하고 꼬리절단평균값을 계산하시오.

20. 다음과 같은 그룹표본이 주어졌다.

|  |  |  |
| --- | --- | --- |
| 구간 | 그룹중심값 | 빈도수 |
| (145,155] | 150 | 4 |
| (155,165] | 160 | 8 |
| (165,175] | 170 | 6 |
| (175,185] | 180 | 2 |

이 그룹표본의 표본평균, 표본표준편차, 표본비대칭도와 뾰족도를 구하시오.

21. 네무지의 제품을 검사하였는데 무지당 제품개수와 불합격률은 다음과 같다.

|  |  |  |
| --- | --- | --- |
| 무지번호 | 제품개수 | 불합격률 |
| 1 | 100 | 0.05 |
| 2 | 300 | 0.06 |
| 3 | 250 | 0.04 |
| 4 | 150 | 0.03 |

이 네 무지제품의 총체적인 불합격품률을 구하시오.

22. 모집단이 같은 확률로 1, 2, 3, 4, 5를 취한다고 가정하고 그중에서 용량이 4인 표본을 발취할 때 의 분포를 구하시오.

23. 모집단 가 기하분포에 따른다고 하자. 즉 . 여기서 이다. 이 그 모집단의 표본일 때 의 확률분포를 구하시오.

24. 은 로부터 얻은 표본이다. 다음 확률을 구하시오.

①  ② 

25. 모집단이 웨이블분포라고 하면 그 밀도함수는 다음과 같다.

.

이것으로부터 표본 을 발취할 때 도 여전히 웨이블분포에 따른다는것을 증명하고 그 파라메터를 구하시오.

26. 모집단밀도함수를 이라고 하고 를 그 모집단으로부터 발취한 표본이라고 할 때 표본중위수의 분포를 구하시오.

27. 다음 식을 증명하시오.

. 여기서 .

28. 모집단 의 분포함수 가 련속이고 이 그 모집단으로부터 취한 순서통계량이라고 하자. 일 때 다음을 증명하시오.

① 이며 는 모집단 으로부터 얻은 순서통계량이다.

② .

③ 와 의 공분산행렬은 이다. 여기서 .

29. 모집단 가 에 따르며 그 모집단로부터 다음과 같은 한조의 표본관측값을 얻었다.

①  즉 위치에서 을 구하시오.

② 의 에서의 분포함수값을 구하시오.

30. 밀도함수가 다음과 같을 때 용량이 인 표본의 중위수 에 대한 점근분포를 구하시오.

① .

② .

③ 

④ .

31. 모집단 가 쌍파라메터지수분포에 따르는데 그 분포함수는 다음과 같다.



여기서 이고 는 표본의 순서통계량이다. 이때 에 대하여 이 자유도가 2인 분포에 따른다는것을 증명하시오.

32. 모집단 의 밀도함수가 다음과 같다.



가 이 모집단으로부터 얻어진 용량이 5인 순서통계량일 때 와 가 서로 독립이라는것을 증명하시오.

33. ① 와 를 각각 용량이 인 표본의 최소 및 최대순서통계량이라고 할 때 극차 의 분포함수가 라는것을 증명하시오. 여기서 와 는 각각 모집단의 분포함수와 밀도함수이다.

② ①의 결론을 리용하여 모집단이 지수분포 일 때 표본극차 의 분포를 구하시오.

34. 이 로부터 얻어진 표본이고 이 순서통계량일 때 라고 하면 들이 서로 독립이라는것을 증명하시오.

35. 아래의 자료로부터 상자선도를 작성하시오.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 472 | 425 | 447 | 377 | 341 | 369 | 412 | 419 |
| 400 | 382 | 366 | 425 | 399 | 398 | 423 | 384 |
| 418 | 392 | 372 | 418 | 374 | 385 | 439 | 428 |
| 429 | 428 | 430 | 413 | 405 | 381 | 403 | 479 |
| 381 | 443 | 441 | 433 | 419 | 379 | 386 | 387 |

36. 아래의 자료로부터 상자선도를 작성하시오

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 40.6 | 39.6 | 43.8 | 36.2 | 40.8 | 37.3 | 39.2 | 42.9 |
| 38.6 | 39.6 | 40.0 | 34.7 | 41.7 | 45.4 | 36.9 | 37.8 |
| 44.9 | 45.4 | 37.0 | 35.1 | 36.7 | 41.3 | 38.1 | 37.9 |
| 37.1 | 37.7 | 39.2 | 36.9 | 44.5 | 40.4 | 38.4 | 38.9 |
| 39.9 | 42.2 | 43.5 | 44.8 | 37.7 | 34.7 | 36.3 | 39.7 |
| 42.1 | 41.5 | 40.6 | 38.9 | 42.2 | 40.3 | 35.8 | 39.2 |

## 5.4 3대표본분포

여기서는 많은 통계적추론이 정규분포에 기초하고 또한 표준정규우연량을 기준으로 구성한 세가지 유명한 통계량이 실천응용에서 널리 리용된다는것을 보게 된다. 그것은 이 세가지 통계량이 명백한 배경이 가지고있을뿐 아니라 그 표본분포(sampling distribution)의 밀도함수가 명시적인 표시식이 있기때문이다. 그러므로 이것들은 통계학에서 3대표본분포라고 불리운다.

과 가 표준정규분포로부터 얻은 서로 독립인 표본일 때 이 세개 통계량의 작성과 표본분포를 표 5.4.1에 보여주었다.

표 5.4.1세가지 유명한 통계량의 구조와 그 표본분포

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 통계량의 작성 | 표본분포밀도함수 | 기대값 | | 분산 |
|  |  | |  |  |
|  |  | |  |  |
|  |  | |  |  |

이제 이것들을 하나씩 유도하고 설명하게 된다.

### 5.4.1 분포

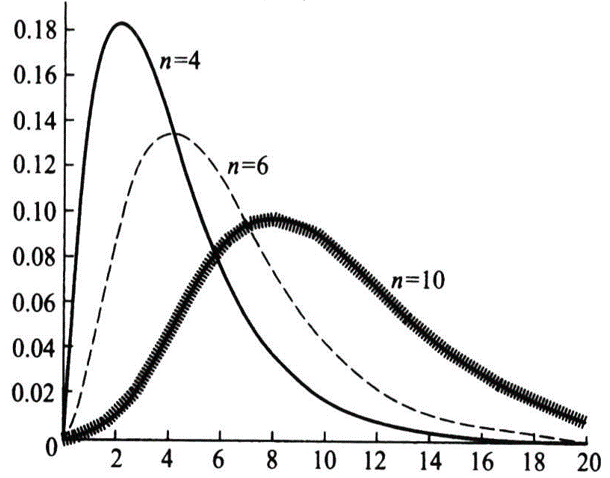
정의 5.4.1 가 독립동일분포하고 그 공동분포가 이라면 의 분포를 자유도가 인 분포라고 부르고 으로 표시한다.

3장에서 이미 언급한것처럼 우연량 에 대하여 이다. 감마분포의 가법성에 의하여 이 성립하게 된다. 이로부터 은 감마분포의 특수경우라는것을 알수 있다. 분포의 밀도함수는 다음과 같다.



이 밀도함수의 그라프는 그림 5.4.1과 같이 부아닌 값만 취하는 치우침분포이다. 그 기대값은 자유도, 분산은 자유도의 2배 즉 이다.

그림 5.4.1 분포의 밀도함수



실례5.4.1 를 정규분포 로부터 얻은 표본이라고 하자. 여기서 는 주어졌다. 이때 통계량



의 분포를 구하시오.

풀기. 라고 하면 는 독립동일분포하는 우연량이고 그 공동분포는 이다. 따라서 정의 5.4.1로부터 이고 의 밀도함수는

.

이것은 감마분포 이다. 식 (5.4.3)과 (5.4.1)사이에 인자 만이 차이난다.

분포가 중요한 리유는 다음과 같은 정리때문이다.

정리 5.4.1 가 정규분포 로부터 얻은 표본이고 그 표본평균과 표본분산이 각각 다음과 같다고 하자.

. 그러면

① 와 은 서로 독립이다.

② .

③ .

증명. 의 동시밀도함수는 다음과 같다.

.

라고 표시할 때 첫행의 모든 원소가 인 차원직교행렬 를 다음과 같이 취하자.

.

라고 하면 이 변환의 야코비행렬식은 1이다. 또한 , 이므로 의 동시밀도함수는 다음과 같다.



따라서 의 성분들은 서로 독립이고 모두 정규분포에 따르며 그 분산은 모두 이지만 평균값은 완전히 같지는 않다. 즉 의 평균값은 0이고 의 평균값은 이므로 결론 ②가 증명된다.

또한 이므로 결론 ①이 증명된다. 이 독립동일분포하고 공동분포가 이므로 . 이렇게 정리가 증명된다.

우연량 가 주어졌을 때 주어진 에 대하여 확률등식 을 만족하는 는 자유도가 인 분포의 분위수이다. 분위수 은 부록 3에서 찾을수 있다. 실례로 에 대하여 부록 3으로부터 을 찾을수 있다.

### 5.4.2 분포

정의 5.4.2 우연량 이 서로 독립이면 의 분포를 자유도가 인 분포라고 부르고 이라고 표시한다. 여기서 은 분자자유도, 은 분모자유도라고 한다.

다음으로 분포의 밀도함수를 두 단계로 유도한다.

첫 단계: 의 밀도함수를 유도한다. 만일 를 각각 의 밀도함수라고 표시한다면 독립우연량들의 상분포의 밀도함수에 대한 공식 (3.3.22)에 의하여 의 밀도함수는 다음과 같다.



변환 를 실시하면 .

마지막 정적분은 감마함수 이고 따라서 

둘째 단계: 의 밀도함수를 유도한다. 의 값이 일 때 에 대하여



이것이 바로 자유도가 인 분포의 밀도함수다. 이 밀도함수의 그라프는 그림 5.4.2와 같이 부아닌 값만 취하는 치우침분포이다.

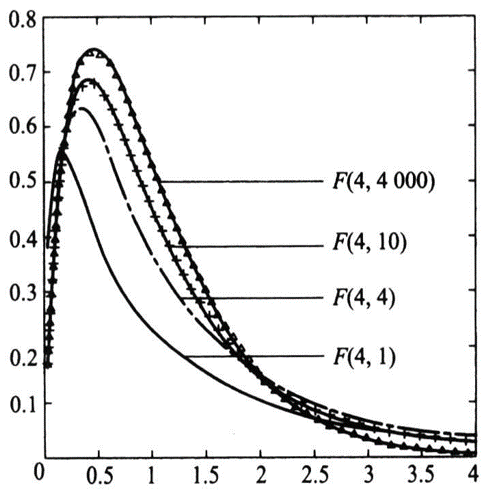


그림 5.4.2 분포의 밀도함수

우연량 가 주어졌을 때 에 대하여 확률등식 을 만족하는 를 자유도가 인 분포의 분위수라고 한다.

분포의 구조로부터 이면 이고 그러므로 주어진 에 대하여



따라서 인데 이것으로부터

.

작은 에 대하여 분위수 는 부록 5에서 찾을수 있으며 분위수 는 식 (5.4.4)를 리용하여 얻을수 있다.

실례 5.4.2 이면 부록 5를 조사하여 를 얻을수 있고 식 (5.4.4)을 리용하면 을 얻을수 있다.

따름 5.4.1 의 표본 와 의 표본 가 서로 독립이고 , 라고 표시하면

.

여기서 이며 특히 이면 이다.

증명. 두 표본이 서로 독립이므로 은 서로 독립이며 정리 5.4.1로부터 , 이라는것을 알수 있다. 따라서 분포의 정의로부터 이다.

### 5.4.3 분포

정의 5.4.3 우연량 과 이 서로 독립이고 이면 의 분포를 자유도가 인 분포라고 하고 으로 표시한다.

아래에서 분포의 밀도함수를 유도한다. 표준정규밀도함수의 대칭성으로부터 과 이 동일한 분포를 가지며 따라서 와 도 동일한 분포를 가진다. 이것은 임의의 정의 실수 에 대하여 이라는것을 의미한다. 따라서 이다.

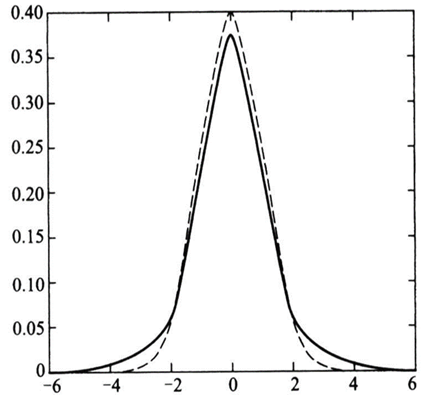
변수의 구조로부터 이라는것을 알수 있으며 웃식의 두변에 에 관한 도함수를 취하면 분포의 밀도함수를 얻을수 있다.



이것이 바로 자유도가 인 분포의 밀도함수이다.

분포밀도함수의 그라프는 세로축대칭(그림 5.4.3)이며 표준정규분포의 밀도함수와 비슷한 모양이지만 봉우리가 표준 정규분포보다 낮고 꼬리부의 확률이 더 크다.

그림 5.4.3 분포와 의 밀도함수







• 자유도가 1인 분포는 표준코씨분포이며 평균이 존재하지 않는다.

• 일 때 분포의 수학적기대값은 0이다.

• 에서 분포의 분산은 이다.

• 자유도가 비교적 클 때(례컨대 ) 분포는 분포로 근사시킬수 있다.

분포는 통계학에서 중요한 분포의 하나로서 표준정규분포와의 미세한 차이는 잉글랜드의 통계학자 고세트(Gosset)에 의해 발견되였다. 1908년까지 통계학은 처음에 사회통계 특히 인구통계에 리용되였으며 그 다음 생물통계문제에 주로 리용되였다. 이러한 문제들의 특징은 자료가 일반적으로 대량이고 자연적으로 수집되며 리용하는 방법은 대부분 중심극한정리에 근거하고 항상 정규분포로 귀착된다는것이다. 당시 통계학계의 권위자였던 피어슨은 정규분포가 인간에게 주어진 유일한 정확한 분포라고 주장하였다. 그러나 20세기에 들어서면서 사람들에 의하여 설계되는 실험조건하에서 얻은 자료의 통계분석문제가 날로 사람들의 주목을 받게 되였다. 이때 자료의 량이 일반적으로 크지 않으므로 중심극한정리에만 의존하는 전통적인 방법이 의문시되게 되였다. 이 방향의 선구자들은 고세트와 피셔이다.

옥스포드대학에서 수학과 화학을 공부한 고세트는 1899년부터 한 양주장에서 화학기술자로 일하며 실험과 자료분석을 했다. 그가 접촉하는 표본의 크기는 4~5개밖에 되지 않았다. 많은 량의 실험자료를 축적하여 고세트는 의 분포가 전통적으로 공인되던 분포와 다르며 특히 꼬리부의 확률에서 큰 차이가 있다는것을 발견하였다. 표 5.4.2는 와 자유도가 4인 분포의 꼬리부확률을 보여준다.

표 5.4.2 와 의 꼬리부확률 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | 0.0455 | 0.0124 | 0.0027 | 0.000465 |
|  | 0.1161 | 0.0668 | 0.0399 | 0.0249 |

고세트는 다른 종류의 분포족이 있을것인가를 의문시하기 시작했지만 그의 통계학적기초로는 이 문제를 해결할수 없었다. 그리하여 1906~1907년 피어슨에게서 통계학을 배운 다음 소량자료에 대한 통계적분석문제를 연구했으며 1908년 잡지 Biometrics에 Student라는 필명으로 론문을 발표하였다. 그 론문에서 그는 다음과 같은 결과를 제시하였다. 이 로부터 얻은 독립동일분포하는 표본이고 이 미지일 때 는 자유도가 인 분포에 따른다. 분포의 발견은 통계학에서 획기적인 의의를 가진다. 정규분포가 지배하던 국면을 타파하고 소표본통계추론의 새로운 시대를 열었으며 이때부터 소표본통계분석이 많은 통계학연구자들의 관심을 끌었다. 사실 고세트의 증명에는 결함이 있었는데 피셔가 이것을 알고 1922년에 완벽하게 증명하였으며 분포의 분위수표를 작성하였다.

따름 5.4.2 이 로부터 얻은 표본이고 이 각각 그 표본의 평균과 분산일 때



증명. 정리 5.4.1의 ②로부터 을 알수 있다. 식 (5.4.6)의 왼변을 로 고쳐쓸수 있다.

분자는 표준정규변수이고 분모의 두제곱뿌리안의 항은 자유도가 인 변수를 자유도로 나눈것이며 분자와 분모가 서로 독립이기때문에 분포의 정의에 따라 라는것을 알수 있다.

따름 5.4.3 따름 5.4.1의 기호를 그대로 리용할 때 이라고 하고 라고 표시하면

.

증명. 이고 가 독립이므로 이고 따라서 .

정리 5.4.1로부터 이고 그것들이 서로 독립이므로 가법성으로부터 .

와 이 서로 독립이므로 분포의 정의에 의하여 식 (5.4.7)을 얻을수 있다.

우연량 일 때 주어진 에 대하여 확률식 를 만족시키는 는 자유도가 인 분포의 분위수이다. 분위수 는 부록 4에서 찾을수 있다. 실례로 라고 하면 부록 4에서 를 찾을수 있다.

분포의 밀도함수는 0대칭이므로 그 분위수들사이에는 다음과 같은 관계가 있다.

.

례를 들어 이다.

이 비교적 클 때(례를 들어 ) 분포의 분위수는 표준정규분포의 분위수로 교체될수 있다. 실례로 이다.

련습문제 5.4

1. 모집단 에서 용량이 인 표본을 발취할 때 그 표본평균값이 (5.6, 9.6)에 놓일 확률이 0.95보다 작지 않자면 이 적어도 얼마여야 하는가?

