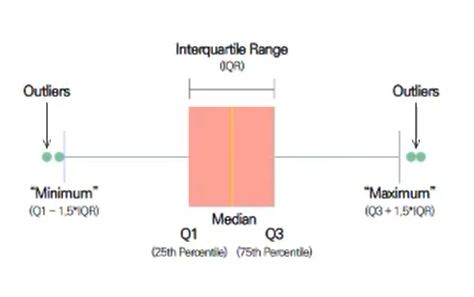
**1강. 통계학이란?**

1. 통계학이란
   1. 타당하고 신뢰할 수 있는 자료의 효율적 수집
   2. 수집된 자료를 표, 그림, 수치로 정리 분석
   3. 의사결정을 위해 분석한 결과 활용
2. 개체와 변수
   1. 개체(unit): 통계작업의 대상이 되는 관찰 및 실험단위
   2. 변수(variable): 통계쩍 관심이 되는 특성에 이름 붙인 것
   3. 변수값: 변수가 취할 수 있는 값
   4. 관측값: 개체로부터 실제 측정한 변숫값
3. 모집단과 표본
   1. 모집단(population): 통계적 관심의 대상이 되는 모든 개체의 집합
   2. 모수(parameter): 모집단의 특성을 나타내는 수 (모평균, 모분산, 모비율)
   3. 표본(sample): 모집단의 부분집합
4. 자료수집방법
   1. 관찰: 어떤 집단이나 상황에 개입하지 않고 있는 그대로를 기술
   2. 실험: 연구대상에 처리(treatment)를 가하고 그로 인한 결과를 살펴 봄
5. 교란변수(confounding variable), 교란인자: 두 변수의 관계를 교란시키는 인자
   1. 제어방법
      1. 무작위화: 각 군에 동일하게 분포하게 함
      2. 제한: 표준기준과 제외기준을 통해 표본집단에 들어올 때부터 제한
      3. 짝짓기: 교란요인을 가진 대상을 각 군에 동일하게 할당
6. 상관관계와 인과관계
   1. 상관관계(non-causal relationship): 변수 사이의 연관성 파악
   2. 인과관계(causal relationship): 원인변수(독립변수)가 결과변수(종속변수)에 미치는 영향 파악
7. 계량화 경향 🡪 통계학이 필요한 이유
8. 기술통계학과 추측통계학
   1. 기술통계학(descriptive statistics): 조사된 자료를 표나 그림, 수치를 이용하여 요약함으로써 자료 특징 파악
   2. 추측통계학(inferential statistics):: 관련된 자료를 이용하여 미지의 성질을 추론함으로써 새로운 정보 획득
9. 통계자료분석의 단계
   1. 자료의 수집 🡺 자료의 요약(기술통계학) 🡺 미지 성질의 추론(추측통계학) 🡺 의사결정

