1. Introduction

비모수 방법은 모수에 (parameter) 대한 어떠한 가정(assumption) 없이 통계 가설 검정하는 것이다. 모수에 대한 가정이란 일반적으로 모집단 분포에 대한 가정을 의미한다. 가설검정을 위해 계산되는 통계량은 순서(rank) 통계량이고, 가장 많이 사용되는 통계량은 중앙값(median)이다. 비모수의 경우 통계량에 대한 분포가 따로 주어지거나 p-value 개념을 사용한다. 다음 용어들은 비모수와 같은 개념으로 사용된다.

distribution free, assumption free, statistical inference based on ranks

1.1. Nonparametric?

John Arbuthnot에(1710) 의해 처음 소개된 후 1942년에 Wolfowitz가 비모수라는 용어를 사용할 때까지 큰 발전이 없었으나 그 후 물리학, 생화학, 사회 과학 분야에서 널리 사용되기 시작했다. 비모수 방법은 모수에 대한 가정이 성립하지 않거나 표본이 적은 경우 주로 사용되는 방법이다.

1.1.1. advantage (장점)

- 모집단 분포에 대한 가정에 불필요하여 잘못된 결론을 내릴 가능성이 적다.
- 계산이 용이하고 사용 시 수리적, 통계적 배경이 많이 필요하지 않다.
- 통계량이 순서(rank)에 의해 계산되므로 측정 도구의 정확성이 요구되지 않는다.
- 표본을 많이 추출할 수 없는 경우 사용 가능하다.

1.1.2. disadvantage (단점)

- 모수적 방법에 비해 얻는 정보가 제한적이고 결과 해석이 어렵다. 그러므로 모수적 방법도 사용 가능하다면 모수적 방법을 사용하는 것을 권한다.

1.2. statistical terms

1.2.1 descriptive and inferential (기술 통계학과 추론 통계학)

자료부터 정보를 얻는 과정에 관한 학문이 통계학은 자료 수집 방법, 수집된 자료를 그래프나 숫자로 요약 정리하는 기술(descriptive) 통계학과 자료에서 계산된 통계량(statistic)을 이용하여 모수(parameter)를 추정하거나 검정하는 추론(inferential) 통계학으로 나뉜다. 비모수 방법은 추론 통계의 한 분야이다.

1.2.2. population and sample (모집단과 표본)

자료 분석에서 관심의 대상이 되는 사람이나 개체들 전체를 모집단(population)이라 하고 분석을 위하여 추출한 일부를 표본(sample)이라 한다. 우리나라 대학생들의 신체적 특성에 관심이 있는 경우 모집단은 우리나라 대학생이 된다. 물론 모집단의 구성원은 명확히 정의되어야 한다. 예를 들면, 2000년 1학기 등록 학생 등…

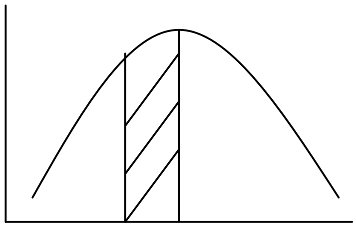
모집단 전체를 조사하는 것이 때로 시간, 비용면에서 비효율적이거나 불가능한 경우 적절한 표본 추출 방법을 이용하여 모집단 구성원의 일부만을 추출하여 분석한다. 이 모집단 일부를 표본(sample)이라 한다. 비모수 방법은 표본 자료의 크기가 작은 경우 (소표본, n≤20) 주로 이용된다.

1.2.3. parameter and statistic (모수와 통계량)

모집단의 확률 분포 함수의 형태를 결정하는 값을 모수(parameter)라 한다. 모평균 μ, 모집단 표준 편차 σ, 모집단 상관 계수 ρ 등이 모수의 예이다. 모수는 미지(unknown)이므로 이 값은 통계량(statistic)에 의해 추정(estimation)된다. 통계량은 표본 자료로부터 계산된 값으로 표본 평균 $\bar{x}$, 표준편차 s 등이 그 예다. 비모수 방법에서 가장 관심을 갖는 통계량은 중앙값(median)이다. 중앙값은 순서통계량으로부터 구해지므로 비모수 방법을 rank에 기반을 둔 추정 방법이라 한다.

