학습 정리

팀	빅나물	구성원	이지현, 조영현

일정	발제자	주제
1일차(5/28)	이지현	-통계 기초(자료의 형태 및 분포)

주요 내용 요약

자료(변수)의 두 가지 형태

1) categorical(범주형) : 명목/순서
   - 도수분포표(Frequency table)
   - 막대그래프(Bar graph)
     -각 범주가 하나의 막대로 표현됨
   - 파이 차트(Pie chart)
     -각 범주는 파이의 한 slice로 표현됨
     -보통 %를 사용하여 모두 더해서 1이 되도록 함

2) quantitative(양적) : 연속/이산
   - Graphical
     -Dotplot, Stemplot, Histogram, Boxplot, … ,Line graph
     -전체적인 분포의 패턴과 그 패턴으로부터 벗어난 극단적 관측치들(outliers)을 살펴봄
   - 수치적
     -대표값(Center of distribution)
     :산술평균, 중앙값, 최빈값(범주형도 가능)
     -산포도(Spread of distribution)
     :범위, 사분위범위(IQR), … ,표준편차

대표값(center)

1) 산술평균(mean)
   -계산이 쉽고 수학적으로 다루기 쉬움
   -모든 관측치를 사용하므로 특이값에 영향을 많이 O
2) 중앙값(median)
   -관측한 자료를 순서대로 배열하여 가장 중앙에 있는 값
   -순위를 사용해 중앙에 있는 값만 사용하므로 특이값에 영향을 받지 X
3) 최빈값(mode)
   -관측치 가운데 가장 여러번 나타난 값
   -여러개 존재하거나 존재하지 않을 수 있고 중심을 잘 대변하지 못하는 경우가 많음
   -이산변수에 주로 사용하고, 범주형 자료에도 사용가능

=>분포가 한쪽으로 치우쳐있는 경우나 특이값들이 있는 경우 중앙값이 더 적합하고 그렇지 않은 경우 대부분 산술평균이 적합

산포도(spread)

1) 범위(range)
   -최대값-최소값
   -간단하지만 특이값에 큰 영향을 받음

2) 4분위 범위
-특이값에 영향 받지 않음
3) 표준편차
-가장널리 이용되며 통계적 추론에 유용
-산술평균처럼 특이값에 영향을 받음

사분위 범위(IQR)
1) 백분위수 (percentile, quantile)
: p백분위수란 p%의 관측치는 이 값 아래에 있고 나머지는 이 값보다 위에 있게 되는 값을 말함
2) 중앙값 : 50 백분위수
3) Q1=25 백분위수=제1사분위수(first quartile)
4) Q3=75 백분위수=제3사분위수(third quartile)
5) IQR = Q3-Q1
6) 다섯 숫자 요약(Five-number summary) : min Q1 median Q3 max

상자그림(Boxplot)
1) 다섯숫자 요약의 graphical result
2) 상자는 중앙 50%의 자료를 표시
3) 여러 개의 분포를 한 눈에 비교 할 때 유용함
4) 그리는 방법 :
-Q1과 Q3로 끝나는 상자를 그림(상자의 길이 = IQR)
-상자 안에 줄을 그어 중앙값을 표시
-Q3+1.5IQR보다 크거나 Q1-1.5IQR보다 작은 값은 * 또는 다른 symbol로 표시(outliers)
: “1.5IQR criterion”
-상자의 끝에서 Outlier가 아닌 값 중에 가장 큰 값과 가장 작은 값까지 줄을 그음

확률변수와 분포
1) 확률과 임의성
-어떤 현상이 '랜덤'하다는 것
현상의 개별적인 결과를 예측할 수는 없으나, 여러 번 반복 시행 시 그 결과가 규칙적인 분포를 따르게 되는 것
- 랜덤한 현상 or 실험의 어떤 결과가 나올 확률
매우 여러 번 반복 시행 시, 이 결과가 나오는 비율, 즉 상대빈도
- 확률변수
랜덤한 현상 or 실험의 결과로 결정되는 수치적인 양
매번 시행 때마다 다른 값을 가질 수 있으며 일정한 확률분포를 가짐(이산형/연속형)

2) 이산확률변수
- 이산확률변수 x는 유한, 또는 셀 수 있는 무한의 값만 가짐.
확률분포는 모든 가능한 값에 그 값이 나올 확률을 대응시키는 확률분포표나 확률 히스토그램으로 표현.