**2강. 자료의 요약**

1. 범주형 변수와 계량형 변수
   1. 범주형 변수(질적 변수)
      1. 명목척도(성, 인종, 혈액형) 🡺 부여한 숫자 의미 X
      2. 순서척도(암의 병기, 학년) 🡺 평균으로 구하기는 어려움
   2. 계량형 변수(양적 변수)
      1. 구간척도(수능 성적, 섭씨온도) 🡺 실질적 0이 없으며 배수의 개념 X
      2. 비척도(몸무게) 🡺 실질적 0점
         1. 이산형(셀 수 있음), 연속형(셀 수 없음)
2. 범주형 자료 요약
   1. 표를 이용한 요약
      1. 도수분포표 🡪 내림차순이 바람직함
      2. 분할표 🡪 두 범주형 자료 간 연관성 분석 분석
   2. 그림을 이용한 요약
      1. 막대그림(bar chart)
      2. 띠그림
      3. 버블차트
      4. 모자이크 그림 🡪 다변량 자료(multivariate data)
3. 계량형 자료의 요약
   1. 표에 의한 요약
      1. 도수분포표 🡪 누적도수도 추가하는 경우 많음
   2. 그림을 이용한 요약
      1. 히스토그램 🡪 막대그림과 달리 면적도 의미를 가짐
      2. 즐기-잎 그림(stem and leap plot) 🡪 자료의 값을 살리면서 히스토그램 효과, 자료가 너무 크면 사용하기 어려움
      3. 점도표(dot plot)
      4. 꺾은선 그래프(line graph) 🡪 변화 또는 추이를 살펴볼 때, 결측치가 있으면 기울기 왜곡 가능성
4. 이변량 자료의 요약
   1. 그림에 의한 요약
      1. 산점도(scatter plot)
      2. 버블차트 🡪 버블 크기 고려
5. 대푯값: 중심위치의 측도(measure of central tendency)
   1. 평균(mean)
      1. 산술평균(arithmetic mean) 🡪 이상값에 왜곡받기 쉬움
      2. 절사평균(trimmed mean) 🡪 양 극단점을 제거하고 평균
      3. 조화평균(harmonic mean) 🡪 평균시속, 단가 등에서 사용
      4. 가중평균(weighted mean) 🡪 각 자룟값의 중요도가 달라 그에 따라 가중값을 두어 구한 평균
      5. 기하평균(geometric mean) 🡪 평균증가율, 평균성장률, 평균이자율 등의 기하급수적인 변동에 대한 평균을 구할 때
   2. 중앙값(median) 🡪 예외적으로 큰 값들의 영향력을 배제하기 위함
   3. 최빈값(mode) 🡪 계량형 자료에서는 최빈값이 권장 안됨
6. 산포도 🡪 흩어진 정도의 측도(measure of dispersion)
   1. 분산(variance), 표준편차(standard devation)
   2. 변동계수(coefficient of variation)
      1. 자료의 상대적인 변동폭을 비교하는 측도
      2. 표준편차를 산술평균으로 나눔
   3. 범위: 최대값 – 최소값
   4. 사분위 범위: 제3사분위수 - 제 1사분위수 🡪 이상값 영향 덜 받음
7. 상자그림
   1. 중앙값에 실선
   2. 1사분위수와 3사분위수에 실선을 긋고 이를 이어주는 상자
   3. 상자 끝점에서 1.5x사분위범위(1스텝) 만큼 떨어진 값을 구해 범위 내 가장 가까운 자료값까지 선으로 연결
   4. 벗어난 것은 이상값(outlier)로 간주하여 별도표기



* 1. 두 계량형 변수 간 직선의 연관성 측도로 자료는 짝을 이룬 형태여야 함
  2. 공분산(covariance): 두 변수의 편차의 fsdasfdadff 🡪 단위 때문에 두 변수 사이의 직선관련성을 나타내는fs 측도fsda로 사용 어려움
  3. 상관계수(corrlation coefficient): 공분산을 각각의 표준편차로 나누어 단위 변화를 상쇄 🡺 대칭적인 연관이 있는 자료의 상관계수는 0이므로 직선의 연관성만 파악하는 척도로만 사용

3장. 확률

1. 기본개념
   1. 확률:추론이 얼마나 믿을만한가 제시하는 근거
   2. 표본공간: 시행에서 나올 수 있는 모든 결과들의 집합
   3. 사건(event) 표본공간의 부분집합
   4. 확률의 공리적 정의
      1. 확률은 0과 같거나 0보다 큰 값을 갖음
      2. 전체에 대한 확률은 1
      3. 서로 배반인 사건들의 합사건의 확률은 각 사건에 대한 확률의 합과 같음
   5. 조건부확률: 어떤 조건이 주어지거나 사전정보가 있는 경우 그것을 감안할 수 있는 확률의 개념
2. 독립사상(independent envents)
   1. 곱사상의 확률이 각 사상의 확률의 곱이 성립이 안되면 두 사상은 독립
   2. 복원 추출: o
   3. 비복원 추출: x
3. 전확률정리
   1. 분할: 표본공간을 겹치지 않게 자르는 개념
4. 베이즈 정리
   1. 분할된 표본공간 S 위에 임의의 사상 A가 주어졌을 때, 특정 하위사상에서 발생했을 확률
   2. 전확률공식으로 분모의 확률값 계산