1.2.4. random variable and probability density function (확률 변수와 분포 함수)

단순 임의 추출(simple random sample) 방법으로, 즉 모집단 개체들이 추출될 확률이 모두 같은 경우, 측정 결과들의 모임을 확률 변수(random sample)라 한다. 변수라는 의미는 각 개체에 따라 측정 값이 달라지기 때문이다. 확률 변수가 가지는 값들에 대한 정보(확률)를 나타내는 것이 확률 분포 함수이다. 예를 들면 몸무게 측정치 확률 변수는 정규 분포를 따른다고 가정하게 되는데 이때 평균이 70이고 표준 편차가 10인 것을 알고 있다면 몸무게가 60이상이고 70미만이 확률을 구할 수 있게 된다.



이처럼 확률 분포 함수를 알면 통계적 가설 검정이 가능하게 된다. 비모수 방법은 통계량(확률 변수)의 확률 분포 함수를 알지 못해 모수 방법을 사용하지 못하는 경우 사용하는 방법이다.

1.2.5. measurement and categorical (측정형 변수와 분류형 변수)

측정할 수 있거나 셀 수 있는 측정형(measurable 혹은 numerical) 변수와 개체나 집단을 분류하는데 사용되는 분류형(categorical) 변수로 나누어진다.

측정형 변수는 측정 도구를 이용하여 측정할 수 있는 키, 몸무게, IQ, 혈압이나 셀 수 있는 자료 자녀 수, 9시에서 10시 사이에 A국도를 이용하는 자동차 수 등을 측정형 변수라 한다. 분포 함수 이론에서는 키, 몸무게 같이 측정 가능 값이 연속일 경우 연속형(continuous) 변수, 자녀 수나 IQ와 같이 불연속 값을 갖는 경우 이산형(discrete) 변수로

구별하기도 한다. 측정형 변수를 구간(interval)과 비율(ratio)로 나누기도 한다. 몸무게 100(kg)은 50의 두 배라고 말할 수 있으나 오늘 기온이 30도라 해서 15도보다 두 배 더 더울 것이라고 말하는 것은 의미가 없다. 몸무게나 키처럼 어느 값이 다른 값의 몇 배라고 표현할 수 있는 변수를 비율 단위 측정형 변수라 하고 그런 표현이 무의미한 경우를 구간 단위 측정형 변수라 한다. 그러므로 구간 측정형에서는 0이 의미가 없게 된다. 구간 단위 측정형 변수의 대표적인 예는 특정 제품에 대한 소비자 만족도나 선호도와 같이 개인의 평가 성향, 기준의 불분명 등으로 인하여 측정 불가능한 경우 사용하는 설문 조사의 Likard 척도 (5점 척도)이다.

성별, 혈액형, 인종, 직업, 수종, 학년, 일의 숙련 정도 등이 관심의 대상을 구별하기 위해 사용되는 변수를 분류형 변수라 한다. 분류형 변수는 순서형 변수와 분류형 변수로 나눠는데, 변수가 가질 수 있는 값들이 크기에 따른 순서를 갖는다면 이를 순서형(ordinal) 변수라 하고 그렇지 않은 경우를 명목형(nominal) 변수라 한다. 성별, 인종, 직업 등은 명목형 변수이며 A, B, C, D, F로 나눈 문자 학점과 상, 중, 하로 분류한 소득 수준은 순서형 변수이다.