3) 연속확률변수
- 어떤 구간 안의 모든 값을 다 취할 수 있음. 확률분포는 확률밀도함수로 표현.
- 연속형의 경우 확률은 각 구간에 할당. 밀도함수 아래 면적
- 확률밀도함수는 양의 값을 가짐. 전체 구간 적분한 값=1
- 연속확률변수가 단 하나의 값을 가질 확률은 0이다.

정규분포

1) 확률과 임의성

- 정규분포는 연속형 분포 가운데 가장 많이 쓰이는 확률분포
- 정규분포~N(모평균,분산)
  μ: 분포 가운데. 분포의 위치.
  σ: 분포의 퍼짐 정도
  μ에 대칭인 종 모양 분포.
  μ±σ에서 볼록, 오목이 바뀜.
- σ가 작으면 평균 주위 가깝게, 크면 분포 넓게 퍼진 형태
- 표준정규분포: μ=0, σ=1

2) 68-95-99.7 Rule

모든 데이터가 (μ-σ,μ+σ) 안에 들어올 확률이 68%
μ±2σ: 95%
μ±3σ:99.7%

3) 정규분포의 표준화

모든 정규분포는 같은 형태적 성질을 갖기 때문에 표준화해 N(0,1) 얻을 수 있음.
표준화 후, 확률표 이용해 확률계산도 가능.
$Z=(X-\mu)/\sigma$

모집단과 표본 (표본⊂모집단)

1) 모집단(population)
   -어떤 연구에서 실제 관심 있는 집단으로 흔히 전체를 모두 연구하기 어려움
   ex) 모든인간, 전국의 모든 근로자, 전국의 모든 유권자,... ,모든 금붕어
2) 모수 (parameter)
   -모집단의 특성을 나타내는 숫자
   -미지의 고정된 상수
3) 표본(sample)
   -모집단의 일부분으로서 실제로 연구자가 자료를 수집하여 연구하는 부분
   -표본추출이 잘 되어야 연구전체가 의미 있어짐
4) 통계량(statistic)
   -표본의 특성을 나타내는 숫자
   -표본에 따라 다른 값을 갖는 확률변수
   -모수를 추정하는 데에 사용됨

표본분포(Sampling distribution)

1) 통계량의 표본분포
   -확률변수인 통계량의 확률분포
   -모집단에서 표본의 크기가 n(정해진 숫자)인 모든 표본이 뽑혔다고 가정했을 때, 각 표본에서 계산된 통계량이 가지는 값들의 분포
   -이론적인 분포이고, 실제 관측하는 분포는 아님
2) 표본평균의 분포
   -평균이 μ이고 표준편차가 σ인 모집단에서 표본크기 n인 표본을 많이 추출 했다고 가정 할 때, 어떤 표본에서의 평균은 μ보다 크고, 어떤 표본에서의 평균은 μ보다 작아지면서 표본분포가 생성 될 것임

3) 모집단의 평균이 μ이고 표준편차가 σ일 때:
   -표본평균의 평균은 모집단의 평균과 같음

μxbar = μ (불편추정량)

-표본평균의 표준편차는 모집단의 표준편차보다 작으며 표본의 크기가 증가함에 따라 1/ n½ 의 비율로 줄어듬

σxbar = σ / n½

표본분포와 중심극한정리

1) 중심극한정리

- 평균이 μ, 표준편차 σ인 모집단에서 임의 표본을 뽑을 때, 표본의 크기 n이 크면 표본평균의 표본분포는 근사적으로 정규분포를 따름. 그 평균은 μ, 표준편차는 σ/√n

2) 표본의 크기는 얼마나 커야하는가?

- 필요한 표본의 크기는 모집단의 형태에 따라 다름.
- 모집단이 정규분포와 많이 다를수록 표본의 크기는 더 커야함.
  - 모집단의 분포가 한 쪽으로 치우쳐 있고 약한 특이치들이 존재하는 경우: 25개 정도의 표본이 있으면 표본분포의 정규성을 가정할 수 있음
  - 모집단의 분포가 매우 치우쳐 있고 심한 특이치들이 존재하는 경우: 40개 정도의 표본 정규성 어느 정도 만족
  - 많은 경우 n=25~40은 아주 큰 표본의 크기가 아니므로, 대부분의 경우에 표본평균의 분포를 생각할 때 정규분포를 가정해도 큰 무리 X.

참고 사이트

러닝패킷 : 통계의 기초

http://www.kocw.net/home/search/kemVie w.do?kemId=694004