2. 이 로부터 얻은 표본일 때 이 얼마여야 이 성립하겠는가?

3. 정규모집단 로부터 서로 독립인 용량이 각각 15, 20인 두개의 표본을 발취하고 표본평균값을 각각 라고 할 때 을 구하시오.

4. 정규모집단 로부터 용량이 20인 표본을 발취할 때 을 구하시오.

5. 이 로부터 얻어진 표본이고 일 때 을 구하시오.

6. 이 로부터 얻어진 표본일 때 임의의 에 대하여 이 성립하는 최소의 상수 를 구하시오.

7. 우연량 에 대하여 를 증명하시오.

8. 우연량 에 대하여 가 베타분포에 따른다는것을 증명하고 그 파라메터를 구하시오.

9. 의 표본 에 대하여 의 분포를 구하시오.

10. 의 표본 에 대하여 상수 를 결정하여 이 만족되게 하시오.

11. 의 표본 와 의 표본 가 주어졌을 때 임의의 동시에 0이 아닌 상수 에 대하여 이 성립한다는것을 증명하시오. 여기서 이다.

12. 이 의 표본이고 일 때 이 분포에 따르게 되는 상수 를 구하고 그 분포의 자유도를 지적하시오.

13. 분산이 같은 두개의 독립적인 정규모집단으로부터 용량이 각각 15, 20인 표본을 발취하고 그 표본분산이 각각 이라고 할 때 을 구하시오.

14. 를 모집단 의 표본으로 할 때 의 분포를 구하시오.

15. 를 의 표본, 를 각각 표본평균, 표본분산이라고 할 때 이 성립하는 를 구하시오.

16. 모집단 가 에 따르고 그 모집단으로부터 발취한 표본 의 표본평균을 로 표시하자. 이때 통계량 의 수학적기대값을 구하시오.

17. 우연량 라면 일 때 라는것을 증명하고 를 구하시오.

18. 우연량 라면 일 때 라는것을 증명하고 를 구하시오.

19. 는 어떤 련속모집단의 표본이고 그 모집단의 분포함수 는 련속이며 엄격히 증가하는 함수이다. 이때 통계량 가 에 따른다는것을 증명하시오.

20. 는 의 표본이고 는 표본분산이다. 이때 를 만족하는 최소의 을 구하시오.

21. 가 독립동일분포하며 에 따른다. 이고 라고 표시하면 와 분포사이의 관계를 찾고 의 밀도함수를 구하시오. (풀기방향: 직교변환 를 리용하시오.)

22. 가 서로 독립이며 이라고 하자. 일 때 들이 서로 독립이고 이라는것을 증명하시오. (풀기방향: 라고 놓고 변환 을 리용하시오.)

## 5.5 충분통계량

### 5.5.1 충분성의 개념

통계량을 구성하는것은 표본을 가공하여 나쁜것은 버리고 좋은것만 취하며 표본을 간소화하여 통계적추론에 편리하도록 하는것을 말한다. 하지만 가공하는 과정에 표본속의 관심있는 정보가 잃어지지 않는가? 만일 어떤 통계량이 표본속의 관심있는 모든 정보를 포함한다면 이 통계량은 앞으로의 통계적추론에 쓸모가 있을것이다. 이것이 바로 충분통계량의 직관적인 의미로서 피셔가 1922년에 정식으로 제시하였다. 그것은 천문학자 에딩턴(Eddington)과 표준편차를 추정하는 문제를 가지고 론의하던 과정에 착상되였다. 가 의 독립동일분포하는 표본이라고 할 때 를 추정하려면 표준편차 를 리용해야 한다고 피셔는 주장했고 에딩턴은 다음과 같은 평균절대편차를 리용해야 한다고 주장하였다.



피셔는 에 표본속의 에 대한 모든 정보가 포함되여있는 반면에 는 그렇지 않다고 보았다. 다시 말하여 는 충분통계량이지만 는 아니기때문에 를 리용해야 한다고 하였다.

충분통계량에 대한 엄밀한 정의를 내리기에 앞서 한가지 실례를 들어보자.

실례 5.5.1 한 선수의 사격명중률 를 알아보기 위하여 그 선수에 대한 관측을 하였다. 10번의 사격 결과 3번째, 6번째를 제외한 나머지 8번을 명중시켰다. 이런 관측결과에는 두가지 정보가 들어있다.

① 10번 사격하여 8번 명중시킨다.

② 두번의 실패는 각각 3번째, 6번째 사격에서 나타난다.

두번째 정보는 그 선수의 명중률을 알아내는데 도움이 되지 않는다. 첫번째, 두번째 실패하고 나머지는 명중했다고 할 때 비록 표본관측값은 다르지만 제공하는 명중률 에 대한 정보는 모두 같다. 그러므로 대부분의 실천문제들에서 시행의 순서정보는 일반적으로 모집단이나 그 파라메터를 료해하는데는 그다지 중요하지 않다.

일반적으로 이 선수를 번 관측하여 를 얻는다고 하자. 매 는 0 또는 1을 취하는데 명중이면 1, 실패하면 0이다. 라고 하면 는 관측된 명중회수이다. 이러한 경우 를 기록하고 리용해도 명중률과 관련된 어떠한 정보도 손실되지 않는다. 통계적으로는 이렇게 표본가공과정에 정보를 손실하지 않는것을 "충분성"이라고 한다.

우에서 직관적으로 "충분성"에 관한 설명을 했다면 이제부터는 확률적측면에서 분석을 해보겠다. 알다싶이 표본 는 표본동시분포 를 가지고 이 분포는 표본의 에 관한 모든 정보를 포함하고있다. 통계량 도 표본분포(sampling distribution) 를 가진다. 이제 통계량 로 원래의 표본 를 대신하면서도 와 관련된 아무런 정보도 잃지 않으려면 표본분포 가 처럼 와 관련된 모든 정보를 포함하여야 한다. 다시 말하여 통계량 가 값을 가진다는 조건밑에서 표본 의 조건부분포를 생각할 때 다음과 같은 두가지 경우가 있을수 있다.

• 가 파라메터 에 의존하고 이 조건부분포가 여전히 의 정보를 포함한다.

• 가 파라메터 에 의존하지 않고 이 조건부분포가 의 정보를 포함하지 않는다.

두번째 경우는 조건 의 출현이 표본동시분포 로부터 조건부분포 로 되게 하고 와 관련된 정보가 없어진다. 이것은 와 관련된 정보가 모두 통계량 속에 포함되여있다는것을 의미한다. 의 값을 알면 표본속에 있는 의 모든 정보를 알게 된다. 이것이 바로 통계량이 충분성을 가진다는 의미이다.

실례 5.5.2 모집단은 2점분포 이고 은 표본이다. 라고 할 때 의 값이 주어지면 임의의 한개 조 에 대하여 다음과 같은 식이 성립한다.



이 조건부분포는 과 무관계하다. 라고 하면 가 앞의 두 표본관측값만 리용하였기때문에 분명히 표본에 있는 에 대한 모든 정보를 포함하지 않는다. 가 주어진 다음 임의의 조 에 대하여



이 분포는 미지파라메타 에 의존하는데 이것은 표본의 에 대한 정보가 통계량 에 완전히 포함되여있지 않다는것을 의미한다.

우의 례로부터 통계량이 표본속의 가치있는 정보를 잃지 않는다는것을 나타내려면 조건부분포가 미지파라메터와 무관계하다는것을 보여주면 된다는것을 직관적으로 알수 있다. 이로부터 충분통계량의 정의를 줄수 있다.

정의 5.5.1 를 분포함수가 인 어떤 모집단에서 발취한 표본이라고 하자. 통계량 을 의 값이 결정된 다음 의 조건부분포가 와 무관계할 때 **충분통계량**이라고 부른다.

정의 5.5.1의 미지파라메터 는 1차원일수도 있고 다차원일수도 있으며 조건부분포는 조건부분포렬이나 조건부밀도함수로 나타낼수 있다.

실례 5.5.3 를 에서 얻은 표본이고 라고 하면 이다. 변환 을 실시할 때 그 야코비행렬식은 이고 의 동시밀도함수는 다음과 같다.



따라서 조건부밀도함수 는 다음과 같다.



이 분포는 와 무관계하며 이것은 가 의 충분통계량이라는것을 의미한다.

### 5.5.2 인자분해정리

통계학에는 **충분성준칙**이라고 하는 다음과 같은 하나의 기본준칙이 있다. 즉 충분통계량이 존재하는 경우에는 모든 통계적 추론이 충분통계량에 기초하여 진행될수 있으며 이것은 통계적추론과정을 간소화할수 있다는것이다. 그러나 정의 5.5.1로부터 직접 통계량이 충분한가를 검증하는것은 일반적으로 조건부분포의 계산이 쉽지 않기때문에 어렵다. 다행히도 통계학자 네이만(Neyman)이 제시한 인자분해정리를 리용하여 간단히 통계량이 충분한가를 검증할수 있다. 편리상 두가지 분포류형에 통용되는 개념인 **확률함수**를 도입하겠다. 우연량 의 확률함수 는 련속인 경우에는의 확률밀도함수, 리산인 경우에는 확률분포렬을 나타낸다.

정리 5.5.1 모집단확률함수가 이고 그 모집단으로부터 표본 이 발취되였다고 하자. 그러면 (는 1차원일수도 있고 다차원일수도 있다)이 충분통계량이기 위한 필요충분조건은 두개의 함수 가 존재하여 임의의 와 관측값 에 대하여



이 성립하는것이다. 여기서 는 통계량 의 값을 취하는 방식으로 표본에 의존한다.

증명. 여기서는 일반적인 결과에 대한 증명은 하지 않고 리산우연량에 대한 증명을 한다. 이때 이다.

먼저 필요성을 증명하자. 가 충분통계량이라고 하면 일 때 와 는 무관계하다. 이것을  또는 로 표시하고 라고 하면 일 때 이다. 그리하여



여기서 이고 는 와 무관계하다. 이렇게 필요성이 증명되였다.

계속하여 충분성을 증명하자.



이므로 임의의 와 에 대하여 이 만족되고



이 분포는 와 무관계하며 충분성이 증명된다.

실례 5.5.4 는 모집단 에서 발취한 표본으로서 모집단밀도함수는 다음과 같다.



이때 표본동시밀도함수는 다음과 같다.



, 이므로 웃식을 다음과 같이 고쳐쓸수 있다.



로 취하고 라고 하면 인자분해정리로부터 은 의 충분통계량이라는것을 알수 있다.

실례 5.5.5 는 모집단 에서 발취한 표본이고 는 미지이다. 이때 동시밀도함수는



를 취하고 라고 놓으면 인자분해정리로부터 이 충분통계량이다. 더 나가서 이 통계량과 이 1:1대응한다는것을 알수 있다. 그러므로 정규모집단에서 널리 리용되는 은 의 충분통계량이다. 사실상 이 실례로부터 다음과 같은 분해를 쉽게 얻을수 있다.



다음과 같은 더 일반적인 명제가 있다.

정리 5.5.2 통계량 가 충분통계량이라고 하자. 어떤 함수 가 존재하여 를 로 표시할수 있으면 통계량 도 충분통계량이다.

증명. 는 충분통계량이기때문에 인자분해정리에 의해 다음과 같이 분해된다.



이므로 웃식은 다음과 같이 변환할수 있다.

.

여기서 이다. 이것은 통계량 도 충분통계량이라는것을 의미한다.

련습문제 5.5

1. 를 기하분포 로부터 발취한 표본이라고 할 때 이 충분통계량이라는것을 증명하시오.

2. 를 뽜쏭분포 의 한개 표본이라고 할 때 이 충분통계량이라는것을 증명하시오.

3. 모집단이 다음과 같은 리산분포이고 은 그 모집단으로부터 얻은 표본이다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | … |  |
|  |  |  | … |  |

① 순서통계량 이 충분통계량이라는것을 증명하시오.

② 가운데서 와 같은 개수를 로 표시하면 가 충분통계량이라는것을 증명하시오.

4. 를 정규분포 의 표본이라고 하면 이 충분통계량이라는것을 증명하시오.

5. 가 의 표본이라고 할 때 충분통계량 한개를 구하시오.

6. 를 웨이블분포 의 표본(은 기지)이라고 할 때 충분통계량 한개를 구하시오.

7. 를 파레토(Pareto)분포 의 표본(은 기지)이라고 할 때 충분통계량 한개를 구하시오.

8. 를 라플라스(Laplace)분포 의 표본이라고 할 때 충분통계량 한개를 구하시오.

9. 가 독립동일분포하고 이 다음과 같은 분포에 따를 때 대응한 충분통계량을 구하시오.

① 부의 2항분포: 이고 는 기지.

② 리산평등분포: 이고 은 미지.

③ 로그정규분포: .

④ 레일리(Rayleigh)분포: .

10. 가 정규분포 의 표본이라고 하자.

① 가 주어졌을 때 의 충분통계량을 하나 구하시오.

② 가 주어졌을 때 의 충분통계량을 하나 구하시오.

11. 가 평등분포 의 표본일 때 충분통계량 한개를 구하시오.

12. 가 평등분포 의 표본일 때 충분통계량을 구하시오.

13. 가 감마분포족 의 표본일 때 의 충분통계량을 구하시오.

14. 가 베타분포족 의 표본일 때 의 충분통계량을 구하시오.

15. 가 분포족 에서 발취한 단순우연표본일 때 이 충분통계량이라는것을 증명하시오.

16. 은 정규모집단 의 표본, 은 다른 정규모집단 의 표본이며 이 두 표본은 서로 독립이다. 이때 의 충분통계량을 구하시오.

17. 은 정규분포족 의 2차원표본이다. 이때 의 충분통계량을 구하시오.

18. 2차원우연량 이 평균값벡토르가 령벡토르이고 공분산행렬이 인 2차원정규분포에 따른다. 2차원통계량 가 2차원정규분포족의 충분통계량이라는것을 증명하시오.

19. 가 쌍파라메터지수분포 의 표본이라고 할 때 가 충분통계량이라는것을 증명하시오.

20. 우연량 에서 모든 들이 서로 독립이고 는 주어진 상수라고 할 때 이 충분통계량이라는것을 증명하시오.

# 제6장 파라메터추정

앞장에서는 주로 몇가지 대표적인 통계량의 표본분포와 충분통계량을 취급하였다. 통계량을 받아들이는 목적은 관심있는 문제에 대해 통계적추론을 하기 위해서이고 실천에서 우리가 관심을 가지는 문제는 대부분 분포족들의 미지파라메터와 관련된것이다. 이 장에서부터 파라메터의 추정과 검정에 대해 론의하며 6장에서는 파라메터추정문제를 취급한다.

여기서 파라메터는 다음과 같은 세가지 류형의 미지파라메터를 가리킨다.

• 분포에 포함된 미지파라메터 . 례를 들어 2점분포 에서의 확률 , 정규분포 의 .

• 분포에 포함된 미지파라메터 의 함수. 례를 들면 정규분포 에 따르는 우연량 가 어떤 주어진 값 를 초과하지 않는 확률 은 미지파라메터 의 함수라는것, 단위제품당 결함수 가 뽜쏭분포 에 따르는데 결함이 없는 단위제품합격확률 는 미지파라메터 의 함수라는것이다.

• 분포의 여러가지 특징수도 모두 미지파라메터이다. 례를 들면 평균 , 분산 , 분포중위수 등을 들수 있다.

일반적으로 로 파라메터를 표시하는데 파라메터 의 모든 가능한 값으로 구성되는 모임을 파라메터공간이라고 하며 로 표시한다. 파라메터추정문제는 표본에 근거하여 적당한 통계량을 구성하고 여러가지 미지파라메터들을 추정하는 문제이다.

파라메터추정에는 점추정과 구간추정의 두가지 형태가 있다. 여기서는 점추정으로부터 시작한다.

## 6.1 점추정의 개념과 불편성

### 6.1.1 점추정과 불편성

정의 6.1.1 를 모집단의 한개 표본이라고 하고 미지파라메터 를 추정하는데 리용되는 통계량 을 의 **추정량**이라고 한다. 또는 의 **점추정**, 간단히 **추정**이라고 한다.

여기서 어떻게 를 구성하는가는 명백한 규정이 없고 다만 그것이 일정한 합리성을 만족시키기만 하면 된다. 가장 흔히 볼수 있는 합리성의 요구는 불편성이다.

정의 6.1.2 는 의 추정이며 의 파라메터공간을 라고 하자. 만일 임의의 에 대하여



이면 를 의 **불편추정**(unbiased estimation)이라고 하고 그렇지 않으면 **편향추정**이라고 한다.

불편성요구를 으로 고쳐쓸수 있다. 이것은 불편추정에는 계통오차(bias)가 없다는것을 의미한다. 우리가 으로 을 추정할 때 표본의 우연성으로 하여 와 사이에는 반드시 편차가 생긴다. 이런 편차는 어떤 때는 정수, 어떤 때는 부수이며 클수도 있고 작을수도 있다. 불편성은 이런 편차를 평균하면 0이 된다는것을 의미한다. 이것이 바로 불편추정의 의미이다. 만일 추정이 불편성을 만족시키지 않으면 몇번을 리용하든 평균값은 파라메터의 진짜값과 일정한 거리를 가지는데 이 거리가 바로 계통오차 이다.

실례6.1.1 모든 모집단에 대하여 표본평균은 모집단평균의 불편추정이다. 모집단차모멘트가 존재할 때 표본차원점모멘트 는 모집단차원점모멘트 에 대한 불편추정이다. 그러나  차중심모멘트의 경우는 그렇지 않다. 례를 들어 표본분산 는 모집단분산 의 불편추정이 아니다. 그것은 정리 5.3.2로부터 이기때문이다.

이에 대하여 다음과 같은 두가지 설명이 있다.

① 표본크기가 무한대로 갈 때 인데 이것을 이 의 **점근불편추정**이라고 한다. 이것은 표본용량이 비교적 클 때 을 의 불편추정으로 볼수 있다는것을 의미한다.