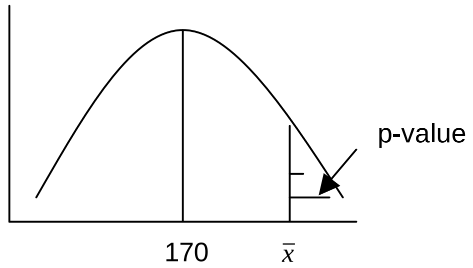
비모수 방법은 측정형 변수에 대한 분석 방법이고 분류형 변수의 경우는 비율에 대한 가설 검정에만 이용된다.

1.2.6. statistical hypothesis (통계적 가설)

- 통계적 가설: 모집단에 대해 알고자 하는 것에 대한 서술을 가설(hypothesis)이라 하고 이를 통계적으로 옳고 그름을 판단할 수 있도록 설정한 것을 통계적 가설이라 한다. 통계적 가설은 귀무 가설과 대립 가설로 나뉜다. 귀무 가설은 일반적으로 "차이가 없음", "영향을 미치지 않음", "모평균=100" 등으로 표시된다. 그러므로 우리가 선택하고자 하는 가설은 대립 가설이 되고, 이를 연구 가설이라고도 부른다.
- 유의 수준: 가설 검정의 결과는 귀무가설을 채택(대립 가설 기각) 혹은 기각(대립 가설 채택) 두 가지로 나타나므로 아래와 같은 표를 만들 수 있다. 두 오류를 최소화하는 검정 방법을 찾아야 하지만 실제 두 오류를 동시에 최소화할 수 있는 방법은 존재하지 않는다. 그리하여 통계적 가설 검정에서는 1종 오류를 (type I error) 고정시키고 2종 오류를 최소화하는 검정 방법을 사용한다. 가설 검정하기 전에 분석자가 임의로 설정한 1종 오류의 값을 유의 수준이라 (significant level) 하고 α로 표시한다. $(1-\alpha)$를 신뢰 수준이라 (confidence level) 한다.

실제 상황 / 검정 결과	귀무가설(참)	대립가설(참)
귀무 가설 채택	옳은 판단	2종 오류(β)
귀무가설 기각	1종 오류(α)	옳은 판단

▪ **p-value: 자료로부터 구해진 통계량에 의해 귀무가설을 기각할 최소의 유의 확률로 관측된 유의 수준이라** (observed significant level) **한다. 그러므로 p-값이 설정한 유의 수준보다 크다면 귀무 가설을 채택하고 작으면 귀무가설을 기각하게 된다.**



▪ **검정력: 설정된 기각역에 의해 계산된** (1-β)**가 검정력**(power)**이다. 즉, 모수 값에 따라 대립 가설이 참일 경우 대립 가설을 채택하는 확률을 검정력이라 한다. 우리의 관심이 대립 가설에 있으므로 대립 가설이 참일 경우 채택하는 것에 관심을 갖게 된다. 모평균 검정 시 기각역을** {$\bar{x}$>120}**이라 하면 검정력은** $\Pr(\frac{\bar{x}-u}{s/\sqrt{n}} < \frac{120-u}{s/\sqrt{n}})$ **이고 모수**(μ)**에 따라 각각 계산된다.**

▪ **가설 검정과 신뢰구간: 가설 검정과 양측 신뢰구간은 일대일 대응관계에 있다. 계산된 통계량과 그 통계량의 분포에 의해 계산된** 95% **양측 신뢰구간은 유의 수준** 5%**로 하여 가설 검정을 하는 경우 귀무 가설이 기각되지 않는 모수들의 모임과 일치한다. 예를 들면, 모평균에 대한** 95%**신뢰 구간이** (100, 110)**인 경우 귀무가설** $H_0 : m = 105$ **는 유의수준** 5%**하에서 기각되지 않는다.** 95% **신뢰구간의 의미는 모수가 신뢰구간에 포함될 확률이** 95%**라는 것이 아니라 자료로부터 구한 통계량으로부터 신뢰 구간들 중** 95%**는 모수를 포함하고 있을 것이다.**