4장. 확률변수와 확률분포

1. 확률변수
   1. 표본공간의 각 원소에 하나의 수치를 부여함(하나의 함수임)
   2. 이산확률변수🡺 개별값에 0이 아닌 확률값이 대응될 수 있음
   3. 연속확률변수🡺 개별값의 확률은 모드 0이고 구간에 대한 확률만 의미 있음
   4. F(X): 확률변수 X의 확률질량함수, 확률변수의 확률분포를 나타내는 함수
2. 확률변수의 기댓값(expectation)
   1. 평균: 확률분포의 (무게)중심
   2. E(a) = IntegralXf(x)
   3. 기대값의 성질
      1. E(a) = a
      2. E(X + b) = E(X) + b
      3. E(aX) = aE(X)
      4. E(aX + b) = aE(X) + b
3. 분산
   1. 확률분포가 흩어진 정도를 나타내는 측도
   2. E(x^2) – m^2
   3. 분산의 성질
      1. Var(a) = 0 = E(0) = 0
      2. Var(X + b) = Var(X)
      3. Var(aX) = a^2\*Var(X)
      4. V(aX + b) = a^2\*Var(X)
4. 베르누이 분포
   1. 베르누이 시행: 시행의 결과가 두가지만 있는 시행
   2. 베르누이 확률변수: 베르누이 시행에서 성공을 1, 실패를 0으로 두는 확률변수
   3. 베르누이 확률분포: x: 0 🡪 f(x): 1-p / x: 1 🡪 f(x): p
   4. 평균: E(X) = p
   5. 분산: Var(X) = p(1-p)
5. 조합(Combination): n개중에서 r개를 뽑는 경우의 수
   1. 텍스트이(가) 표시된 사진

      자동 생성된 설명텍스트, 장치, 게이지이(가) 표시된 사진

      자동 생성된 설명
6. 이항분포
   1. 성공률 p인 베르누이시행을 독립적으로 n회 반복시행하여 그 중 성공의 횟수를 확률변수 Y
   2. n번을 시행하여 y번 성공할 확률 g(y) = nCy \* P^y \* (1-p)^(n-y)
   3. 이항확률변수 평균: np
   4. 이항확률변수 분산: np(1 – p)
   5. 예: 동전을 5번 던져서 나온 앞면의 횟수 🡺 Y ~ B(5,1/2) 🡺 g(y) = 5Cy(1/2)^y(1/2)^(5-y), y = 0, 1, 2 … 5
7. 정규분포(normal distribution, 연속확률분포)
   1. Gauss 분포라고도 함
   2. 텍스트이(가) 표시된 사진

      자동 생성된 설명
   3. 평균이 크면 오른쪽으로 이동, 분산이 작으면 뾰족(첨도가 높다)
   4. 95%관찰점들은 평균으로부터 2\*편차 이내에 있음
   5. 비모수적 검정을 사용할 때
      1. 표본의 크기가 작음
      2. 측정값이 정규분포를 따르지 않음
      3. 자료가 순위변수일 때
8. 표준정규분포(standard normal distribution)
   1. 표준화: (자료 – 평균) / 표준편차 🡺 표준 정규분포 N(0, 1)을 따른다
   2. -2 ~ 2 🡺 95%
   3. 표준정규분포에 적합한 데이터인지 확인 🡺 QQ plot 🡺 직선 형태 따라야 함