② 을 다음과 같이 수정한다.



그러면 은 무집단분산의 불편추정이다. 이런 간단한 수정방안이 일부 경우에 자주 리용된다. 식 (6.1.2)에 정의된 을 표본분산이라고 하는데 일반적으로 보다 더 자주 리용된다. 이것은 일 때 이므로 로 를 추정하는것은 작아지는 경향이 있으며 특히 소표본의 경우 을 리용하여 를 추정한다.

불편성은 불변성을 만족하지 않는다. 즉 이 의 불편추정일 때 가 의 선형함수가 아닌 이상 는 일반적으로 의 불편추정이 아니다. 례를 들어 는 의 불편추정이지만 는 의 불편추정이 아니다. 아래에서 정규분포를 실례들어 설명한다.

실례6.1.2모집단이 이고 는 표본이다. 분명히 는 의 불편추정이다. 이제 가 의 불편추정인가를 알아보겠다. 정리 5.4.1로부터 이고 그 밀도함수는 이다.

따라서



그러므로 .

이로부터 는 의 불편추정이 아니며 수정방식을 리용하여 가 의 불편추정이라는것을 알수 있다. 여기서 는 수정곁수이다. 표 6.1.1에 부분적인 값들을 제시하였다. 일 때 이라는것을 증명할수 있다. 이것은 가 의 점근불편추정이며 따라서 표본용량이 클 때에는 수정되지 않은 도 좋은 추정이라는것을 의미한다.

표 6.1.1 정규표준편차의 수정곁수표

%FontSize=14
%TeXFontSize=14
\documentclass{article}
\pagestyle{empty}
\begin{document}
\[
\begin{tabular}{cc||cc||cc||cc||cc}
\hline$n$ & $c_n$ & $n$ & $c_n$ & $n$ & $c_n$ & $n$ & $c_n$ & $n$ & $c_n$ \\
\hline & & 7 & $1.0424$ & 13 & $1.0210$ & 19 & $1.0140$ & 25 & $1.0105$ \\
2 & $1.2533$ & 8 & $1.0362$ & 14 & $1.0194$ & 20 & $1.0132$ & 26 & $1.0100$ \\
3 & $1.1284$ & 9 & $1.0317$ & 15 & $1.0180$ & 21 & $1.0126$ & 27 & $1.0097$ \\
4 & $1.0854$ & 10 & $1.0281$ & 16 & $1.0168$ & 22 & $1.0120$ & 28 & $1.0093$ \\
5 & $1.0638$ & 11 & $1.0253$ & 17 & $1.0157$ & 23 & $1.0114$ & 29 & $1.0090$ \\
6 & $1.0509$ & 12 & $1.0230$ & 18 & $1.0148$ & 24 & $1.0109$ & 30 & $1.0087$ \\
\hline
\end{tabular}
\]
\end{document}

편차가 큰것은 흔히 예측이 나쁘다는것으로 인정되는데 학자들은 편차를 줄이기 위한 여러가지 방법을 제기하였다. 다음의 잭크나이흐(Jackknife, 刀切)법은 1949년과 1956년에 퀴누유(Quenouille)가 제안하였고 1958년에 튜키(Tukey)가 이름을 붙였다.

실례6.1.3(Jackknife) 가 표본 에 기초한 파라메터 의 추정량이고 를 만족한다고 하자. 가 표본에서 를 빼버린 벡토르를 표시한다면 의 잭크나이흐통계량(절삭통계량)은 다음과 같이 정의된다.



이 통계량이 이라는 성질을 가지며 그 분산이 증가하지 않는다는것을 증명할수 있다.

례를 들어 모집단이 , 이 그 표본이라고 하고 라고 놓으면 는 의 추정으로서 이 성립한다.

다음은 잭크나이흐법을 적용해보겠다. 이므로

.

을 검증할수 있다. 잭크나이흐법에 대한 상세한 설명은 참고문헌 [20]을 열람할수 있다.

모든 파라메터가 불편추정을 가지는것은 아니다. 파라메터에 불편추정이 있을 때 파라메터를 **추정할수 있다**고 하고 그렇지 않을 때 **추정할수 없다**고 한다.다음은 추정할수 없는 실례이다.

실례6.1.4 모집단이 2점분포 이고 은 표본이다. 이제 파라메터 가 추정할수 없다는것을 설명하여보자.

우선 은 충분통계량이며 이다. 만일 가 의 불편추정이면  또는 이다.

이것은 의 차방정식으로서 최대로 개의 실수뿌리를 가지기때문에 모든 에 대하여 성립될수 없으므로 파라메터 는 추정할수 없다.

그리고 만일 어떤 가 의 불편추정이면 라고 할 때 재기대값공식으로부터 을 알수 있다. 이것은 가 의 불편추정이라는것을 의미하는데 앞에서 설명한것처럼 이것은 불가능하다.

### 6.1.2 유효성

파라메터를 추정할수 있을 때 불편추정은 많을수 있는데 그가운데서 어떻게 선택할수 있는가? 직관적인 방법은 그 추정이 파라메터의 진짜 값주위에서 적게 파동할수록 더 좋은것이고 파동의 크기는 분산으로 나타낼수 있기때문에 불편추정량의 분산크기를 선택의 기준으로 하는것이다. 이것이 바로 유효성(efficiency)이다.

정의 6.1.3 가 의 불편추정들이고 임의의 에 대하여 이고 적어도 하나의 에 대하여서는 웃식의 안같기부호가 엄격히 성립할 때 가 보다 유효하다고 한다.

실례6.1.5 이 모집단평균이 , 분산이 인 모집단으로부터 발취한 표본일 때 , 는 모두 의 불편추정량인데 이다. 분명히 이기만 하면 은 보다 유효하다. 이것은 일부 자료만 리용하기보다 전체 자료의 평균을 리용하여 모집단평균을 추정하는것이 더 유효하다는것을 보여준다.

실례6.1.6 이 평등분포하는 모집단 으로부터 얻은 표본일 때 일반적으로 최대관측값 으로 를 추정(실례 6.3.5를 참고)한다. 이므로 은 불편추정량이 아니지만 점근불편추정량이며 수정하면 의 불편추정량 을 얻을수 있다. 그리고 

한편 모집단평균이 이므로 표본평균을 리용하여 모집단평균을 추정(6.2 참고)할수 있으며 의 다른 불편추정량 를 얻을수 있고 이때 

두개를 비교하면 일 때 이 보다 더 유효하다는것을 알수 있다.

련습문제 6.1

1. 는 어떤 모집단에서 발취한 용량이 3인 표본이다. 다음의 통계량들이 모두 의 불편추정량이라는것을 증명하고 분산이 존재할 때 어느 추정량이 가장 유효하지 않은가를 지적하시오.

① .

② .

③ .

2. 가 로부터 얻은 표본이며 가 의 불편추정이라는것이 알려졌을 때 가 의 불편추정인가를 알아보시오.

3. 가 파라메터 의 불편추정이고 일 때 는 의 불편추정이 아니라는것을 증명하시오.

4. 이 모집단 으로부터 얻은 표본이다. 이 의 불편추정이 되는 상수 를 결정하시오.

5. 는 모집단 의 단순우연표본이다. 이때 와 이 모두 의 불편추정이라는것을 증명하고 어느것이 더 유효한가를 판단해보시오.

6. 가 평등분포 에 따른다면 이 모두 의 불편추정이라는것을 증명하고 어느것이 더 유효한가를 판단해보시오.

7. 평균이 , 분산이 인 모집단으로부터 용량이 각각 인 두 독립표본을 발취하였는데 그 표본평균은 각각 이다. 임의의 상수 에 대하여 는 모두 의 불편추정이라는것을 증명하고 가 최소로 되는 를 구하시오.

8. 를 평균이 , 분산이 인 모집단 로부터 얻은 표본이고 이 의 임의의 선형불편추정량일 때 와 의 상관곁수가 라는것을 증명하시오.

9. 대의 계기가 설치되여있는데 째 계기의 측정표준편차가 라는것을 알고있다. 이런 측정기구들을 리용하여 독립적으로 어떤 물리적량 를 각각 한번씩 관측하여 를 얻었다. 측정계기들에 모두 계통오차가 없다고 하면 가 어떤 값을 가져야 가 의 불편추정량으로 되고 그 분산을 최소화할수 있는가?

10. 가 의 표본이라고 할 때 는 불편추정량을 가지지 않는다는것을 증명하시오. (풀기방향: 가 에서 미분불가능하다는것을 리용하시오.)

11. 이 모집단 으로부터 얻은 표본이다. 표준편차 의 추정량으로서 다음과 같은 통계량을 생각한다.



과 이 모두 의 불편추정으로 되는 상수 을 구하시오.

## 6.2 모멘트추정과 일치성

### 6.2.1 치환원리와 모멘트추정

1900년에 피어슨이 치환원리를 제기하였는데 후에 이 방법을 모멘트법이라고 하였다. 치환원리는 다음의 두가지를 의미한다.

• 표본(원점 또는 중심)모멘트에 의하여 모집단모멘트를 교체한다.

• 모집단모멘트의 함수를 표본모멘트의 함수로 교체한다.

이 교체원리에 따라 모집단의 분포형식이 주어지지 않은 경우에도 여러가지 파라메터에 대한 추정값을 얻을수 있다. 실례로

• 표본평균 로 모집단평균 를 추정한다.

• 표본분산 로 모집단분산 를 추정한다.

• 사건 가 발생한 잦음률을 리용하여 의 발생확률을 추정한다.

• 표본의 분위수로 모집단분위수를 추정한다. 특히 표본중위수로 모집단중위수를 추정한다.

실례 6.2.1 어떤 형의 차량 20대에 대하여 연유 5리터당 주행거리(단위:km)를 기록한 자료가 다음과 같다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 29.8 | 27.6 | 28.3 | 27.9 | 30.1 | 28.7 | 29.9 | 28.0 | 27.9 |
| 28.4 | 27.2 | 29.5 | 28.5 | 28.0 | 30.0 | 29.1 | 29.8 | 29.6 |

이것은 용량이 20인 표본관측값이며 대응하는 모집단은 이 형의 차량 연유 5리터당 주행거리로서 그 분포형태는 알려지지 않았지만 모멘트법으로 평균, 분산, 중위수 등을 추정할수 있다. 이 례에 대하여 계산하면 이다.

이로부터 모집단의 평균, 분산, 중위수의 추정값이 각각 28.695, 0.9668, 28.6이라고 볼수 있다.

모멘트법에 의한 추정의 통계적원리(치환원리)는 아주 간단하고 명백하기때문에 많은 사람들이 모두 인정하고 널리 리용하고있다. 그 본질은 경험적분포함수로 모집단분포를 교체하는것인데 그 리론적기초는 글리문코의 정리이다.

### 6.2.2 확률함수가 기지일 때 미지파라메터의 모멘트추정

모집단의 확률함수 가 주어지고 는 미지의 파라메터 혹은 파라메터벡토르, 는 표본이라고 하자. 모집단의 차원점모멘트 가 존재한다면 에 대하여 가 모두 존재하며 또 가 의 함수 로 표시할수 있다면 모든 에 대한 모멘트추정 을 얻을수 있다. 여기서 은 차까지의 원점모멘트로서 이다. 더 나아가서 만일 의 함수 를 추정하려면 직접 의 모멘트추정 을 얻을수 있다. 일 때 일반적으로 표본평균값으로부터 출발하여 미지파라메터를 추정할수 있으며 라면 미지파라메터들을 1차, 2차 원점모멘트(또는 중심모멘트)를 리용하여 추정할수 있다.

실례6.2.2모집단이 밀도함수가 인 지수분포이고 은 표본이다. 일 때 이므로 이고 의 모멘트추정은 이다.

또한 이므로 그 거꿀함수는 이고 치환원리로부터 의 모멘트추정을 로 취할수도 있다. 여기서 는 표본표준편차이다. 이것은 모멘트추정이 유일하지 않을수 있다는것을 의미하며 이때 일반적으로 저차의 모멘트를 리용하여 미지파라메터를 추정한다.

실례6.2.3 는 로부터 얻은 표본이며 는 모두 미지파라메터이다. 이면 이므로 이고 따라서 의 모멘트추정 을 얻을수 있다.

만일 평등분포 로부터 용량이 5인 표본 4.5, 5.0, 4.7, 4.0, 4.2을 발취한다고 하자. 계산하면 이고 의 모멘트추정은 이다.

### 6.2.3 일치성

점추정값은 통계량이고 따라서 우연량이며 결국 표본크기가 일정하게 주어지면 점추정값이 파라메터의 실지 값과 완전히 같아질수는 없다. 그러나 충분한 관측값을 얻었다면 글리문코의 정리에 따라 표본량이 증가함에 따라 경험적분포함수가 실제 분포함수에 접근하며 따라서 이때 추정량이 파라메터의 실지 값에 접근하도록 할수 있다. 이것이 일치성(consistency)이며 그 정의는 다음과 같다.

정의 6.2.1 를 미지파라메터, 은 의 추정량, 은 표본용량이라고 하자. 만일 에 대하여



이 만족되면 를 파라메터 의 **일치추정**이라고 한다.

일치성은 추정에 대한 가장 기본적인 요구의 하나이며 어떤 추정량이 표본용량이 계속 커짐에 따라 추정정확도를 임의로 지정하는 수준으로 보장할수 없다면 그것은 믿기 어려다. 일반적으로 일치성요구조건을 충족하지 못하는 추정은 고려하지 않는다.

만일 표본용량 에 의존하는 추정량 을 하나의 우연량렬로 본다면 일치성은 가 로 확률수렴한다는것이며 따라서 추정의 일치성을 증명할 때 확률수렴의 성질과 여러가지 큰수의 법칙을 리용할수 있다.

실례6.2.4 이 정규분포 로부터 얻은 표본렬이면 신친의 큰수법칙과 확률수렴의 성질로부터 다음과 같은것을 알수 있다.

• 는 의 일치추정이다.

• 는 의 일치추정이다.

• 도 의 일치추정이다.

따라서 파라메터의 일치추정은 한개뿐이 아니라는것을 알수 있다.

추정의 일치성을 판단할 때 다음의 두 정리가 매우 쓸모있다.

정리 6.2.1 은 의 추정량이다. 만일



이면 은 의 일치추정이다.

증명. 에 대하여 체븨쉐브부등식으로부터 

한편 로부터 이 충분히 클 때 이다. 이때 만일 이면 다음과 같다.

. 그러므로 이고 동등하게 이다.

이로부터 이고 정리는 증명된다.

실례6.2.5 가 모집단 으로부터 얻은 표본일 때 이 의 일치추정이라는것을 증명하시오.

증명. 순서통계량의 분포로부터 의 밀도함수는 이고 따라서



그러므로 정리 6.2.1로부터 는 의 일치추정이라는것을 알수 있다.

정리 6.2.2만일 이 각각 의 일치추정이고 이 련속함수라면 는 의 일치추정이다.

증명. 함수 의 련속성으로부터



또한 의 일치성으로부터 주어진 에 대하여  따라서



식 (6.2.3)에 의하여 이며 이다. 의 임의성으로부터 정리가 증명된다.

큰수의 법칙과 정리 6.2.2로부터 모멘트추정은 일반적으로 일치성을 만족한다. 례를 들어

• 표본평균은 모집단평균의 일치추정이다.

• 표본표준편차는 모집단표준편차의 일치추정이다.

• 표본변이곁수 는 모집단변이곁수의 일치추정이다.

실례 6.2.6 한가지 실험의 가능한 결과가 세가지이며 그 발생확률은 각각 , 이다.

차례의 실험을 거쳐 관측한 세가지 결과의 발생회수는 각각 이며 잦음률치환법으로 를 추정할수 있다. 에 관한 세가지 서로 다른 식 이 있으므로 다음과 같은 잦음률치환추정을 얻을수 있다. 

큰수의 법칙으로부터 이 각각 의 일치추정이므로 정리 6.2.2으로부터 이 세가지 추정이 모두 일치추정이라는것을 알수 있다.

련습문제 6.2

1. 한 무지의 전자부속품가운데서 임의로 8개를 뽑아 수명을 측정하였는데 아래와 같은 자료(단위:시간)를 얻었다.

1050, 1100, 1130, 1040, 1250, 1300, 1200, 1080.

이 부속품의 평균수명과 수명분포의 표준편차에 대하여 모멘트추정을 구하시오.

2. 모집단 으로부터 용량이 10인 표본을 우연발취했는데 표본값들은 0.5, 1.3, 0.6, 1.7, 2.2, 1.2, 0.8, 1.5, 2.0, 1.6이다. 파라메터 에 대한 모멘트추정을 구하시오.

3. 모집단분포렬이 다음과 같고 이 표본일 때 미지파라메터에 대한 모멘트추정을 구하시오.

① 이고 이 미지파라메터.

② .

4. 모집단밀도함수가 다음과 같고 이 표본일 때 미지파라메터에 대한 모멘트추정을 구하시오.

① .

② .

③ .

④ .

5. 모집단 에 대하여 번 관측하여 그중 번의 관측값이 정수라는것을 알았다. 잦음률치환법으로 를 추정하시오.

6. 두 교정원 A, B가 서로 독립적으로 같은 원고를 교정하였다. 교정을 마친 후 A는 개, B는 개의 오자를 발견하였다. 모멘트추정법으로 다음과 같은 미지파라메터를 추정하시오.

① 이 원고의 오자총수.

② 발견되지 않은 오자수.

7. 모집단 가 2항분포 에 따르는데 여기서 는 미지파라메터이다. 이 의 표본일 때 미지파라메터 에 대한 모멘트추정을 구하시오.