📖 **Homework #1 [due 3월 8일]**

1) **한남대학교 신입생의 키가 170이상인가를 알아보기 위하여 한남대학생 중 100명을 무작위로 추출하여 키를 측정한 결과 평균 172cm, 표준편차 10cm이었다. 유의수준 5%에서 통계 가설을 검정하려고 한다.**
   - **이 경우 모수와 통계량은 무엇인가?**
   - **통계적 가설을 설정하시오.**
   - **p-value를 (관측된 1종 오류 값) 구하고 가설 검정 결과를 해석하시오.**
   - **표본 평균이 172cm 이상이면 귀무 가설을 기각하려 한다. 모수가 171, 172, 173인 경우 검정력을 구하시오.**

- 표본 평균이 171.65cm 이상이면 귀무 가설을 기각하려 한다. 모수가 171, 172, 173인 경우 검정력을 구하시오.
- 95% 신뢰구간을 구하고 위의 검정 결과와 비교하시오.

2) 한남대학교 신입생의 남녀 비율이 같은가를 알아보고자 하여 20명을 무작위로 추출하여 조사하였더니 남자 비율이 0.6이었다.
   - 이 경우 모수와 통계량은 무엇인가?
   - 통계적 가설을 설정하시오.
   - p-value를 구하고 가설 검정 결과를 해석하시오.

1.3. order statistic

1.3. order statistic (순서통계량)

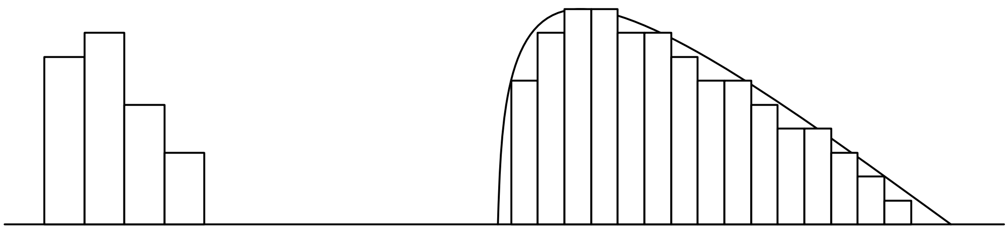
1.3.1. review

확률변수: 실험에서 발생할 수 있는 결과와 숫자를 대응시키는 것을 확률 변수라 (random variable) 한다. $X(c) = x$: c=실험 결과, x=실수 (real number)

- (예) 주사위 하나 던지는 실험: $X(c) = i \text{ where } i = 1, 2, 3, 4, 5, 6$

확률 분포 함수: 확률 변수의 값들에 확률을 대응시켜 주는 함수 관계 확률 분포 함수 (probability density function)이라 한다.

- (예1) 주사위 하나 던지는 실험: $f(x) = 1/6 \text{ where } x = 1, 2, 3, 4, 5, 6$
- (예2) 연속 확률 분포 함수 $f(x) = \dfrac{1}{\Gamma(r/2)2^{r/2}} x^{(r/2-1)} e^{-x/2)}$



누적 분포 함수: (cumulative distribution function) $F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(x)dx$

확률 표본: 동일한 확률 분포 함수로부터 독립적으로 추출한 확률변수 $X_1, X_2, ..., X_n$ 들을 확률 표본 (random sample)이라 한다. 표본 추출방법에서 단순임의 추출에 의해 측정된 관측치들이 확률 표본의 예이다.

1.3.2. definition

크기가 n인 확률 표본 $X_1, X_2, ..., X_n$ 을 크기 순으로 나열한 후 가장 작은 값을 $X_{(1)}$, 그 다음 작은 값을 $X_{(2)}$ … 이렇게 지정하면 $X_{(1)} < X_{(2)} < ... < X_{(n)}$ 가 성립한다. $X_{(1)}, X_{(2)}, ..., X_{(n)}$ 을 순서 통계량 (order statistic) 이라 한다.