**5장. 표집분포**

1. 기본개념
   1. 추론: 통계량으로 모수 계산 🡸 근거: 표집분포
   2. 확률표본(random sample): 크기 n의 확률분포 🡺 Xi들은 모집단과 동일한 분포를 가지면서 서로 독립인 확률변수
   3. 통계량(statistic)
      1. 확률표본이 주어지면 관측 가능한 함수(평균, 분산, 비율)
      2. 미지의 모수가 포함된 변수는 통계량 X
2. 표본평균의 분포
   1. 표본평균의 표본크기가 클수록 모평균에 가까워짐
   2. 표본분산 = 모분산/표본크기 🡺 표본크기가 클수록 신뢰성이 올라간다
   3. 표본크기가 충분히 크면 표본은 정규분포를 근사적으로 따른다 🡺 중심극한의 정리
3. 표본비율의 분포
   1. 표분비율 = Y / n
   2. N이 충분히 크면 표준정규분포를 따른다 (np 5이상, n(1-0) 5 이상)
4. 표본분산의 분포
   1. 카이제곱분포: 표준정규확률변수들을 제곱해서 더한 제곱합이 따르는 분포
   2. 표본분산(S^2)에 대해서 ( (n-1)\*S^2 )/모분산 ~ 카이제곱분포를 따름
5. 정규모집단이 아닐 때
   1. 크기가 n인 random sample을 뽑으면 X바는 근사적으로 정규분포 따름 (중심극한정리)
6. T 분포
   1. 모표준편차를 모르는 경우 표본표준편차를 사용하게 되는데, 이 경우 표준정규분포보다 꼬리가 두꺼운 t 분포를 따른다 (패널티)
   2. 표준정규분포와 편차만 다름 (T 분포 분산이 항상 큼)
   3. 자유도가 커지면 표준정규분포와 가까워짐
   4. T분포를 읽으려면 자유도도 고려해야함
7. 요약테이블이(가) 표시된 사진

   자동 생성된 설명
8. F 분포
   1. 확률분포 U와 V가 서로 독립이고 카이제곱 분포에 따르면 F = (U/m) / (V/n)
   2. 분산분석 ANOVA(analysis of varience)에서 사용
   3. F분포표는 두 확률분포의 자유도, 확률을 고려해야함 (3차원)

**6장. 추정**

1. 점추정
   1. 점추정: 모수를 하나의 값으로 추정
   2. 점추정량이 가져야할 바람직한 성질
      1. 불편성(unbiasedness): 추정량이 추정하고자 하는 모수의 참값을 중심으로 분포하는 성질 (E(t^) = t) ⬄ 편의 추정량
      2. 효율성(efficiency): 분산이 작은 추정량이 더 좋음
      3. 일치성(consistency): 표본의 크기가 커지면 점추정량이 추정하고자 하는 모수로 수렴하는 성질
      4. 충분성(sufficiency): 표본의 정보를 모두 활용하는 성질(중간값은 충분통계량이 아니라고 볼 수 있음)
   3. 표본평균, 표본비율, 표본분산은 점추정량의 바람직한 성질을 모두 갖춘 좋은 점 추정량
2. 구간추정
   1. 신뢰도, 신뢰수준: 모수를 구간 내의 포함시킬 가능성 🡺 신뢰구간(confidence interval)
   2. 좌우대칭인 경우 양쪽에 똑같이 꼬리확률을 설정하고 계산
   3. 구간추정 방법
      1. 가장 좋은 점추정량 선정(ex, 모평균이 궁금할땐 표본평균)
      2. 점추정량과 모수로 이루어진 함수 설정 🡪 잘 알려진 분포
         1. 분포가 알려진 함수
         2. 미지의 모수는 추정하려는 것만 포함
   4. 모평균 추정
      1. 정규모집단에서 모분산을 알 때 (드문 경우)
         1. 모분산 사용
         2. 표준정규분포 따름
      2. 정규모집단에서 모분산을 모를 때 (현실에서 많이 사용)
         1. 표본 분산 사용
         2. t분포 따름
         3. 신뢰구간 길이 길어짐
      3. 임의의 모집단일 때, 모분산 모름, 대표본 (n 30이상)
         1. 표준정규분포 따름
         2. 표본분산은 모분산으로 수렴
      4. 텍스트이(가) 표시된 사진