## 6.3 최대우도추정과 EM알고리듬

최대우도추정(MLE)은 1821년 도이췰란드의 수학자 가우스에 의해 처음으로 정규분포에 대하여 제안되였지만 1922년 피셔가 다시 이런 원리를 제기하고 일부 성질을 증명한 다음부터 널리 리용되기 시작하였다. 이 절에서는 최대우도추정의 정의와 계산, 그리고 복잡한 경우에 MLE의 효과적인 알고리듬인 EM알고리듬을 취급하며 최대우도추정의 점근정규성을 소개한다.

### 6.3.1 최대우도추정

최대우도추정원리를 직관적으로 잘 리해하기 위하여 두개의 실례를 들어보자.

실례 6.3.1 겉모양이 꼭같은 두개의 상자가 있는데 상자1에는 99개의 흰 공과 1개의 검은 공이 있고 상자2에는 99개의 검은 공과 1개의 흰 공이 있다. 이제 우연적으로 상자 한개를 선택하고 그중에서 한개의 공을 뽑았는데 흰 공이였다. 이 공이 어느 상자에서 꺼낸것인가?

풀기. 어떤 상자든 상자에서 공을 꺼낼 때 두가지 결과를 얻을수 있다. 흰 공을 꺼낸것을 A로, 검은 공을 꺼낸것을 B로 표시하자. 만일 선택한것이 상자1이라면 A가 발생할 확률은 0.99이지만 상자2를 선택하였다면 그 확률은 0.01이다. 지금 첫 시행에서 A가 발생하였으므로 사람들의 자연적인 생각은 "이 흰 공이 상자1에서 꺼낸듯하다"는것 또는 시행조건이 결과 A에 유리한것을 생각하면 이 공이 상자1에서 꺼낸것이라고 추론할수 있다. 이런 생각은 사람들의 경험적사실에도 잘 부합된다. 여기서 "듯하다"는것이 바로 "최대우도"의 본질이며 이 생각을 "최대우도원리"라고 부른다.

우의 실례에서 가정한 자료는 매우 극단적인데 일반적으로 다음과 같이 생각할수 있다. 두 상자에 각각 100개의 공이 있고 상자1에서 흰 공의 비률은 , 상자2에서 흰 공의 비률은 이며 이라고 하자. 이제 우연히 한개 상자를 선택하고 그중에서 공을 하나 임의로 뽑을 때 그 공이 흰색이라고 가정하자. 만일 두 상자가운대서 한개를 선택하라고 하면 상자1의 흰 공 비률이 상자2보다 높기때문에 최대우도원리에 의하여 그 공을 상자1에서 꺼낸것이라고 판단하여야 한다.

실례 6.3.2 어떤 제품을 합격품과 불합격품의 두 종류로 나누는데 우연량 를 리용하여 검사하여 알아낸 불합격품수를 표시한다. 그러면 은 합격품, 은 불합격품을 나타내며 는 2점분포 에 따른다. 여기서 는 미지의 불합격률이다. 이제 개 제품을 발취하고 합격품인가를 검사하여 표본 을 얻었다. 이런 관측값이 얻어질 확률은



가 미지이므로 최대우도원리에 따라 식 (6.3.1)으로 표시되는 확률이 최대로 되는 를 선택하여야 한다.

식 (6.3.1)을 미지파라메터 의 함수로 보고 로 표시하며 **우도함수**(likelihood function)라고 한다. 즉

.

식 (6.3.2)의 최대값점을 구하는것은 어렵지 않다. 두변에 로그를 취하고 에 관한 도함수를 구하여 0이라고 놓으면 우도방정식이라고 하는 다음과 같은 방정식이 얻어진다.



이것을 풀면 의 **최대우도추정** 이 얻어진다.

실례6.3.2로부터 최대우도추정을 구하는 기본적인 사고방식을 알수 있다. 리산모집단의 경우 표본관측값 이 있다고 할 때 그 관측값이 출현하는 확률을 쓸수 있는데 그것은 일반적으로 로 표시되는 어떤 한개 또는 여러개의 파라메터에 의존한다. 그 확률을 의 함수로 보고 로 표시하며 역시 우도함수라고 부른다. 즉 . 최대우도추정을 구하는것은 바로 가 최대로 되는 의 추정값 을 찾는것이다.

련속모집단의 경우 표본관측값 가 나타날 확률은 항상 0이다. 그러나 동시확률밀도함수를 리용하여 우연량이 그 관측값주위에서 나타날 가능성이 얼마나 큰가를 나타낼수 있으며 이것을 역시 우도함수라고 한다. 따라서 다음과 같은 정의를 줄수 있다.

정의 6.3.1 모집단의 확률함수를 라고 하자. 여기서 는 미지의 한개 파라메터 또는 여러개 미지파라메터들로 구성된 파라메터벡토르이며 는 파라메터공간, 는 그 모집단으로부터 얻은 표본이다. 표본의 동시확률함수를 의 함수로 보고 로 간단히 로 표시하자.



이때 를 표본의 **우도함수**라고 한다. 만일 통계량 가



를 만족하면 를 의 **최대우도추정**이라고 하고 간단히 MLE(maximum likelihood estimate)라고 표시한다.

는 의 단조증가함수이므로 로그우도함수 를 최대화하는것과 를 최대화하는것은 동등하다. 일반적으로 으로부터 출발하여 의 최대우도추정을 찾는다. 이 미분가능한 함수일 때 보통 도함수를 구하여 최대우도추정을 구하며 이때 로그우도함수에 대한 미분이 더 간단하다.

최대우도추정의 정의로부터 과 동시밀도함수가 와 무관계한 인수만 차이날 때 그 추정에 아무런 영향도 미치지 않는다는것을 알수 있으며 따라서 에서 와 무관계한 인수를 제거할수 있다.

실례 6.3.3(실례 6.2.6의 계속) 실례 6.2.6에서 에 대한 3개의 모멘트추정을 론의하였다. 여기서는 의 최대우도추정을 보겠다.

우도함수는 이고 그 로그우도함수는 이다.

에 관한 도함수를 취하여 0으로 놓으면 우도방정식을 얻을수 있다.

.

이것을 풀면  이므로 는 극대값점이다.

실례6.3.4 정규모집단 에 대하여 는 2차원파라메터, 표본을 라고 하면 우도함수와 그의 로그는 각각



을 두개의 파라메터성분에 관하여 도함수를 취하고 0으로 놓으면 우도련립방정식





을 얻는다. 이 방정식을 풀면 식 (6.3.6)으로부터 의 최대우도추정 을 얻을수 있으며 이것을 식 (6.3.7)에 대입하면 의 최대우도추정 을 얻을수 있다.

2계미분곁수행렬의 정의 비부값성으로부터 우의 추정이 우도함수가 최대값으로 되게 한다는것을 알수 있다.

비록 도함수를 구하는 방법은 최대우도추정을 구하는 가장 일반적인 방법이지만 모든 경우에 다 효과적인것은 아니다. 이 문제는 아래의 실례로 설명할수 있다.

실례6.3.5 가 평등분포 의 표본일 때 의 최대우도추정을 구하시오.

풀기. 우도함수 를 최대로 되게 하려면 먼저 지시함수(indicator function, characteristic function, 示性函数)가 1을 취하여야 하고 다음은 이 가능한껏 커야 한다. 이 의 단조감소함수이므로 의 값은 될수록 작아야 하지만 지시함수가 1이라는것이 가 보다 작을수 없다는것을 의미하므로 의 최대우도추정은 이다.

최대우도추정은 간단하고 쓸모있는 성질을 가진다. 즉 만일 가 의 최대우도추정이면 임의의 함수 에 대하여 이 그의 최대우도추정이다. 이 성질을 최대우도추정의 **불변성**이라고 하며 이것으로 하여 구조가 복잡한 파라메터에 대한 최대우도추정도 쉽게 구할수 있다.

실례6.3.6 가 정규모집단 의 표본이라고 할 때 실례 6.3.4에서 와 의 최대우도추정이 이라는것을 구하였다.

따라서 최대우도추정의 불변성으로부터 다음과 같은 파라메터들의 최대우도추정을 얻을수 있다.

• 표준편차 의 MLE는 이다.

• 확률 의 MLE는 이다.

• 모집단0.90분위수 의 MLE는 이며 여기서 은 표준정규분포의 0.90분위수를 나타낸다.

### 6.3.2 EM알고리듬

MLE는 파라메터를 추정하는 매우 효과적인 방법이지만 분포에 다른 잠재적인 파라메터가 있거나 자료가 불완정 또는 결손되였을 때 MLE을 얻기가 어렵다. 뎀스터(Dempster) 등이 1977년에 제안한 EM알고리듬은 MLE를 구하는 과정을 두 단계로 나누는데 첫 단계는 기대값을 구하는것이고 둘째 단계는 극소값을 구하는것이다. 여기서는 매우 쓸모있는 EM알고리듬을 간단히 설명한다.

실례 6.3.7 한번의 시행에서 네가지 결과가 발생할수 있는데 그 확률은 각각 이다. 여기서 이다. 이제 197번의 시행결과 네가지 결과의 발생회수가 각각 75, 18, 70, 34였다면 의 MLE를 구하시오.

풀기. 네가지 결과가 발생하는 회수를 각각 로 표시한다면 모집단분포가 다항분포이므로 우도함수는 이다. 이것으로부터 의 MLE를 구하는것은 그 로그우도방정식이 3차다항식이므로 비교적 시끄럽다.

이제 2개의 변수 을 받아들여 비교적 간단하게 풀수 있다. 가령 첫번째 결과를 그 발생확률이 각각 , 인 두개 부분으로 나눌수 있다고 하고 두 부분에 속하는 회수를 각각 이라고 하자. 다른 한편 세번째 결과를 발생확률이 각각 인 두개 부분으로 나누고 두 부분에 속하는 회수를 각각 이라고 하자. 분명히 은 우리가 인입한것이며 관측불가능한것으로서 **잠재변수**(latent variable)라고 한다. 또한 자료 를 **완전자료**(complete data)라고 하고 관측한 자료 를 **불완전자료**라고 한다. 이때 완전자료의 우도함수는 이고 그 로그우도는 이다. 만일 가 모두 기지라면 웃식으로부터 쉽게 의 MLE를 얻을수 있지만 유감스럽게도 만 알고 의 값은 알지 못한다.

그러나 와 가 기지일 때 이므로 뎀스터 등은 다음과 같은 두단계로 나누어 반복적으로 풀이를 구할것을 제기하였다.

E단계: 관측자료 와 걸음의 추정값 이 주어진 조건에서 완전자료에 기초한 로그우도함수의 기대값을 구한다. 즉 기대값가운데서 와 관련된 부분은 적분하여 없애버린다.

.

M단계: 의 에 관한 최대값 을 구한다. 즉



인 를 구한다.

이렇게 로부터 로의 한차례의 반복을 수행한다. 식 (6.3.8)과 (6.3.9)을 수렴할 때까지 반복하여 의 MLE를 얻는다.

이 실례에서 E단계는 다음과 같다.



M단계에서는 우의 식에서 에 관한 도함수를 구하고 0으로 놓는다. 즉

.

이것을 풀면 반복식 을 얻을수 있다.

시작할 때 초기값은 임의로 택할수 있다. 실례로 을 택하면 13번 반복한 다음 의 MLE를 0.606747로 계산할수 있다. 반복과정은 아래와 같다.

%FontSize=12
%TeXFontSize=12
\documentclass{article}
\pagestyle{empty}
\begin{document}
\[
\begin{tabular}{ccccc}
\hline $i$ & $\theta^{(i)}$ & $\frac{\theta^{(i)}}{1+\theta^{(i)}} y_3+y_4$ & $\frac{1-\theta^{(i)}}{2-\theta^{(1)}} y_1+y_2$ & $\theta^{(i+1)}$ \\
\hline 0 & $0.5$ & $57.333333$ & $43.000000$ & $0.571429$ \\
1 & $0.571429$ & $59.454545$ & $40.500000$ & $0.594816$ \\
2 & $0.594816$ & $60.107784$ & $39.626214$ & $0.602681$ \\
3 & $0.602681$ & $60.323186$ & $39.325786$ & $0.605357$ \\
4 & $0.605357$ & $60.395987$ & $39.222803$ & $0.606271$ \\
5 & $0.606271$ & $60.420804$ & $39.187529$ & $0.606584$ \\
6 & $0.606584$ & $60.429289$ & $39.175449$ & $0.606691$ \\
7 & $0.606691$ & $60.432193$ & $39.171312$ & $0.606728$ \\
8 & $0.606728$ & $60.433187$ & $39.169896$ & $0.606740$ \\
9 & $0.606740$ & $60.433527$ & $39.169411$ & $0.606744$ \\
10 & $0.606744$ & $60.433644$ & $39.169245$ & $0.606746$ \\
11 & $0.606746$ & $60.433684$ & $39.169188$ & $0.606746$ \\
12 & $0.606746$ & $60.433697$ & $39.169168$ & $0.606747$ \\
13 & $0.606747$ & $60.433702$ & $39.169162$ & $0.606747$ \\
\hline
\end{tabular}
\]
\end{document}

이에 대한 설명은 다음과 같다.

① 실례로 의 분포를 설명하자. 로 첫째 결과가 나타났음을 표시하고 로 각각 우리가 정의한 두개의 사건을 표시하자(). 정의로부터 그것들은 서로 독립이며 , , 이고 이다. 따라서 의 조건밑에서 이다.

② 식 (6.3.8)의 오른변의 기대값은 에 관하여 의 조건에서 구한것이고 기타 파라메터들은 변하지 않으므로 왼변이 과 관련된다.

③ EM알고리듬은 매우 일반적인 조건에서 수렴한다.([22]를 참고)

### 6.3.3 점근정규성

최대우도추정은 일반적으로 점근정규성이라는 좋은 성질을 가진다.

실례 6.3.2 파라메터 의 일치추정 은 0으로 수렴하는 비부인 상수렬 이 있어서 이 표준정규분포로 분포수렴할 때 **점근정규성**을 가진다고 말한다. 이때 가 **점근정규분포** 에 따른다고도 말하며 로 표시하고 를 의 **점근분산**이라고 부른다.

우의 정의에서 0으로 가까워지는 수렬 는 추정량 이 로 확률수렴하는 속도를 표시한다. 왜냐하면 이 0으로 다가가는 속도와 이 로 확률수렴하는 속도가 같을 때(동차일 때)에만 그 비률 의 분포가 안정하게 표준정규분포 으로 수렴하기때문이다.만일 이 0으로 가는 속도가 너무 빠르면 그 비률은 로 되고 너무 뜨면 그 비률은 0으로 된다. 오직 가 0으로 가는 속도가 빠르지도, 느리지도 않을 때만 그 비률이 로 분포수렴할수 있다. 그러므로 가 0으로 가는 속도는 이 로 확률수렴하는 속도이며 따라서 를 점근분산이라고 하는것은 타당하다.

실례6.3.8 모집단을 뽜쏭분포 이라고 할 때 모멘트추정이나 최대우도추정이나 모두 동일한 의 추정 즉 표본평균 을 얻는다.

중심극한정리에 따라 은 로 분포수렴하며 따라서 은 점근정규성을 가진다. 그리고 . 여기서 상수렬 . 이것은 이 로 확률수렴하는 속도가 이라는것을 의미한다. 뒤에서 대부분의 점근정규추정은 모두 의 속도로 추정되는 파라메터에 확률수렴한다는것을 론의한다.

점근정규성에 대한 자세한 론의는 피하고 주로 다음과 같은 두가지를 강조하여둔다. 첫째, 최대우도추정의 점근정규성에 대한 결론을 증명없이 리용한다는것, 둘째, 점근정규성은 서로 다른 일치추정을 비교할 때 자주 리용되며 주로 점근분산의 크기를 가지고 비교된다는것이다.

정리 6.3.1 모집단 의 밀도함수가 이고 는 비퇴화구간이라고 하자. 이제 다음과 같이 가정하자.

① , 편도함수 가 모든 에 대하여 존재한다.

② ,  여기서 은 , 을 만족한다.

③ .

만일 이 그 모집단에서 발취한 표본이면 미지파라메터 에 대한 최대우도추정 이 존재하며 은 일치성과 점근정규성을 가진다. 즉 .

정리 6.3.1은 최대우도추정은 일반적으로 점근정규성을 만족하며 그 점근분산 는 공통적인 형식을 가진다는것을 설명하여준다. 그중에서 는 피셔정보량이다. 이에 대한 구체적인 정의는 다음 절에서 론의한다. 여기서는 모집단분포 에 따라 피셔정보량과 점근분산을 계산하기만 하면 된다.

실례6.3.9 는 의 표본이다.  또는 가 기지일 때 그 모집단분포가 정리 6.3.1의 세가지 조건을 만족한다는것을 검증할수 있다.

① 가 주어졌을 때 의 MLE는 이며 정리 6.3.1의하여 는 점근정규분포에 따른다. 이제 를 구하자.



따라서 이며 이것은 의 정확한 분포와 같다.

② 를 알고있을 때 의 MLE는 . 이제 을 구한다.



따라서 이다.

여러개의 일치추정이 있을 때 일반적으로 점근정규분포의 분산을 가지고 좋은것을 선택한다.

실례6.3.10(실례 6.2.6의 계속) 앞에서 에 대한 다음과 같은 세가지 일치추정을 제시하였다.



이것들 모두가 점근정규추정이라는것을 증명할수 있다. 즉  들은 각각 ([15]를 참고) 이다.

그림 6.3.1에서와 같이 세개의 점근분산을 비교해보면 첫 두개의 추정값은 좋은 점도 있고 나쁜점도 있지만 세번째 추정값은 이것들보다 일치하게 좋다. 그러므로 이것이 바로 최대우도추정이다.