1.3.3. 기초 통계량과의 관계

① 최소값 (minimum) $X_{(1)}$, 최대값 (maximum) $X_{(n)}$

② 범위 (range) $X_{(n)}$ - $X_{(1)}$, midrange: $[X_{(1)} + X_{(n)}]/2$

③ 중앙값 (median) m=$[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}]/2$ (n이 짝수), m=$X_{(\frac{n+1}{2})}$ (n이 홀수)

1.3.4. 순서 통계량 분포

1) $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$의 결합 밀도 함수 (joint distribution function)

$$f(x_1, x_2, \ldots, x_n) = n!\, f(x_1) f(x_2) \cdots f(x_n)$$

2) $X_{(1)}$의 marginal distribution function

$$f(x_1) = n[1 - F(x_1)]^{n-1} f(x_1)$$

3) $X_{(n)}$의 marginal distribution function

$$f(x_n) = n[F(x_n)]^{n-1} f(x_n)$$

4) n이 짝수인 경우 중앙값(median) m의 marginal distribution function

$$f(m) = \frac{(2k)!2}{[(k-1)!]^2} \int_m^\infty [F(2m-v)]^{k-1} [1-F(t)]^{k-1} f(2m-t) f(t) dt$$, n=2k

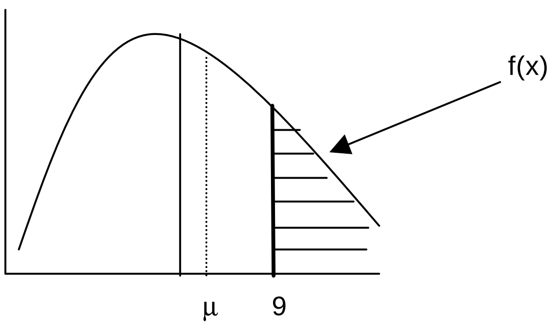
모집단 분포 $f(x) = 2x,\ \ 0 < x < 1$로부터 크기 4의 확률 표본 x_1, x_2, x_3, x_4를 추출했을 경우 $x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}$의 joint pdf, 최소값 $x_{(1)}$, 최대값 $x_{(n)}$의 marginal pdf를 구하시오.

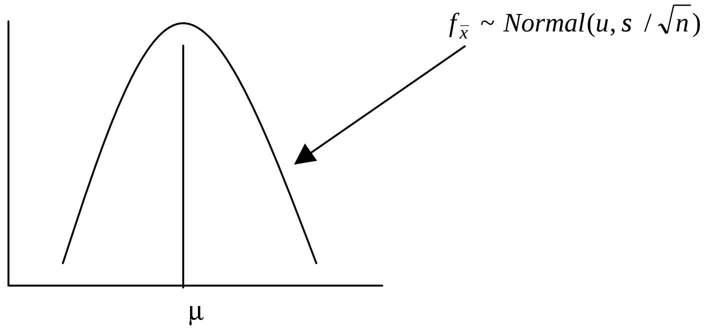
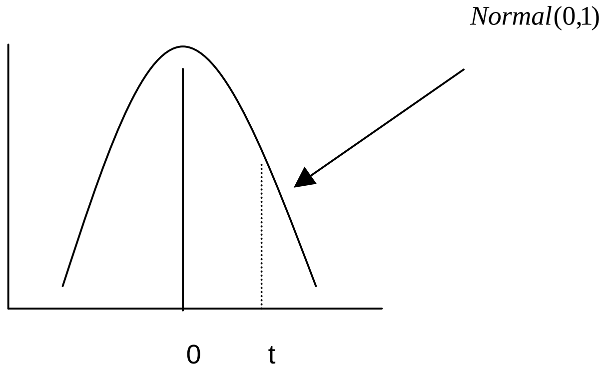
1.4. 언제 비모수 방법을 사용할 것인가? (예제)

one population + location parameter

60세 이상 노인이 일년에 약국을 찾는 횟수에 관심이 있어 자료 조사를 실시하려 한다.