         자동 생성된 설명
   5. 모비율 구간추정
      1. N이 충분히 클때만 가능 (np >= 5 && n(1-p) >= 5) 🡺 p를 P로 두고 품
   6. 모분산 구간추정
      1. 점추정량 S^2 선택
      2. 카이제곱분포
3. 신뢰구간의 의미
   1. 신뢰구간과 표집오차
      1. 퍼센트와 퍼센트포인트
         1. 퍼센트: 50%에서 10%증가 🡺 55%
         2. 퍼센트포인트: 50%에서 10%P 증가 🡺 60%
      2. 신뢰구간이 겹치는지 따짐 🡺 유의한지 확인 (겹치면 유의미한 차이가 없다고 봐야함)
      3. 표집오차 = 표본오차 = (허용)오차범위 = (허용)오차한계
   2. 신뢰도 95%의미 🡺 100번 반복샘플링을 통해 얻은 결과 중 5개는 모수를 포함하지 않음
   3. 모평균 추정에서 표본의 크기: n >= (Z(a/2)\*시그마 / d)^2 🡺 표본오차를 d 이내로 하려면 최소 n명 이상이어야 함
   4. 모비율 추정에서 표본의 크기: n >= P(1-P)\*(Z(a/2)/d)^2
      1. 표본비율을 모르면 P(1-P)의 최대값을 써야함 (1/4)
      2. 따라서 : n >= 1/4\*(Z(a/2)/d)^2
      3. 문제 예시: 재구입의사가 있는 고객의 비율을 파악하고자 할 때 신뢰도 95% 오차한계 3%p 이하로 하려면 표본 크기는 얼마?

**7장. 가설 검정**

1. 가설검정
   1. 귀무가설(H0, null hypothesis, 영가설): 기존의 주장
   2. 대립가설(H1, alternative): 연구자가 입증하고자 하는 새로운 주장
   3. 검정결과
      1. 귀무가설 참 / 귀무가설 채택 🡺 옳은 결정
      2. 귀무가설 참 / 대립가설 채택 🡺 제 1종 오류 (알파) (더 피해가 큼)
      3. 대립가설 참 / 귀무가설 채택 🡺 제 2종 오류 (베타)
      4. 대립가설 참 / 대립가설 채택 🡺 옳은 결정
   4. 제 1종 오류, 제 2종 오류 동시에 줄이는 방법 없음 🡺 제 1종 오류를 제한하는 방법 🡺 가설검정 방법
   5. 유의수준: 제 1종 오류를 범할 확률의 최대 허용 한계
   6. 유의확률(p value) <= 유의수준 이어야 통계적으로 유의
   7. 가설검정 종류
      1. H0: M < 7.4, H1: M > 7.4 🡺 단측검정(one sided test
      2. H0: M > 3.1, H1: M < 3.1 🡺 단측검정
      3. H0: M = 100, H1: M != 100 (M > 100 or M < 100) 🡺 양측검정 (tow\_sided test)
   8. 가설채택 기준
      1. 귀무가설채택
         1. 귀무가설을 입증할 뚜렷한 근거가 자료에 있음
         2. 귀무가설과 대립가설 중 어느 쪽에도 근거가 없음
      2. 대립가설채택
         1. 대립가설을 입증할만한 뚜렷한 근거가 있음
      3. 검정력(Power of test): H1이 참일 때 H1 채택
   9. 가설검정의 절차
      1. 귀무가설과 대립가설 설정
      2. 검정통계량의 함수가 잘 알려진
      3. 유의수준하에서 H0 기각역 설정
      4. 표본자료로부터 H0 기각여부 결정
2. 모평균에 대한 가설검정
   1. 정규모집단에서 모분산을 알때
      1. 단측검정: H0: M = M0 vs H1: M > M0
      2. (X바 – M)/시그마/루트n ~ N(0, 1) 🡺 시그마, n 문제에서 줌
      3. 유의수준하에서 H0 기각역: a = Pr(H1 채택 | H0 참)
      4. (X바 – M)/시그마/루트 > Z0.05 일때 H0 기각, H1 채택
   2. 정규모집단에서 모분산을 모를 때 🡺 t분포 이용, 시그마 대신 S 이용
   3. 임의의 모집단에서 모분산을 모를 때 🡺 n이 크다면 중심극한정리에 의해 시그마대신 이용, 표준정규분포
3. 모비율에 대한 가설검정
   1. (P헷 – P)/(sqrt(p(1-p)/n)) ~ N (0, 1), P: 모비율
4. 모분산에 대한 가설검정
   1. (n-1)\*표본분산/모분산 ~ 카이제곱분포