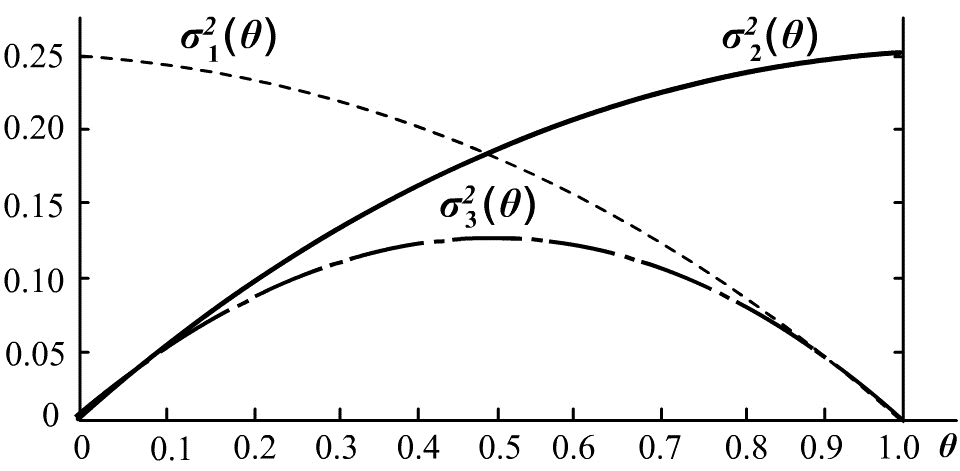


그림 6.3.1세개의 일치추정에서 점근분산의 비교

련습문제 6.3

1. 가 다음과 같은 확률함수를 가지는 모집단의 표본일 때 미지파라메터의 최대우도추정을 구하시오.

① .

② 이고 는 기지, 이다.

2. 가 다음과 같은 확률함수를 가지는 모집단의 표본일 때 미지파라메터의 최대우도추정을 구하시오.

① 이고 는 기지이다.

② .

③ 이고 는 기지이다.

3. 가 다음과 같은 확률함수를 가지는 모집단의 표본일 때 미지파라메터의 최대우도추정을 구하시오.

① .

② .

③ .

4. 한 지질학자가 어떤 호수가주변의 암석성분을 연구하기 위하여 그 지역에서 100개의 표본을 발취하였는데 매 표본은 10개의 암석덩어리이고 그 가운데서 석회석에 속하는 덩어리의 개수를 기록하였다. 이 100개의 관측이 서로 독립이라고 가정하면 이 지역의 암석가운데서 석회석의 비률 의 최대우도추정을 구하시오. 그 지질학자가 얻은 자료는 다음과 같다.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 표본의 암석수 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 표본개수 | 0 | 1 | 6 | 7 | 23 | 26 | 21 | 12 | 3 | 1 | 0 |

5. 유전학연구에서는 흔히 절단된 2항분포로부터 표본발취를 하는데 그 모집단확률함수는 다음과 같다.



이 주어지고 이 표본일 때 의 최대우도추정을 구하시오.

6. 어느한 책에서 한 문장의 단어수 는 근사하게 로그정규분포에 따른다는것을 알고있다. 즉 . 이 책에서 임의로 20개 문장을 뽑아 이 문장들속의 단어수를 세여본데 의하면 다음과 같다.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 52 | 24 | 15 | 67 | 15 | 22 | 63 | 26 | 16 | 32 |
| 7 | 33 | 28 | 14 | 7 | 29 | 10 | 6 | 59 | 30 |

이 책에서 한 문장의 평균단어수 의 최대우도추정을 구하시오.

7. 모집단 에서 은 미지파라메터, 를 이 모집단으로부터 얻은 표본, 를 표본평균이라고 할 때

① 는 파라메터 에 대한 불편추정, 일치추정이라는것을 증명하시오.

② 의 최대우도추정을 구하고 그것이 불편추정인가 일치추정인가를 판단하시오.

8. 밀도함수가 인 모집단으로부터 표본 을 발취하였다. 이때

① 의 최대우도추정 을 구하고 그것이 불편추정인가 일치추정인가를 판단하시오.

② 의 모멘트추정 을 구하고 그것이 불편추정인가 일치추정인가를 판단하시오.

9. 호수에 물고기가 몇마리 있는가를 추정하기 위해 1000마리를 건져내고 표식을 한 다음 다시 호수에 넣었다. 또 150마리를 건져내고 보니 그중 10마리에 표식이 있었다. 호수에 물고기가 몇마리 있어야 150마리 건져낸것가운데 표식을 한것이 10마리일 확률이 가장 큰가?

10. 정규분포 에 대하여 하나의 관측값만 있다면 의 최대우도추정이 존재하지 않는다는것을 증명하시오.

## 6.4 최소분산불편추정

앞에서 이미 론의한것처럼 점추정을 찾는데는 여러가지 방법이 있다. 서로 다른 점추정들을 비교선택하기 위해서는 그것들의 좋고나쁨에 대한 평가기준을 주어야 한다. 통계학에는 추정값을 평가할수 있는 기준이 많은데 동일한 추정값도 다른 기준에 따라 평가하면 완전히 다른 결론을 얻을수 있기때문에 어떤 추정을 평가할 때 먼저 기준이 무엇인가를 설명해야 하며 그렇지 않으면 그에 대한 론의가 아무런 의미도 없다.

### 6.4.1 평균두제곱오차

일치성과 점근정규성은 큰 표본의 경우에 좋고나쁨을 평가할수 있는 중요한 두가지 기준이다. 표본용량이 크지 않을 경우 사람들은 작은 표본에 맞는 기준을 리용하려고 한다. 이 경우 불편추정에 대하여서는 보통 분산을, 편향추정에는 평균두제곱오차를 리용한다.

일반적으로 표본량이 일정할 때 점추정의 좋고나쁨을 평가하는데 리용하는 지표는 항상 점추정값 와 파라메터의 진값 사이의 거리의 함수이며 가장 일반적인것은 함수는 거리의 두제곱이다. 은 우연성을 가지고있기때문에 이 함수에 대하여 기대값을 구할수 있는데 이것이 바로 다음과 같은 평균두제곱오차이다.



평균두제곱오차는 점추정을 평가하는 가장 일반적인 기준이다. 물론 추정의 평균두제곱오차가 작을수록 더 좋다. 한편 가 일정한 상수라는것을 고려하면



따라서 평균두제곱오차는 점추정의 분산과 편차 의 두제곱으로 구성된다. 만일 이 에 대한 불편추정이라면 이며 이때 점추정을 평균두제곱오차에 의하여 평가하는것과 분산을 리용하여 평가하는것은 동일하며 결국 분산을 리용하여 불편추정의 유효성을 추정하는것이 합리적이다. 이 에 대한 불편추정이 아닐 경우 그 평균두제곱오차 를 고려하여야 한다. 즉 분산의 크기뿐아니라 그 편차의 크기도 고려하여야 한다. 다음의 실례는 평균두제곱오차의 의미에서 일부 편향추정이 불편추정보다 좋다는것을 보여준다.

실례6.4.1 실례6.1.6에서 설명한것처럼 평등분포모집단 에 대하여 최대우도추정으로 얻은 의 불편추정은 이고 그의 평균두제곱오차는 이다.

이제 형태를 가지는 의 추정에 대하여 생각하자. 그 평균두제곱오차는 다음과 같다.



도함수를 구하는 방법으로 일 때 웃식의 평균두제곱오차가 가장 작다는것을 쉽게 구할수 있으며 이때 이다.

이것은 가 비록 의 편향추정이지만 일 때 평균두제곱오차 이라는것을 의미한다. 그러므로 평균두제곱오차를 기준으로 하면 편향추정 이 불편추정 보다 더 좋다.

정의 6.4.1 표본 이 주어졌다고 하자. 만일 파라메터 에 대한 추정클라스에 속하는 임의의 추정 에 대하여 파라메터공간 에서 모두



이면 를 그 파라메터추정클라스에서 의 **평등최소평균두제곱오차추정**이라고 부른다.

평등최소평균두제곱오차추정은 실례 6.4.1에서 추정을 의 배수들로 제한한것처럼 일정한 추정클라스에서 진행된다. 만일 추정을 제한하지 않는다면 즉 모든 가능한 추정을 고려한다면 평등최소두제곱오차추정은 존재하지 않으며 의의가 없다. 사실 이 의 모든 추정들가운데서 평등최소평균두제곱오차추정이면 임의의 을 택하여 라고 할 때 그것은 의 특정한 추정으로서 이고 평등최소평균두제곱오차추정 이 에서도 이여야 한다. 그런데 의 임의성으로부터 이 도처에서 0이여야 하는데 이것은 분명히 불가능한것이다.

평등최소평균두제곱오차추정이 일반적으로 존재하지 않으므로 추정에 대하여 일정한 합리적인 요구조건을 제기하는데 앞에서 언급한 불편성이 바로 가장 대표적인것이다.

### 6.4.2 평등최소분산불편추정

평균두제곱오차는 점추정의 분산과 편차의 두제곱으로 구성된다. 이 의 불편추정일 때 평균두제곱오차는 추정의 분산으로 되는데 이때 평등최소평균두제곱오차추정이 바로 평등최소분산불편추정이다. 그 일반적정의는 다음과 같다.

정의 6.4.2 파라메터추정문제에 대하여 이 의 불편추정이라고 하자. 만일 파라메터 에 대한 임의의 추정 에 대하여 파라메터공간 에서 모두 이면 를 의 평등최소분산불편추정(Uniform Minimum Variance Unbiased Estimator)이라고 하고 간단히 UMVUE라고 표시한다.

UMVUE에 대해 다음과 같은 판단규칙이 있다.

정리 6.4.1 가 어떤 모집단으로부터 얻은 표본이고 는 의 불편추정, 이면 가 의 UMVUE이기 위한 필요충분조건은 를 만족하는 임의의 에 대하여 이 성립하는것이다.

이 정리는 의 UMVUE는 임의의 기대값이 0인 불편추정과 비상관이여야 하며 그 거꿀도 성립한다는것을 보여주고있다. 이것은 UMVUE의 중요한 특징이다.

증명. 먼저 충분성을 증명한다. 에 대한 임의의 불편추정 에 대해 라고 하면 이다. 그리하여



이것은 가 의 불편추정클라스에서 분산이 평등최소라는것을 의미한다.

귀유법으로 필요성을 증명하자. 가 의 UMVUE이고 가 을 만족하고 파라메터공간 에 인 이 존재한다면 를 취할 때 이다.

라고 놓으면 이고 따라서 도 의 불편추정으로 된다. 그러나 그 분산에 대하여서는



이것은 가 의 UMVUE라는것과 모순되며 파라메터공간 에서 임의의 에 대하여서도 이 증명된다.

실례6.4.2 는 지수분포 의 표본이다. 인자분해정리로부터 는 의 충분통계량이다. 이므로 는 에 대한 불편추정이다. 이 의 임의의 불편추정이라고 하면  즉



량변에서 에 관한 도함수를 구하면 

이것은 이라는것을 의미하며 따라서 

정리 6.4.1에 의하여 는 의 UMVUE이다.

### 6.4.3 충분성준칙

실례 6.1.6에서 평등분포 의 두개 불편추정 와 의 좋고나쁨을 평가하였는데 좋은 불편추정은 충분통계량의 함수였다. 이것은 우연한것이 아니며 사실 충분통계량과 UMVUE가 존재한다면 UMVUE는 반드시 충분통계량의 함수로 표시할수 있다. 아래에서 이것과 관련된 결론을 소개한다.

정리 6.4.2 모집단확률함수는 , 는 표본, 는 의 충분통계량이면 의 임의의 불편추정 에 대하여 는 의 불편추정이며 이다.

증명. 가 충분통계량이므로 와 는 관계가 없다. 그러므로 는 의 추정(통계량)이고 재기대값공식에 의하여  따라서 는 의 불편추정이다. 그 분산을 보자.  이므로 이다.

웃식의 오른변에서 첫 항은 비부이므로 두번째 결론을 증명할수 있다.

정리 6.4.2는 불편추정이 충분통계량의 함수가 아니면 충분통계량에 관한 조건부기대값을 계산하여 새로운 불편추정을 얻을수 있으며 그 추정의 분산이 원래의 분산보다 작고 이렇게 불편추정의 분산을 줄일수 있다는것을 보여준다. 다시 말하여 의 추정문제는 충분통계량의 함수에서만 생각하면 된다. 이런 원리는 모든 통계적추론문제에서 성립한다. 이것이 바로 **충분성준칙**이다.

실례6.4.3 가 의 표본일 때 (또는 )는 의 충분통계량이다. 을 추정하기 위해 라고 하자. 이므로 는 에 대한 불편추정이다. 이 추정은 두개의 관측값만 리용하므로 좋지 않지만 정리 6.4.2를 리용하여 쉽게 개선할수 있다. 의 충분통계량 에 관한 조건부기대값을 구한다.



여기서 이다. 가 의 불편추정이고 라는것을 검증할수 있다.

### 6.4.4 크라메르-라오부등식

정리 6.3.1에서 지적한것처럼 최대우도추정의 점근분산은 주로 피셔정보량 에 의해 결정된다. 여기서는 먼저 을 소개한 다음 UMVUE를 판단하는데 리용될수 있는 크라메르-라오(Cramer-Rao)부등식에 대해 설명한다.

정의 6.4.3 모집단확률함수가 이고 다음과 같은 조건을 만족한다고 하자.

① 파라메터공간 는 직선우의 한개 열린구간이다.

② 지지모임 는 와 관계없다.

③ 도함수 는 모든 에서 존재한다.

④ 에 대하여 적분과 미분연산의 순서를 교체할수 있다. 즉 .

⑤ 기대값 이 존재한다.

이때



을 모집단분포의 **피셔정보량**이라고 한다.

피셔정보량은 통계학의 기본개념으로서 많은 통계결과들이 피셔정보량과 관계된다. 례하면 최대우도추정의 점근분산, 불편추정의 분산하한 등이 모두 피셔정보량 과 관계된다. 의 여러 성질들은 가 클수록 모집단분포에서 포함하는 미지파라메터 의 정보가 더 많다는것을 의미한다.

실례6.4.4 모집단이 뽜쏭분포 이고 그 분포렬이 라고 하자. 정의 6.4.3의 조건을 만족한다는것을 검증할수 있다. 또한



따라서 이다.

실례6.4.5 모집단이 지수분포이고 그 밀도함수는 이다. 정의 6.4.3의 조건을 만족한다는것을 검증할수 있다. 또한 

따라서 이다.

정리 6.4.3(크라메르-라오부등식) 모집단분포 이 정의 6.4.3의 조건을 만족하고 이 그 모집단의 표본, 은 의 임의의 불편추정이라고 하자. 또한 이 존재하고 에 대하여 의 도함수를 적분기호안에서 처리할수 있다고 하자. 즉



리산모집단에서는 우의 적분을 부호로 고친후에도 등식이 여전히 성립된다고 하자. 그러면

.

식 (6.4.5)을 **크라메르-라오(C-R)부등식**, 을 의 불편추정에 대한 분산의 C-R하계(간단히 의 C-R하계)라고 한다. 특히 의 불편추정 에 대하여 이다.

증명. 련속모집단에 대하여 증명하겠다. 이므로 량변의 에 관한 도함수를 구하면 적분과 미분의 순서를 교체할수 있기때문에

.

로 표시하면 이고 이다.

따라서 식 (6.4.5)로부터 . 그런데 슈와르쯔부등식으로부터 이다. 이렇게 식 (6.4.5)가 증명되며 리산모집단에 대해서도 류사하게 증명할수 있다.

식 (6.4.5)에서 등식이 성립하면 를 의 **유효추정**이라고 하는데 유효추정은 반드시 UMVUE이다.

실례6.4.6 모집단분포렬이 이면 정의 6.4.3의 조건을 모두 만족한다. 그러면 이 분포의 피셔정보량 을 계산할수 있다. 이제 이 그 모집단의 표본이면 의 C-R하계는 이다. 가 의 불편추정이고 그 분산이 이므로 의 분산은 C-R하계에 이른다. 따라서 은 의 유효추정이며 또 의 UMVUE이기도하다.

실례 6.4.7 모집단분포이 지수분포 이면 정의 6.4.3의 조건을 모두 만족한다. 실례 6.4.5에서 이 분포의 피셔정보량이 라는것을 계산하였다. 이제 이 그 모집단의 표본이면 의 C-R하계는 이고 는 의 불편추정이며 그 분산은 로서 C-R하계에 이른다. 따라서 은 의 유효추정이며 또 의 UMVUE이기도하다.

여기서 지적할것은 C-R하계에 도달할수 있는 불편추정(우의 두 경우)은 많지 않다는것이다. 대부분의 경우에 불편추정은 C-R하계에 도달하지 못한다. 아래에서 그 실례를 들어보자.

실례6.4.8 모집단을 정규분포 라고 할 때 정의 6.4.3의 조건을 모두 만족한다. 이제 피셔정보량을 계산하자. 이고 이므로



이제 이 표본이면으로 의 불편추정의 C-R하계는 이고 는 의 불편추정로서 그 분산은 C-R하계에 이른다. 그러므로 는 의 UMVUE이다. 다른 한편 라고 하면 의 C-R하계는 이고 의 불편추정(실례 6.1.2를 참고)은 

이것이 의 UMVUE이며 그 분산이 C-R하계보다 크다는것을 증명할수 있다. 그러므로 모든 의 불편추정의 분산이 C-R하계보다 크다.

련습문제 6.4

1. 모집단확률함수가 이고 는 그 표본, 는 의 충분통계량이라고 하면 에 대한 임의의 추정 에 대하여 라고 할 때 라는것을 증명하시오. 이 결론으로부터 평균두제곱오차의 준칙하에서 충분통계량에 기초한 추정만 고려하면 된다는것을 알수 있다.

2. 이 각각 의 UMVUE일 때 임의의 비령인 상수 에 대하여 이 의 UMVUE라는것을 증명하시오.

3. 가 의 UMVUE이고 가 의 불편추정일 때 이면 라는것을 증명하시오.

4. 모집단 이고 가 표본일 때 가 각각 의 UMVUE라는것을 증명하시오.