	자료분석	수리 통계
모집단 표본	60세 이상 노인이 일년에 약국을 찾는 횟수에 관심을 갖고 대전 A 병원을 찾는 환자들을 대상으로 20명의 자료를 수집하였다.	확률 분포 함수가 $f(x)$인 모집단으로부터 크기 20의 확률 표본을 추출하였다.
자료 관측치	10, 9, 2, 18, 32, 3, 5, 14, 10, 9 6, 8, 1, 15, 12, 9, 4, 7, 8, 3	$x_1, x_2, \ldots, x_{20}$
관심	자료 정리 (elementary statistic) 모수 (중앙 위치 값) 추정 Location parameter: mean, median	$f(x)$ 추정, 그러나 불가능 결국 parameter에 관심
모수	histogram, stem-leaf, box-plot을 통해	$f(x)$에 대한 가정

추정 방법	$f(x)$ 분포 형태를 조사한다. skewed, outlier가 있는지 조사한다.	
	stem-leaf를 그리시오.	
관심 모수	60세 이상 노인들의 일년에 평균적으로 몇 번 병원을 찾는가? population mean(m) 혹은 median(M)	$\int_{-\infty}^{t} f(x)dx = 1/2$ 을 만족하는 t를 찾는다. 결국은 m , M
통계량 (statistic)	sample mean: $\bar{x} = \sum_{i=1}^{20} x_i / n$ sample median: $[x_{(10)} + x_{(11)}]/2$	
	평균과 중앙값을 구하시오.	
모수 추정	$\hat{m} = \bar{x}$, $\hat{M} = [x_{(10)} + x_{(11)}]/2$	
가설	60세 이상 노인들의 일년에 평균적으로 9번 이상 병원을 찾는가? 귀무가설 (null hypothesis): $H_0 : m = 9$ 혹은 $H_0 : M = 9$ 대립가설 (alternative hypothesis): $H_a : m > 9$ 혹은 $H_a : M > 9$	
모집단 확률 분포	 문제는 모집단 확률 분포를 알지 못한다.	
가설 검정 기초	확률 표본으로부터 계산되는 통계량을 이용하여 가설을 검정할 수 밖에 없다. 그러기 위해서는 통계량의 분포를 알아야 한다. 표본 평균의 분포는 구하기 쉬우나 표본 중앙값의 분포는 구하기 어려우므로 통계학에서는 중앙 위치에 대한 정보로 평균을 더 많이 이용한다. 그러나 표본의 크기가 작은 경우 평균을 이용하여 가설을 검정하는 경우 가정을 하게 되는데... 이 가정이 무너진다면... (그래서 비모수 방법 사용)	

표본 평균 분포 (대표본)	▪ $f(x;u,o)$ 로부터 추출한 확률 표본의 표본 평균 $\bar{x}$ 는 평균 u , 표준편차 $s/\sqrt{n}$ 을 따른다. (수리적 증명) ▪ **[Central Limit Theorem]** 표본의 크기 n이 큰 경우 (대표본) 표본 평균의 분포는 normal distribution을 따른다. 
🕮	중심극한 정리의 의미가 무엇인지 확률분포 함수를 이용하여 설명하시오.
가설 검정 (대표본)	검정통계량: $t = \dfrac{\bar{x}-u}{s/\sqrt{n}} \sim \text{normal}(0,1)$ 모집단의 표준편차(σ)를 모를 경우는 표본 표준편차(s)를 이용한다.  *) 검정 통계량의 분포는 귀무가설이 맞다는 가정 하에서 구한 것이다.
가설 검정 (소표본)	▪ 더 이상 중심극한 정리를 사용할 수 없다. ▪ **[Student t-distribution]** 모집단이 정규분포를 따른다면 다음이 성립한다. $t = \dfrac{\bar{x}-u}{s/\sqrt{n}} \sim \text{t(n-1)}$
비모수 방법	소표본이고 모집단이 정규분포를 따른다는 가정이 성립하지 않는다면…

📖 Homework #2 [due 3월 15일]

중심극한 정리를 이해하기 위하여 Simulation 방법을 이용하여 살펴 보자.