8장. t검정

1. 대응 t 검정
   1. 차이가 0인지 아닌지 검정하는 방법
   2. 귀무가설: 사전 대비 사후 값 차이가 없다
   3. 대립가설: 사전 대비 사후 값 차이가 있다.
   4. t분포, 표본평균, 표본편차 사용
   5. 결국 차이를 보는 것이기 때문에 일표본 t검정과 같음
2. 독립 t검정
   1. 서로 독립인 두 모집단에서 자료를 각각 뽑음
   2. 두 확률표본은 독립
   3. H0: M1 = M2, H1: M1 ~= M2 🡺 M1 – M2에 대한 가설검정
   4. E(X – Y) = M1 + M2
   5. Var(X – Y) = Var(X) + Var(Y) 🡺 합동분산 🡺 패널티로 t분포 따름 🡺 등분산성 가정이 필요함.
   6. 자유도: n1 + n2 – 2
   7. 분산이 같지 않다면 Satterthwaite의 formula를 이용해 자유도를 보정 🡺 항상 등분산일 때 자유도보다 작음 🡺 꼬리가 두꺼워져서 H0을 기각시키기가 어려워짐

9장. 일원배치 분산분석

1. 개념: 서로 독립인 세 집단 이상의 평균차이 모수적 검성
2. 가설검정
   1. H0: M1 = M2 … = Mk
   2. H1: notH0 (적어도 1개)
   3. 기본가정
      1. 서로 독립
      2. 정규분포
      3. 분산 같음
   4. 분산분석(ANOVA: analysis of variance): 총 변동을 인자(factor)에 의한 변동과 오차에 의한 변동으로 분해함으로써 인자의 효과 검증
   5. 변동해 분해
      1. 전체 편차(SST) 🡺 오차 에 의한 편차(SSE) + 인자에 의한 편차(SSTrt)
      2. 오차에 의한 편차의 비중이 적을수록 인자의 효과는 유의함
   6. 제곱합의 분해 🡺 F 분석
      1. 테이블이(가) 표시된 사진

         자동 생성된 설명
      2. Df: 자유도, K: 인자 수
      3. F가 클수록 통계값이 유의
      4. 결정계수 🡺 총제곱합 중 처리제곱합의 비율로 모형에 의해 설명되는 부분의 비율 🡺 1에 가까울수록 통계값 유의
   7. 다중비교(multiple comparison)로 인한 사후검정
      1. 귀무가설이 기각되었다면 각 모집단끼리 가설검정을 해야 함
      2. 검정하고자 하는 가설의 개수가 늘어나면 제 1종 오류를 범할 확률도 늘어남 🡺 사후검정 필요
      3. 사후검정: 전체 분석 결과에 대한 유의수준을 0.05로 조정
      4. 사후분석 기법: Bonferroni, Scheffe, Tukey, Duncan 등

10장. 카이제곱 검정

1. 귀무가설: 두 변수는 독립
2. 독립가설: 두 변수는 종속
3. 관측값과 기대빈도의 차이가 클수록 카이제곱 검정 통계값은 크게 나타남
4. 카이제곱 합: 기대값과 관측값 차를 제곱하고 오차로 나눈 것을 다 더함

11장. 단순선형회귀분석

1. 회귀분석
   1. 두 연속형 변수간의 연관성을 밝히는 기법
   2. 먼저 산점도를 그려 탐색적 자료분석을 해야 함
   3. 목적: 회귀계수 추정
   4. 최소제곱추정
   5. 변동의 분해: 총 제곱합 = 회귀제곱합 + 오차제곱합
2. 잔차분석 🡺 회귀분석 진단
   1. 원자료값 – 예측한 값
   2. 더빈 왓슨 통계
3. 표준화 잔차값이 +- 2안에 들어오면 잘 맞는거