5. 모집단 의 피셔정보량이 존재한다고 할 때 2계도함수 가 모든 에 대하여 존재한다면 라는것을 증명하시오.

6. 모집단밀도함수를 이라고 하고 는 표본이다. 이때

① 의 최대우도추정을 구하시오.

② 의 유효추정을 구하시오.

7. 모집단밀도함수를 라고 할 때 에 대한 피셔정보량 를 구하시오.

8. 모집단밀도함수를 라고 하고 을 알고있을 때 에 대한 피셔정보량 를 구하시오.

9. 모집단분포렬이 일 때 에 대한 피셔정보량 를 구하시오.

10. 이 의 표본이고 이 주어졌을 때 이 의 유효추정이고 UMVUE라는것을 증명하시오.

11. 일 때 의 UMVUE를 구하시오.

12. 일 때 의 UMVUE를 구하고 이 UMVUE가 C-R부등식의 하계에 도달하지 못한다는것 즉 유효추정이 아니라는것을 증명하시오.

13. 뽜쏭분포 에 대하여

① 를 구하시오.

② 의 피셔정보량이 과 관계없도록 하는 함수 를 하나 찾으시오.

14. 가 독립동일분포하는 변수이고 이며 , 이라고 하자.

① 의 MLE 을 구하고 이 불편추정인가 아닌가를 판단하시오.

② 의 모멘트추정 를 구하시오.

③ 의 불편추정의 분산의 C-R하계를 구하시오.

15. 가 모집단 의 표본일 때 의 모멘트추정과 최대우도추정이 모두 이며 그것은 불편일치추정이다. 평균두제곱오차의 준칙에서 보다 좋은 추정이 존재한다는것을 증명하시오. (풀기방향: 을 생각하고 평균두제곱오차가 제일 작은것을 찾으시오)

## 6.5 베이스추정

통계학에는 크게 두개의 학파 즉 고전학파와 베이스학파가 있다. 이 책에서는 주로 고전학파의 리론과 방법을 소개하지만 베이스학파에 대해 이 절에서 간단히 소개한다.

### 6.5.1 통계적추정의 기초

앞에서 언급하였지만 통계추론은 표본정보에 근거하여 모집단의 분포나 특징수에 대해 추론한다. 사실상 이것은 통계추론에 대한 고전학파의 견해이다. 여기서 통계추론은 **모집단정보**와 **표본정보**라는 두가지 정보를 리용한다. 그러나 베이스학파는 우의 두가지 정보외에 통계추론이 또다른 제3의 정보 즉 **사전정보**를 리용해야 한다고 주장하였다. 우선 세가지 정보에 대해 설명하자.

① 모집단정보

모집단정보는 바로 모집단분포나 모집단이 속한 분포족이 제공하는 정보이다. 례를 들어 "모집단이 정규분포"라는것을 안다면 많은것을 알고있는것과 같다. 대표적으로 모집단의 모든 모멘트들이 존재하고 모집단밀도함수가 평균에 관하여 대칭이라는것, 모집단의 모든 성질들이 그의 1차, 2차 모멘트들에 의해 결정된다는것을 비롯하여 많은 성숙된 통계추론방법들을 리용할수 있다. 모집단정보는 매우 중요한 정보이며 이것을 얻으려면 많은 품이 든다. 례를 들면 베아링의 수명분포가 웨이블분포라는것을 확인하는데 5년이란 시간을 들여 수천개의 자료를 처리한후에야 이런 결론을 얻었다.

② 표본정보

표본정보란 표본을 발취하여 얻은 관측값이 제공하는 정보를 말한다. 례를 들어 표본관측값이 있으면 그것으로부터 모집단평균이나 분산과 같은 모집단의 일부 특징수를 일정한 범위내에서 대략적으로 알수 있다. 이것들은 가장 "생신한" 정보이며 많을수록 좋다. 표본을 통하여 모집단분포나 모집단의 일부 특징에 대하여 비교적 정확한 통계적추론을 할수 있으면 리상적이다. 표본이 없으면 통계학에 대하여 말할수 없다.

③ 사전정보

표본을 발취하는것을 시행으로 본다면 표본정보는 바로 시행과정에 얻는 정보이다. 실제로 사람들은 시행을 하기전에 경험적으로 또는 자료적으로 어느 정도 맞다든 문제를 료해하고있는데 이런 정보는 통계추론에 유익한것이다. 사전정보란 표본발취(시행)를 진행하기 전부터 존재하는 통계문제에 관한 정보를 말한다. 일반적으로 사전정보는 경험과 과거의 자료로부터 얻어진다. 사전정보는 일상생활과 사업에서도 매우 중요하다. 먼저 례를 들어보자.

실례 6.5.1 어느한 공장에서 매일 개의 제품을 발취하여 제품의 품질이 요구조건에 부합되는가를 확인한다. 제품품질은 불합격률 를 가지고 측정하는데 개 제품가운데서 불합격품의 개수 를 가지고 표시할수도 있다. 생산공정이 련속적이기때문에 매일의 제품품질은 관련이 있다. 즉 현재의 를 추정할 때 이전에 루적된 자료들도 리용하여야 한다. 이러한 과거의 자료들이 바로 사전정보이다. 이런 사전정보를 리용하려면 그에 대한 처리를 진행하여야 한다. 례를 들면 일정한 시간이 흐른 다음 과거의 자료에 기초하여 개 제품속의 불합격품수 에 대하여 다음과 같은 하나의 분포를 작성할수 있다.



이렇게 사전정보를 가공하여 얻은 분포를 앞으로 사전분포라고 부르며 이러한 사전분포는 그 공장에서 과거에 생산한 제품의 불합격률에 대한 종합적인 견해이다.

우의 세가지 정보에 기초하여 통계추론을 진행하는 통계학을 **베이스통계학**이라고 한다. 이것이 고전통계학과 다른것은 사전정보를 리용하는가 안하는가 하는데 있다. 베이스통계학은 모집단정보와 표본정보를 리용하는 동시에 사전정보의 수집과 발굴, 처리에 주목하여 그것을 수량화하고 사전분포를 형성하며 통계추론에 리용한다. 사전정보를 무시하는것은 일종의 랑비라고도 볼수 있지만 어떤 경우에는 불합리한 결론을 이끌어내기도 한다.

**베이스학파의 견해**는 다음과 같다. 임의의 미지량 는 모두 우연량으로 볼수 있으며 하나의 확률분포로 묘사할수 있다. 이 분포를 **사전분포**라고 한다. 표본을 얻으면 모집단분포, 표본과 사전분포는 베이스공식을 통하여 결합되여 미지량 에 관한 새로운 분포(사후분포)를 얻을수 있다. 임의의 에 관한 통계추론은 모두 의 사후분포에 기초하여 진행되여야 한다.

미지량을 우연량으로 볼수 있는가 없는가를 놓고 고전학파와 베이스학파가 오래동안 론의하였다. 임의의 미지량은 모두 불확정성을 가지고 있고 그 불확정성의 정도를 설명할 때 확률과 확률분포가 가장 좋은 언어이므로 그것을 우연량으로 보는것이 타당하다. 지금 고전학파도 이 관점을 반대하지 않고 있다. 유명한 고전통계학자 레만(Lehmann, E.L.)은 자기의 책 "점추정리론"에서 "통계문제의 파라메터를 우연량으로 보는것은 그것을 미지파라메터로 보는것보다 더 타당하다."고 지적하였다. 두 학파간에 지금은 어떻게 여러가지 사전정보를 리용하여 합리적으로 사전분포를 확정하겠는가 하는것을 놓고 론쟁을 하고 있다. 이 문제는 일부 경우에는 쉽게 해결되지만 많은 경우에 해결하기 어렵다(문헌 [11]을 참고).

### 6.5.2 베이스공식의 밀도함수형식

베이스공식의 사건형식은 1.4에서 설명하였다. 여기서는 다시 우연량의 확률함수를 리용하여 베이스공식을 설명하고 베이스학파의 구체적인 견해를 소개한다.

① 모집단이 파라메터 에 의존하는 확률함수를 고전통계에서는 로 표시한다. 이것은 파라메터공간 의 서로 다른 가 각이한 분포에 대응된다는것을 나타낸다. 그러나 베이스통계에서는 그 확률함수를 로 표시하는데 이것은 우연량 가 어떤 일정한 값을 취할 때 모집단의 **조건부확률함수**를 나타낸다.

② 파라메터 의 사전정보에 따라 사전분포 를 결정한다.

③ 베이스의 관점에서 표본 의 생성은 두 단계로 진행되여야 한다. 먼저 사전분포 로부터 하나의 실체 를 생성한다. 이것은 관측자들이 볼수 없다. 다음은 로부터 한조의 표본 을 생성한다. 이때 그것의 동시조건부확률함수는 이다. 이 분포는 모집단정보와 표본정보를 종합하고 있다.

④ 은 가상적으로 생성한것이므로 여전히 미지이다. 그것은 사전분포 에 따라 생성하였다. 사전정보를 종합하여 넣기 위하여서는 만 고려해서는 안되고 의 다른 값들이 발생할 가능성도 같이 고려해야 한다. 그리하여 를 리용하여 종합한다. 그러면 표본 와 파라메터 의 동시분포는 .

이 동시분포는 모집단정보, 표본정보, 사전정보의 세가지 리용가능한 모든 정보를 종합하였다.

⑤ 목적은 미지파라메터 에 대해 통계추론을 하는것이다. 표본정보가 없을 때 사전분포에 근거하여 를 추론할수 밖에 없다. 표본관측값 이 있으면 에 근거하여 를 추론할수 있다. 만일 를 로 분해하자. 여기서 은 의 주변확률함수이다.

.

즉 에는 에 대한 정보가 들어있지 않다. 그러므로 에 대하여 추론할수 있는것은 조건부분포 뿐이다. 그의 계산공식은

.

이 조건부분포를 의 **사후분포**라고 하는데 이것은 모집단, 표본, 사전지식에서 에 대한 모든 정보를 포함한다. 식 (6.5.3)이 바로 밀도함수로 표시된 베이스공식이며 이것은 모집단과 표본을 리용하여 사전분포 를 조정한 결과로서 보다 의 실지 정황을 더 정확히 반영한다.

### 6.5.3 베이스추정

사후분포 에 의하여 를 추정하는데는 일반적으로 세가지 방법이 있다.

• 사후분포의 밀도함수를 최대화하는 값을 의 점추정으로 하는 **최대사후추정**.

• 사후분포의 중위수를 의 점추정으로 하는 **사후중위수추정**.

• 사후분포의 평균값을 의 점추정으로 하는 **사후기대값추정**.

가장 많이 리용하는것은 사후기대값추정인데 일반적으로 베이스추정이라고도 하며 로 표시한다.

실례 6.5.2 사건 가 한번의 시행에서 발생할 확률이 인데 이것을 추정하기 위하여 번의 독립적인 관측을 진행하였고 그가운데서 가 번 발생하였다. 분명히 이고 이다.

만일 시행을 하기전에 사건 에 대해 아무것도 알지 못했다면 그 발생확률 에 대한 정보가 하나도 없다. 이 경우에 베이스는 "평등하게 모른다"는 원칙에서 평등분포 을 의 사전분포로 리용할것을 제안하였다. 왜냐하면 (0, 1)의 모든 점들이 로 될 기회가 동등하기 때문이다. 베이스의 이러한 제의를 **베이스가설**이라고 부른다. 그러면 베이스공식을 리용하여 의 사후분포를 구할수 있다. 구체적으로 우선 와 의 동시분포 를 구한다.

그다음 의 주변분포를 구한다.



마지막으로 의 사후분포를 구한다.



이 결과는 을 보여주며 그 사후기대값추정은



만일 사전정보를 리용하지 않고 모집단정보와 표본정보만 리용한다면 사건 의 발생확률에 대한 최대우도추정은 이다. 이것은 베이스추정과 다른것이다. 어떤 경우에 베이스추정은 최대우도추정보다 더 합리적일수 있다. 례를 들면 제품발취검사에서 불합격품과 합격품만을 구별하고 로 불합격률을 나타낸다고 하자. 질이 좋은 제품더미에서 발취된 제품들이 대체로 합격품인데 "검사한 3개 제품이 모두 합격품"이라는것과 "검사한 10개 제품이 모두 합격품"이라는 두 사건이 사람들에게 주는 인상은 같지 않고 두번째 사건이 첫번째 사건보다 더 믿음직하다. 이런 차이를 의 최대우도추정 에서는 두 사건이 모두 0이므로 반영하지 못하지만 베이스추정 에서는 과 으로 반영되게 된다. 비슷하게 질이 나쁜 제품더미에서는 발취된 제품들이 대체로 불합격품인데 "검사한 3개 제품이 모두 불합격품"이라는것과 "검사한 10개 제품이 모두 불합격품"이라는 두 사건가운데서 두번째 사건이 사람들에게 질이 더 나쁘다는 인상을 준다. 이런 차이를 에서는 두 사건이 모두 1이므로 반영하지 못하지만 에서는 각각 , 로 차이나게 된다. 이렇게 극단적인 경우에 베이스추정은 최대우도추정보다 사람들의 리해에 더 부합된다.

실례6.5.3 는 정규분포 의 표본이고 은 기지이며 는 미지이다. 가령 의 사전분포가 정규분포 로서 와 이 모두 기지라면 의 베이스추정을 구하시오.

풀기. 표본 의 분포와 의 사전분포는 다음과 같다.



이로부터 와 의 동시분포를 알수 있다.



여기서 이다. 이제 라고 놓으면 .

가 모두 와 관련이 없으므로 표본의 주변밀도함수를 쉽게 구할수 있다.



베이스공식을 리용하면 사후분포를 구할수 있다.



이것은 표본이 주어지면 의 사후분포가 라는것을 의미한다. 즉



사후분포평균이 그 베이스추정으로 된다.  이것은 표본평균 와 사전평균 의 무게평균이다. 모집단분산 이 비교적 작거나 표본량 이 클 때 의 무게가 커지며 사전분산 이 비교적 작을 때 의 무게가 커진다. 이것은 사람들의 일반적인 경험에 부합된다.

### 6.5.4 공액사전분포

베이스공식으로부터 알수 있는것처럼 전체적인 베이스통계추론은 사전분포가 확정된 후에는 리론상 어려울것이 없다. 사전분포를 결정하는데는 여러가지 방법이 있는데 가장 널리 리용되는 사전분포의 한 종류인 공액사전분포를 여기서 소개한다.

정의 6.5.1 는 모집단분포 의 파라메터, 는 그 사전분포이다. 의 임의의 표본을 관측한 후 사후분포 와 가 같은 분포족에 속하면 이 분포족을 의 **공액사전분포(족)**이라고 부른다.

실례 6.5.4 실례 6.5.2에서 이 베타분포의 특수경우인 이며 그에 대응하는 사후분포는 베타분포 라는것을 보았다. 좀 더 일반적으로 의 사전분포가 이고 가 모두 기지라면 베이스공식을 리용하여 사후분포가 라는것을 구할수 있다. 이것은 베타분포가 베르누이시행에서 성공확률의 공액사전분포라는것을 말해준다.

마찬가지로 실례6.5.3로부터 분산이 주어졌을 때 정규모집단평균의 공액사전분포가 정규분포라는것을 알수 있다.

련습문제 6.5

1. 어떤 책 한 페지에 있는 오자의 수가 뽜쏭분포 에 따르는데 는 1.5, 1.8의 두 값을 취할수 있고 사전분포는 이다.

현재 한 페지를 검사한 결과 세개의 오자를 발견하였다면 의 사후분포를 구하시오.

2. 모집단은 평등분포 이고  의 사전분포는 평등분포 이다. 현재 세개의 관측값 11.7, 12.1, 12.0이 있다면 의 사후분포를 구하시오.

3. 는 모집단분포렬이 인 기하분포로부터 얻은 표본이고 의 사전분포는 이다.

① 의 사후분포를 구하시오.

② 만일 4번의 관측값이 4, 3, 1, 6이라면 의 베이스추정을 구하시오.

4. 뽜쏭분포의 평균값 의 공액사전분포가 감마분포라는것을 검증하시오.

5. 정규모집단분산(평균은 기지)의 공액사전분포가 거꿀감마분포(가 감마분포에 따르면 는 거꿀감마분포에 따른다고 한다)라는것을 검증하시오.

6. 를 모집단 로부터 발취한 표본이라고 하자.

① 만일 의 사전분포가 평등분포 이면 의 사후분포를 구하시오.

② 만일 의 사전분포가 이면 의 사후분포를 구하시오.

7. 를 모집단 로부터 발취한 표본이라고 하자. 의 사전분포를 감마분포 라고 하면 의 사후기대값추정을 구하시오.

8. 를 모집단 로부터 발취한 표본이라고 하자. 의 사전분포는 밀도함수가 인 파레토분포이며 여기서 은 기지상수이다.

① 파레토분포는 의 공액사전분포라는것을 검증하시오.

② 에 대한 베이스추정을 구하시오.

9. 지수분포 에서 미지파라메터 의 사전분포를 감마분포 라고 하자. 사전정보로부터 사전평균이 0.0002이고 사전표준편차는 0.01이라는것을 알고 있다. 사전분포를 확정하시오.

10. 는 모집단밀도함수가 인 제곱합렬분포에서 발취된 표본이다. 이때

① 만일 가 기지라면 의 공액사전분포는 파레토분포라는것을 증명하시오.

② 만일 가 미지라면 의 공액사전분포는 감마분포라는것을 증명하시오.

11. 어떤 사람이 매일 아침 뻐스정류소에서 기다리는 시간(단위:분)은 에 따른다. 여기서 는 미지이며 그 사전분포를 라고 하자.

만일 이 사람의 기다리는 시간이 3일동안 각각 5, 3, 8분이였다면 의 사후분포를 구하시오.

12. 정규모집단 에서 용량이 100인 표본을 우연발취하고 의 사전분포가 정규분포라고 한다면 사전분포의 표준편차가 무엇이든 사후분포의 표준편차는 1/5보다 작다는것을 증명하시오.