[방법]

1) 모집단이 치우진 형태를 갖는 χ^2분포라고 가정하자.
2) 다음 방법에 의해 지수 분포로부터 크기 20의 확률 표본을 추출하자.

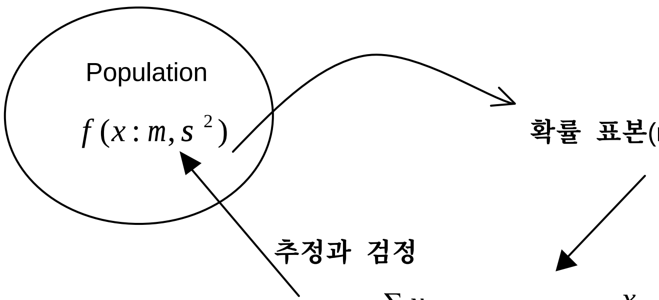
```
data one;
   do i=1 to 20;
      xc2=2*rangam(0, 0.5);
      xc6=2*rangam(0, 5);
      output;
   end;
run;
proc means data=one mean std;
        var xc2 xc6;
run;
```

3) xc2 ~ $\chi^2(\lambda=2)$, xc6 ~ $\chi^2(\lambda=10)$을 따른다.
4) 각 (xc2, xc6) 표본들의 평균을 계산하다.
5) 표본 평균의 분포를 구하기 위하여(히스토그램) 2)-4)작업을 1000번 반복한다.
6) 표본 평균의 분포 함수(히스토그램)을 그리고 비교 해석하시오.

2. One sample

표본의 통계량을 (statistic) 이용하여 모수(parameter)를 추정(estimate)하거나 모수에 대한 통계적 가설을 (statistical hypothesis) 검정하는 것을 통계적 추론이라 (statistical inference) 하고 관심이 되는 모수가 하나인 경우를 일변량 추론이라 한다.

일변량 추론의 주요 관심은 변수 자료의 중앙이 (location parameter) 어디인가 하는 것이고 중앙 위치로는 평균(mean)과 중앙값(median)이 주로 사용된다.



확률 표본(random sample): $(x_1, x_2, \ldots, x_n)$

추정과 검정

계산

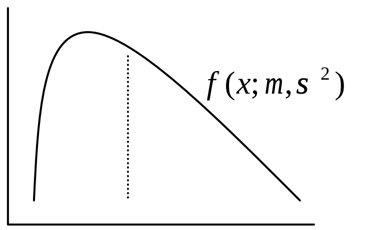
statistic: 평균 $\bar{x} = \frac{\sum x_i}{n}$, 중앙값 $m = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$ $or = x_{((n+1)/2)}$

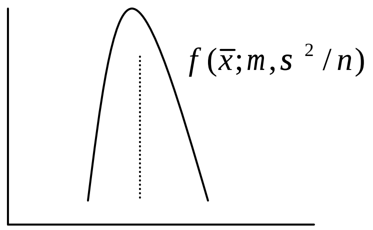
표준 편차 $s = \sum (x_i - \bar{x})^2 / (n-1)$

2.1. parameteric procedure (모수적 방법)

모수적 방법은 표본 평균을 (sample mean) 이용하여 모평균에 대한 추론을 실시한다.

$$E(\bar{x}) = m, \quad V(\bar{x}) = s^2 / n$$





2.1.1. 대표본 (large sample)

1) 중심극한 정리 (central limit theorem)