13. 우연량 가 확률분포가 인 부의 2항분포에 따른다고 가정하면 그 성공확률 의 공액사전분포족이 베타분포족이라는것을 증명하시오.

14. 어떤 제품더미에서 100개를 우연검사하는데 3개의 불합격품을 발견하였다. 그 제품의 불합격률 의 사전분포를 베타분포 라고 하면 의 사후분포를 구하시오.

## 6.6 구간추정

파라메터의 점추정은 하나의 구체적인 수값을 제공하는데 계산과 리용에는 편리하지만 그 정확성은 점추정에 의하여 설명할수 없으며 그 분포에 의하여 반영해야 한다. 실천에서 한 점추정의 정확성을 측정하는 가장 직관적인 방법은 미지파라메터의 구간을 주는것이며 이것이 바로 구간추정의 개념이다.

### 6.6.1 구간추정의 개념

이 모집단의 파라메터이고 이 표본이라면 구간추정은 바로 두개의 통계량 와 을 구하여 이고 표본관측값을 얻은 다음의 의 추정이 구간 안에 놓이게 하는것이다. 표본의 우연성으로 하여 구간 이 미지파라메터 를 피복할 가능성은 확정할수 없지만 일반적으로 이 구간이 를 피복할 확률 를 최대한 크게 한다. 구간의 길이가 클수록 좋아지므로 이 문제를 해결하기 위해 먼저 구간 이 를 피복할 확률(앞으로 믿음성수준이라고 부른다)을 미리 설정한다. 이렇게 되여 믿음구간의 개념이 도입된다.

정의 6.6.1 을 모집단파라메터, 파라메터공간은 , 은 표본이다. ()가 주어졌을 때 두개의 통계량 와 가 있어서 임의의 에 대해



이 성립하면 우연구간 을 의 **믿음수준이 인 믿음구간** 또는 간단히 의 **믿음구간**이라고 한다. 또한 을 각각 의 (량측)**믿음하한, 믿음상한**이라고 한다.

믿음수준 ****는 잦음률로 설명할수 있다. 의 믿음구간 을 다량으로 중복리용할 때 매번 표본관측값들이 서로 다르기때문에 얻은 구간도 차이난다. 매 구체적인 관측값에 대하여 는 안에 놓일수도, 놓이지 않을수도 있다. 평균적으로 보면 이러한 다량의 구간관측값가운데서 적어도 ****%가 를 포함한다. 다음 실례의 그림 6.6.1과 6.6.2에서 이러한 잦음률의미를 직관적으로 보여주었다.

실례6.6.1 가 의 표본일 때 의 ****믿음구간은 이다. 여기서 는 각각 표본평균값과 표본표준편차이다. 이 믿음구간을 얻는 방법은 6.6.3에서 취급하며 여기서는 믿음구간과 믿음수준의 의미를 설명하는데 리용한다.

을 취하면 이고 웃식은 로 된다.

라고 가정하면 우연모의법으로 로부터 용량이 10인 표본을 생성할수 있다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 14.85 | 13.01 | 13.50 | 14.93 | 16.97 |
| 13.80 | 17.95 | 13.37 | 16.29 | 12.38 |

이 표본으로 을 계산할수 있다.

따라서 의 구간추정은 이다. 이 구간은 의 진값 15를 포함한다. 이 방법을 100번 반복하면 100개의 표본이 얻어지며 따라서 100개의 구간이 얻어진다. 이 100개의 구간을 그림 6.6.1에 그려보았다. 그림 6.6.1에서 볼수 있듯이 100개의 구간가운데서 91개는 파라메터의 진값 15을 포함하고 나머지 9개는 진값을 포함하지 않는다. 이것이 믿음수준 에 대한 해석이다.

를 취하면 이고 따라서 의 0.5믿음구간은 이다.

우의 표본에 대하여 의 구간추정은 이다.

이 구간은 파라메터의 진값을 포함하며 비슷하게 이러한 구간들은 100개 얻을수 있다. 그림 6.6.2에서 볼수 있듯이 100개의 구간가운데서 50개는 진값 15을 포함하고 나머지 50개는 진값을 포함하지 않는다. 이것이 믿음수준 에 대한 해석이다. 만일 다른 100개의 표본을으로 바꾸었다고 해서 꼭 50%가 진값을 포함한다고는 할수 없지만 거의 비슷해야 한다. 례컨대 49개나 51개는 모두 합리적이다.

정의 6.6.1에서는 부등식을 리용하여 구간추정에 대한 정의를 주었는데 이것은 일반적으로 모집단이 리산분포인 경우를 고려한것이다. 모집단이 련속분포일 경우에는 믿음수준을 충분히 리용하기 위하여 등식을 리용한다.

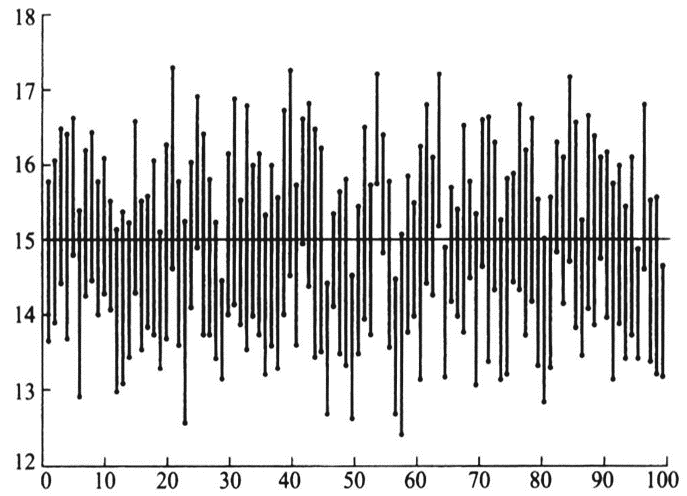


그림 6.6.1 의 0.90믿음구간

정의 6.6.2 정의 6.6.1의 기호를 그대로 리용하면 ()가 주어졌을 때 임의의 에 대해



이면 를 의 동등믿음구간이라고 한다.

일부 실천 문제들에서 사람들은 미지파라메터의 하한이나 상한에만 관심을 가진다. 실례로 어떤 제품의 평균수명을 취급할 때 그것이 길수록 좋기때문에 그의 0.90믿음하한이 얼마인가 하는데 관심을 돌리고 그것이 제품의 질을 평가한다고 생각한다. 일반적인 정의는 다음과 같다.

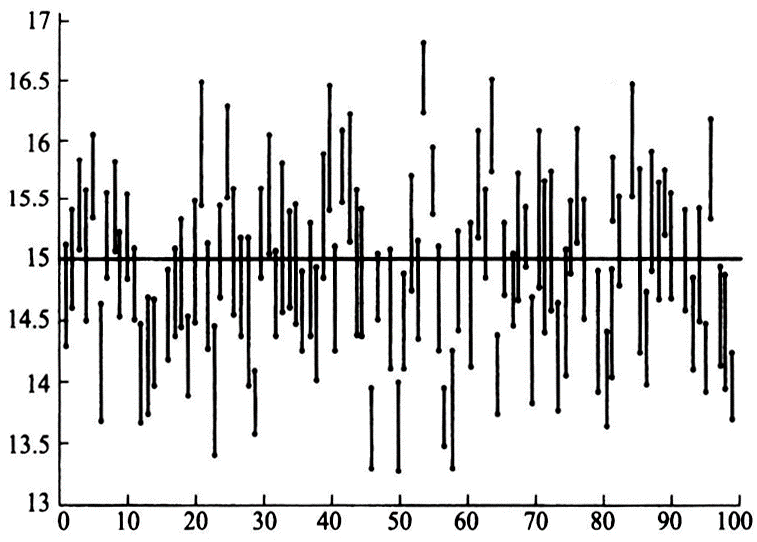


그림 6.6.2 의 0.50믿음구간

정의 6.6.3 는 통계량라고 하자. ()가 주어졌을 때 임의의 에 대해



이면 를 의 믿음수준이 인 **(단측)믿음하한**이라고 한다. 또한 모든 에 대하여 등식이 성립하면 를 의 **동등믿음하한**이라고 한다.

마찬가지로 어떤 약품의 독성과 같은 지표들은 될수록 작으면 한다. 그리하여 믿음상한의 개념을 받아들인다.

정의 6.6.4 는 통계량라고 하자. ()가 주어졌을 때 임의의 에 대해



이면 를 의 믿음수준이 인 **(단측)믿음상한**이라고 한다. 또한 모든 에 대하여 등식이 성립하면 를 의 **동등믿음상한**이라고 한다.

단측믿음하한과 단측믿음상한은 모두 믿음구간의 특수경우이다. 따라서 믿음하한과 상한의 위치를 찾는 방법으로 믿음구간을 구할수 있다. 이제는 믿음구간을 찾는 방법을 소개한다.

### 6.6.2 중심축량법

미지파라메터 에 대한 믿음구간을 구성하는 가장 일반적인 방법은 중심축량(Pivot, 枢轴量)법이며 그 절차는 다음과 같다.

① 표본과 의 함수 를 구성하여 의 분포가 에 의존하지 않도록 한다. 일반적으로 이런 성질을 가진 를 중심축량이라고 한다.

② 두 상수 를 적당히 선택하여 주어진 ()에 대하여



이 만족되도록 한다. 리산의 경우에는 웃식의 같기부호가 로 바뀌운다.

③ 를 동등변환하여 형식의 부등식을 얻을수 있으면



인데 이것은 이 의 동등믿음구간이라는것을 의미한다.

우에서와 같이 믿음구간을 구성하는데서 중요한것은 중심축량 이므로 이 방법을 **중심축량법**이라고 한다. 중심축량의 선택은 일반적으로 의 점추정으로부터 출발한다. 식 (6.6.5)를 만족시키는 이 많을 때 식 (6.6.6)에서의 평균길이 가 짧게 하는 방향에서 선택한다. 물론 가 최단인 를 찾을수 있으면 제일 리상적이다. 그러나 적지 않은 경우에 이렇게 하기 어렵다. 이때는 두개의 꼬리부 확률이 각각 이 되게 하는 를 선택한다. 즉



이러한 믿음구간을 **등꼬리믿음구간**이라고 한다. 실천에서 널리 쓰이는 믿음구간은 대부분 등꼬리믿음구간이다.

실례6.6.2 는 평등모집단 의 표본이다. 주어진 ()에 대하여 의 동등믿음구간을 구하시오.

풀기. 중심축량법을 리용하여 세단계로 나누어 진행한다.

① 의 최대우도추정이 표본의 최대순서통계량 이며 의 밀도함수가 라는것을 알고있다.

이것은 와 무관계하므로 을 중심축량 로 택한다.

② 의 분포함수가 이므로 이다. 따라서 이 성립하는 적당한 를 선택할수 있다.

③ 부등식을 변형하여 의 동등믿음구간이 이라는것을 쉽게 구할수 있다. 이 구간의 평균길이는 이다. 이로부터 인 조건하에서 , 일 때 이 최소값을 취한다는것을 알수 있다. 이것은 가 에 대한 이 종류의 구간추정가운데서 믿음수준이 인 최단믿음구간이라는것을 의미한다.

### 6.6.3 한개 정규모집단파라메터에 대한 믿음구간

정규모집단 은 가장 일반적인 분포이며 여기서는 이 분포의 두 파라메터에 대한 믿음구간을 론의한다.

1. 가 주어졌을 때 의 믿음구간

이 경우 의 점추정은 이고 그 분포는 이므로 중심축량을 로 선택할수 있다. 는 을 만족시켜야 한다. 부등식을 변형하여 을 얻을수 있으며 그 구간의 길이는 이다. 표존정규분포는 단봉우리대칭이므로 그림 6.6.3에서 알수 있듯이 인 조건에서 일 때 가 최소로 된다. 따라서 의 동등믿음구간은

.

이것은 를 중심으로 하고 반경이 인 대칭구간으로서 보통 로 표시한다.

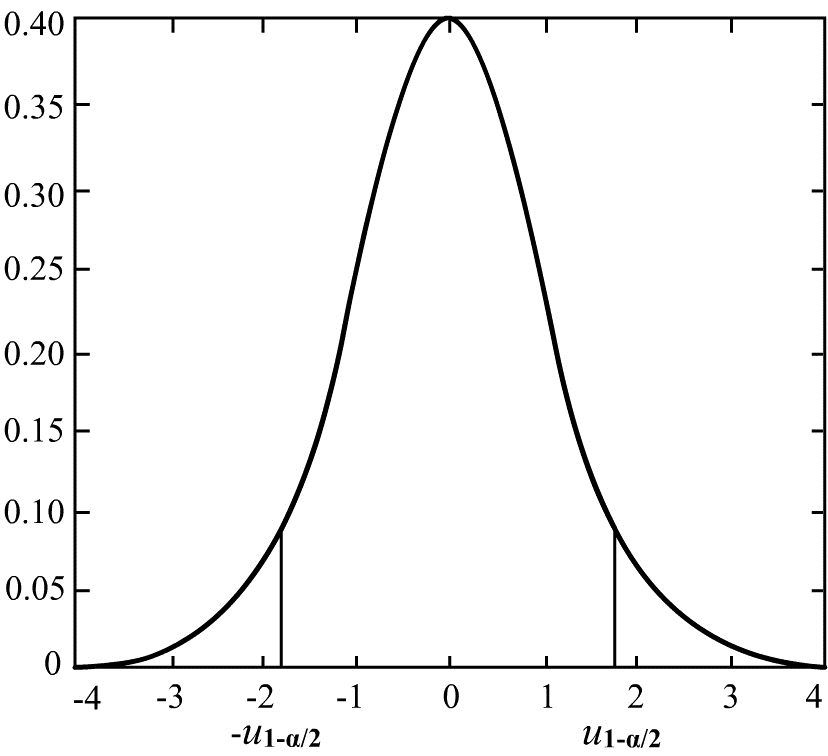


그림 6.6.3 표준정규분포의 도식

실례6.6.3 천평으로 어떤 물체의 질량을 9번 측정한 결과 평균값은 (그람)이다. 천평으로 잰 질량이 정규분포에 따르고 그 표준편차가 0.1(그람)이라면 그 물체질량의 0.95믿음구간을 구하시오.

풀기.여기서 이고 이므로 그 물체질량 의 0.95믿음구간은

.

따라서 질량의 0.95믿음구간은 [15.3347, 15.4653]이다.

실례6.6.4 모집단이 일 때 구간의 길이가 1.2를 넘지 않는 의 0.95믿음구간을 얻으려면 표본량이 얼마여야 하는가?

풀기. 문제의 조건으로부터 의 0.95믿음구간은 이다. 이 구간의 길이는 로서 표본량에만 의존하고 표본의 구체적인 값과는 관계없다. 를 풀면  이므로 이고 . 즉 표본량이 적어도 11일 때 의 0.95믿음구간의 길이가 1.2를 넘지 않는다.

2. 가 주어지지 않았을 때 의 믿음구간

이때 통계량을 리용할수 있다. 그것은 이므로 가 중심축량으로 될수 있다. 앞에서와 비슷한 방법으로 의 믿음구간을 구하면



여기서 는 의 불편추정이다.

실례 6.6.5다이야의 수명이 정규분포에 따른다고 할 때 어떤 다이야의 평균수명을 추정하기 위하여 12개의 다이야를 우연적으로 뽑아 수명(단위:1만km)을 측정한 자료가 다음과 같다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 4.68 | 4.85 | 4.32 | 4.85 | 4.61 |
| 5.20 | 4.60 | 4.58 | 4.72 | 4.38 |

평균수명의 0.95믿음구간을 구하시오.

풀기. 정규모집단의 표준편차는 미지이며 분포를 리용하여 평균의 믿음구간을 구할수 있다. 이 실례에 대하여 계산하면 이다. 를 취하고 수표를 검사하면 . 따라서 평균수명의0.95믿음구간은 이다.

실천에서 바퀴의 수명은 길수록 좋으므로 평균수명의 믿음하한을 알면 된다. 즉 단측믿음하한을 구한다. 이므로 부등식을 변환하여 의 믿음하한을 구하면 . 를 대입하여 계산하면 평균수명 의 믿음하한은 4.5806(km)이다.

3. 의 믿음구간

이 경우에 가 기지인가, 미지인가에 따라 두가지 경우로 갈라 의 믿음구간을 론의할수 있지만 현실에서 이 미지일 때 가 기지인 경우는 극히 드물다. 그러므로 가 미지인 조건에서 의 믿음구간을 론의한다.

중심축량은 쉽게 구할수 있다. 은 표본분산 으로 추정할수 있다는것을 앞에서 언급하였다. 5.4에서 이라는것을 증명하였다. 그런데 분포가 치우침분포이므로 평균길이가 최단인 구간을 구하는것은 힘들고 일반적으로 등꼬리믿음구간을 구하는 문제로 바꾸어 고찰한다. 를 두 부분으로 꼭같이 갈라 분포의 량끝자름면적이 각각 가 되도록 한다. 즉 분포의 분위수와 분위수(그림 6.6.4를 참고)를 리용한다. 그것들은 를 만족한다.

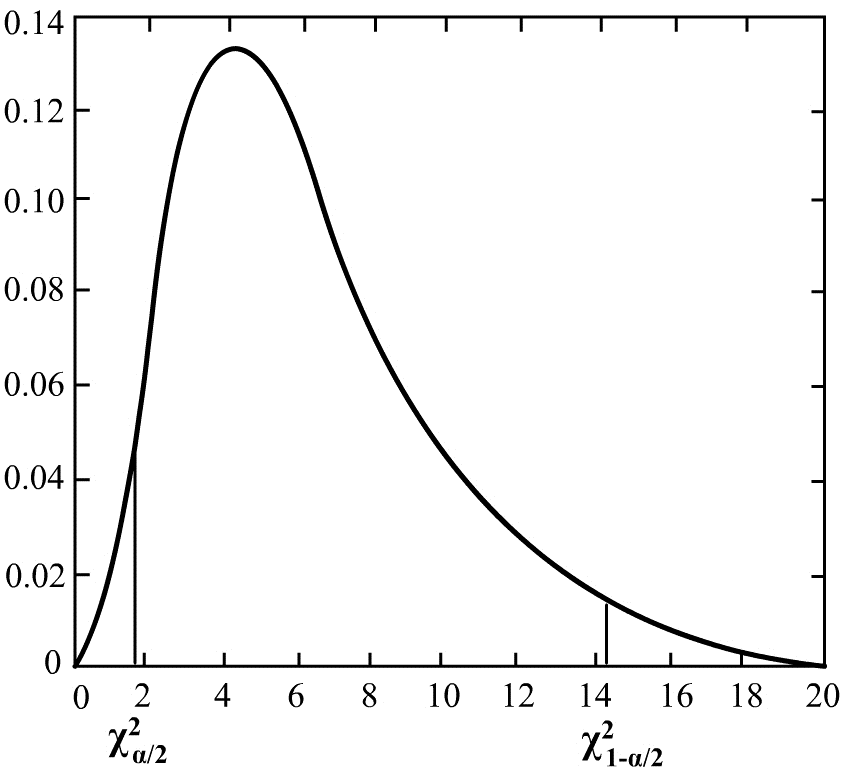


그림 6.6.4 분포믿음구간을 보여주는 도식

이렇게 의 믿음구간을 얻을수 있다.

.

식 (6.6.10)의 두변에 제곱뿌리를 취하면 표준편차 의 믿음구간을 구할수 있다.

실례 6.6.6 어느한 공장에서 생산되는 부속품질량은 정규분포 에 따른다. 현재 그 공장에서 생산하는 부속품가운데서 9개를 뽑아 그 질량(단위:그람)을 측정한 자료는 다음과 같다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 45.3 | 45.4 | 45.1 | 45.3 | 45.5 | 45.7 | 45.4 | 45.3 | 45.6 |

모집단표준편차 에 대한 0.95믿음구간을 구하시오.

풀기. 자료로부터 을 계산할수 있다. 이므로 수표로부터 . 이것을 식 (6.6.10)에 대입하면 의 0.95믿음구간은

.

따라서 의 0.95믿음구간은 [0.121, 0.345]이다.

### 6.6.4 큰표본의 믿음구간

어떤 경우에는 중심축량과 그 분포를 찾기가 어렵다. 표본량이 충분히 크면 점근분포를 리용하여 근사믿음구간을 구성할수 있다. 대표적인 실례는 비률 에 대한 믿음구간이다.

가 2점분포 의 표본일 때 의 믿음구간을 구하려고 한다. 중심극한정리로부터 표본평균 의 점근분포는 이고 따라서 .

는 근사중심축량으로 리용될수 있다. 주어진 에 대하여 표준정규분포의 분위수 를 리용하면 을 얻을수 있다. 괄호안의 사건은 와 같다. 라고 표시하면 우의 부등식은 로 변화된다. 왼변의 2차식에 대하여 판별식을 생각하면



이 2차식의 그라프가 축과 사귀는 점이 두개(그림 6.6.5을 참고)이다. 이 사귐점들의 가로자리표를 라고 표시하면 . 여기서 는 그 2차식의 두개의 뿌리로서







그림 6.6.5 2차 3항식과 그 뿌리를 보여주는 도식

이 비교적 크기때문에 실천에서는 일반적으로 을 생략하고 믿음구간을 다음과 같이 근사하게 구할수 있다.



실례 6.6.7 어떤 시행을 120번 하였는데 사건 가 36번 발생하였다. 사건 의 확률 에 대한 0.95믿음구간을 구하시오.

풀기. 여기서 이고 이므로 의 0.95(량쪽)믿음하한과 상한은 각각 다음과 같다.



따라서 근사믿음구간은 [0.218, 0.382]이다.

비슷한 방법으로 뽜쏭분포의 파라메터 의 믿음구간을 결정할수 있다(련습문제 6.6의 7번문제).

### 6.6.5 표본량 확정

통계학에서는 표본량이 클수록 추정정확도가 높지만 큰 표본량은 많은 자원과 시간을 필요로 한다. 때문에 실천에서는 일정한 요구조건하에서 적어도 얼마만한 표본량이 필요한가 하는 문제에 관심을 가진다. 이것이 바로 표본량의 확정문제이다.

표본량을 확정하는데는 여러가지 방법이 있으며 경우에 따라 서로 다른 방법을 리용한다. 여기서는 비률 를 추정하는데 필요한 표본량을 소개한다. 현실문제에서는 불합격품률, 흡연률, 출생하는 애기들가운데서 남자애의 비률, 지지률과 같이 비률을 추정하는 문제가 많이 제기된다. 이러한 경우에 적어도 얼마만한 표본량이 있어야만 얻어진 추정의 정확도를 보장할수 있는가?

구체적인 례를 들어 그 문제에 대해 설명해보자.

실례 6.6.8 어느한 방송회사에서 TV로 방영하는 한 체육프로의 시청률 를 조사하려고 할 때 의 믿음구간의 길이가 을 초과하지 않게 하려면 최소 몇명의 리용자를 조사해야 하는가?

이것은 전형적인 발취조사문제인데 믿음수준 를 **보장확률**, 믿음구간의 반경(길이의 절반) 을 **절대오차**라고 부른다.

풀기. 이것은 2점분포의 비률 의 믿음구간문제이다. 식 (6.6.11)에 의하면 의 근사믿음구간의 반경이 인데 이것은 하나의 우연량이다. 그러나 이므로 임의의 관측값에 대하여 . 따라서 의 믿음구간반경이 을 초과하지 않는다. 이제 의 믿음구간반경이 을 초과하지 않게 하려면 이면 된다. 결국



이것이 비률 를 추정할 때 표본량을 확정하는 공식이다. 실례로 라면

.

이것으로부터 체육프로시청률 의 0.95믿음구간반경이 0.02를 넘지 않기 위해서는 적어도 2401명의 시청자를 대상으로 조사해야 한다는것을 알수 있다. 다시 말하여 적어도 2401명의 시청자를 조사하여야만 얻은 추정값 와 진값 의 차이가 0.02보다 크지 않도록 0.95의 확률로 담보할수 있다.

표 6.6.1에 일반적으로 요구되는 표본량에 대한 결과를 보여주고 있다. 이로부터 믿음구간의 길이가 작고 믿음수준이 높을수록 보다 큰 표본량이 필요할수 있다는것을 알수 있다.

표 6.6.1 일부 (, )쌍에 대한 표본량

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | | | |
| 0.005 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.1 | 0.2 |
| 0.8 | 65695 | 16424 | 4106 | 1825 | 1027 | 657 | 165 | 42 |
| 0.85 | 82891 | 20723 | 5181 | 2303 | 1296 | 829 | 208 | 52 |
| 0.9 | 108222 | 27056 | 6764 | 3007 | 1691 | 1083 | 271 | 68 |
| 0.95 | 153659 | 38415 | 9604 | 4269 | 2401 | 1537 | 385 | 97 |
| 0.99 | 265396 | 66349 | 16588 | 7373 | 4147 | 2654 | 664 | 166 |

참고: 표 6.6.1의 결과는 정확한 정규분포의 분위수를 리용하는 쏘프트웨어에 의해 직접 계산된 것이므로 정규분포의 분위수근사값을 리용하는 결과와 약간 다르다.

례를 들어 믿음수준이 0.99 이고 믿음구간의 길이가 0.01을 초과하지 않으려면 표본량이 66349가 되여야 하는데 이것은 매우 큰것이다. 그러나 만일 비률 에 대해 어느 정도 알고있다면 표본량에 대한 요구를 줄일수 있다. 실례로 어떤 TV프로든 시청률이 모두 그리 높지 않다는것을 알 때(가령 ) 큰수의 법칙에 따라 라고 근사하게 생각할수 있고 그때 이므로 의 믿음구간반경 의 상한을 로부터 로 변경시킬수 있다. 이제 의 믿음구간의 반경이 을 초과하지 않게 하려면 이면 되고 따라서 을 얻는다. 이것은 표본량을 크게 줄인다. 실례로 믿음수준이 0.99일 때 믿음구간길이가 0.01을 초과하지 않게 하려면 필요한 표본량이 66349로부터 42463으로 줄어든다.

### 6.6.6 두 정규모집단에 대한 믿음구간

가 의 표본이고 는 의 표본이고 두 표본이 서로 독립이라고하자. 는 각각 그것들의 표본평균이고 , 는 각각 표본분산이다. 아래에서 두 평균값의 차와 두 분산의 비에 대한 믿음구간을 론의한다.

1. 의 믿음구간

이에 대한 몇가지 특수경우는 이미 원만히 해결되였다. 아래에서 이 문제를 몇가지 경우로 나누어 서술하는데 그것들사이의 차이와 처리방법에 주목하여야 한다.

1) 이 기지일 때

이때 이며 중심축량은 .

이전에 여러번 리용된 방법을 리용하면 에 대한 믿음구간을 얻을수 있다.



2) 가 미지일 때

이때



이 성립한다. 이 서로 독립이므로 다음과 같이 분포 에 따르는 중심축량을 구성할수 있다.



라고 표시하면 의 믿음구간은 이다.

3) 이 기지일 때

이때의 처리방법은 2)의 처리와 완전히 류사한데 다음의 관계식만 주의하면 된다.



이 서로 독립이기때문에 다음과 같이 분포 에 따르는 중심축량을 구성할수 있다.



라고 표시하면 의 믿음구간은 이다.

4) 이 모두 클 때의 근사믿음구간

만일 에 대한 정보가 없다면 이 모두 클 때 중심극한정리로부터



따라서 의 믿음구간은 이다.

5) 일반적인 경우의 근사믿음구간

에 대한 정보가 없고 도 크지 않은 경우 에 대한 정확한 믿음구간을 구하는것은 유명한 베렌스-피셔(Behrens-Fisher)문제로서 1929년에 현실에서 제기되여 현재까지도 학자들이 연구하고 있다. 여기서는 한가지 근사방법을 소개한다. 라고 하고 근사중심축량 을 취한다.

이때 는 에 따르지 않지만 근사하게 분포에 따른다. 여기서 은 로 결정된다. 이 일반적으로 옹근수가 아니므로 과 가까운 옹근수로 대치한다. 따라서 근사적으로 이고 의 근사믿음구간은 이다.

실례 6.6.9 두 밀품종의 수확고를 비교하기 위해 같은 농사방법으로 일치한 조건을 가진 18개 시험포전을 선정하여 시험하였다. 품종 A를 파종한 8개 시험포전과 품종 B를 파종한 10개 시험포전의 단위당 수확고는 다음과 같았다.(단위:kg)

품종 A: 628 583 510 554 612 523 530 615

품종 B: 535 433 398 470 567 480 498 560 503 426

매 품종의 단위면적당 수확고가 정규분포에 따른다고 할 때 두 품종의 단위면적당 수확고의 차이에 대한 믿음구간을 구하시오. 여기서 로 취한다.

풀기. 가 품종 A의 단위면적당 수확고, 는 품종 B의 단위면적당 수확고라고 하자. 표본자료로부터 계산하면



두가지 경우으로 나누어 론의한다.

① 두 품종의 단위면적당 수확고의 표준편차가 같다면 두 표본의 구간을 리용할수 있다. 이때



따라서 의 0.95믿음구간은 이다.

② 두 품종의 단위면적당 수확고의 분산이 같지 않으면 근사구간을 리용할수 있다.이때



따라서 의 0.95믿음구간은 이다.

2. 의 믿음구간

이고 이 서로 독립이므로 우연량을 모방하여 다음과 같은 중심축량을 구성할수 있다.



주어진 믿음수준 에 대하여 이므로 부등식변환을 통하여 에 대한 다음과 같은 믿음구간을 얻을수 있다.

.

실례 6.6.10 어떤 작업장에서 두대의 선반으로 부속품을 가공하는데 그 부속품의 직경이 정규분포에 따른다고 가정한다. 이제 두 선반으로 가공한 제품가운데서 각각 5개, 6개를 우연검사하여 얻은 직경(단위:cm)의 자료는 다음과 같다.

선반 A: 5.06 5.08 5.03 5.00 5.07

선반 B: 4.98 5.03 4.97 4.99 5.02 4.95

두 선반으로 가공한 부속품직경의 분산비 의 0.95믿음구간을 구하시오.

풀기. 이라고 하고 =0.95를 취하면 수표로부터

.

자료로부터 계산하면 이고 따라서 믿음구간의 량끝은 각각



따라서 의 0.95믿음구간은 [0.1574, 10.8861]이다.

련습문제 6.6

1. 어느한 공장에서 생산하는 화학섬유의 강도는 표분편차가 인 정규분포에 따른다. 이제 용량이 인 표본을 발취하여 그 강도를 측정한 다음 계산해보니 표본평균값이 였다. 이 제품무지의 평균강도에 대한 0.95믿음구간을 구하시오.

2. 모집단 에서 은 기지이다. 표본량 이 얼마여야 의 95%믿음구간의 길이가 보다 크지 않도록 담보할수 있는가?

3.모집단 로부터 0.50, 1.25, 0.80, 2.00이라는 표본을 발취하고 가 에 따른다고 할 때

① 의 95%믿음구간을 구하시오.

② 의 수학적기대값에 대한 95%믿음구간을 구하시오.

4. 한개의 측정계기로 어떤 물리적량을 9번 측정하여 표본평균 , 표본표준편차 를 얻었다.

① 측정표준편차 의 크기는 측정계기의 정밀도를 반영한다. 의 0.95믿음구간을 구하시오.

② 그 물리적량의 진값에 대한 0.99믿음구간을 구하시오.

5. 어느한 재료의 압력견딤세기는 이다. 이제 우연적으로 10개의 시료를 선택하여 시험을 진행한 결과 다음과 같은 자료를 얻었다.

482, 493, 457, 471, 510, 446, 435, 418, 394, 469.

① 평균압력견딤세기 에 대한 95%믿음구간을 구하시오.

② 이 주어졌을 때 에 대한 95%믿음구간을 구하시오.

③ 에 대한 95%믿음구간을 구하시오.

6. 한 무지의 상품에서 우연적으로 80개를 선택하여 검사하였을 때 11개의 불합격품을 발견하였다면 이 상품무지에서 불합격품률의 0.90믿음구간을 구하시오.

7. 는 뽜쏭분포 의 표본이다. 의 근사추정구간이 다음과 같다는것을 증명하시오.

.

8. 어느한 상점에서 어떤 상품의 월판매량은 뽜쏭분포에 따른다. 상점에서 상품을 합리적으로 구입하기 위해서는 판매상황을 알아야 한다. 그 상점의 지난 기간 일부 판매량에 대한 기록자료는 다음과 같았다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 월판매량 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 월수 | 1 | 6 | 13 | 12 | 9 | 4 | 2 | 1 |

월평균판매량의 믿음수준이 0.95인 믿음구간을 구하시오.

9. 모집단 과 로부터 각각 용량이 인 독립인 표본을 발취하고 계산을 통해 을 얻었다.

① 이 주어졌을 때 의 95%믿음구간을 구하시오.

② 이 주어졌을 때 의 95%믿음구간을 구하시오.

③ 에 대해 아무것도 모를 때 의 95%근사믿음구간을 구하시오.

④ 의 95%믿음구간을 구하시오.

10. 사람들의 키가 정규분포에 따른다고 하자. 두 지역 A, B에서 18~25살의 녀성들에 대한 키를 측정하였는데 지역 A에서는 10명을 검사하였는데 표본평균이 1.64m, 표준편차가 0.2m였고 지역 B에서도 10명을 검사하였는데 표본평균이 1.62m, 표준편차가 0.4m였다.

① 두 정규모집단분산비의 믿음수준이 95%인 믿음구간을 구하시오.

② 두 정규모집단평균차이의 믿음수준이 95%인 믿음구간을 구하시오.

11. 모집단 의 밀도함수가 인데 은 미지파라메터이다. 를 이 모집단의 단순우연표본이라고 할 때 의 믿음구간을 구하시오.

12. 어떤 전자제품의 수명이 밀도함수가 인 지수분포에 따른다고 하자. 이 제품무지로부터 용량이 9인 표본을 발취하고 측정한 수명(단위:1000시간)은 다음과 같다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 15 | 45 | 50 | 53 | 60 | 65 | 70 | 83 | 90 |

평균수명 의 믿음수준이 0.9인 믿음구간과 믿음상한, 믿음하한을 구하시오.

13. 모집단 의 밀도함수가 다음과 같다.

.

가 이 모집단의 단순우연표본이라면 위치파라메터 의 근사믿음구간을 구하시오.

14. 가 정규모집단 의 단순우연표본일 때 의 믿음구간이 주어진 길이 보다 길지 않게 하려면 표본량 이 최소한 얼마이여야 하는가?

15. 가 정규모집단 의 단순우연표본일 때 이 중심축량이라는것을 증명하시오. 여기서 는 기지상수이다.

16. 이 의 표본일 때 의 믿음구간을 구하시오. (풀기방향: 이 중심축량이라는것을 증명하고 대응한 밀도함수를 구하시오.)

17. 은 평등분포 의 단순우연표본이다. 를 그의 순서통계량이라고 할 때

① 의 믿음구간을 구하시오.

② 의 믿음구간을 구하시오.

18. 이고 은 미지이며 두 표본이 독립일 때 의 믿음구간을 구하시오.(풀기방향: 이라고 놓고 의 분포가 와 무관계하다는것을 증명한 다음 대응한 밀도함수를 구하시오.)

19. 모집단 의 밀도함수가 이고 는 그 모집단의 단순우연표본이다.

① 의 분포가 와 무관계하다는것을 증명하고 그 분포를 구하시오.

② 의 믿음구간을 구하시오.