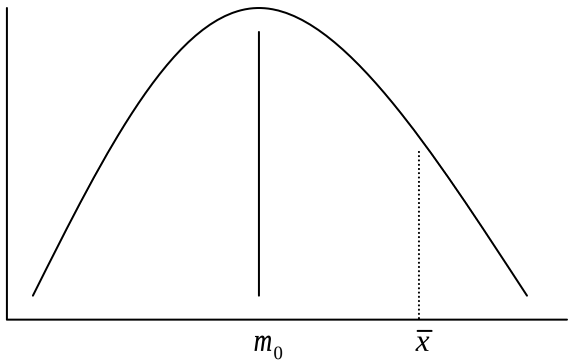
표본의 크기 n이 큰 경우 표본 평균 $\bar{x}$ 의 분포는 모집단 분포 $f(x)$ 관계없이 정규분포를 따른다.

2) 가설 검정 (hypothesis testing)

- 귀무 가설 (null hypothesis): 모평균 $m = m_0$ (하나의 값)
- 대립 가설 (alternative hypothesis): $m \neq m_0$ (two-sided) 혹은 $m > m_0$ (one-sided)
- 검정 통계량 (test statistic): 가설 검정에서 검정 통계량의 분포는 귀무 가설 하에서 구

해진다. 자료로부터 계산된 표본 평균이 귀무 가설에 설정된 분포 하에서 일어날 확률이 매우 적다면 (이를 계산된 유의 수준, 혹은 p-value라 함) 귀무 가설을 기각한다.

$$T = \frac{\bar{x} - m_0}{s/\sqrt{n}} \sim Normal(0,1)$$



- 결론 (conclusion): 계산된 검정 통계량 T 값이 기각역에 (critical region) 속하거나 계산된 유의 수준이 (p-값) 미리 설정한 유의 수준보다 (1종 오류, α) 작다면 귀무 가설을 기각 (대립 가설 채택) 한다. two-sided (양측) 가설 검정에 의해 귀무 가설이 기각되면 같은 유의 수준에서는 (α) one-sided (단측) 가설 검정에서도 귀무 가설이 기각된다.

3) 신뢰 구간 (confidence interval)

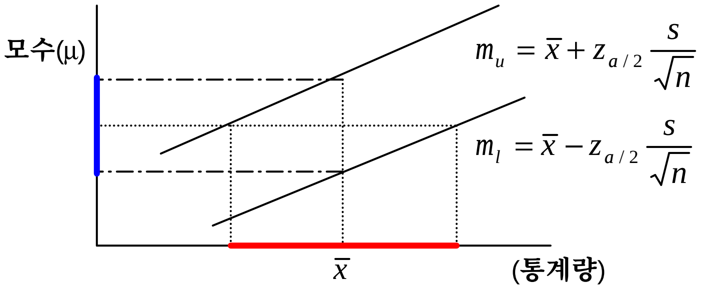
- 통계량의 분포를 이용하여 모수에 대한 100(1-α)% 신뢰 구간을 구한다.

$$\bar{x} \pm z_{a/2} \frac{s}{\sqrt{n}}$$

하한 (lower limit): $m_l = \bar{x} - z_{a/2} \frac{s}{\sqrt{n}}$, 상한 (upper limit): $m_u = \bar{x} + z_{a/2} \frac{s}{\sqrt{n}}$

- (예) 95% 신뢰 구간의 의미는 모수가 계산된 신뢰구간에 있을 가능성이 (확률) 95%라는 의미가 아니다. 모집단으로부터 표본을 추출하고 신뢰 구간을 구하고, 또 다시 표본을 추출하여 신뢰 구간을 구하고... 이렇게 100개의 신뢰 구간을 구할 경우 그 중 95개의 신뢰 구간만이 모수를 포함하고 있을 것이다.

4) 100(1-α)% 신뢰 구간과 가설 검정간에는 일대일 대응 관계가 있다.



∥ Example ∥ 다음은 A 고등학교 학생들의 몸무게를 알아 보기 확률 표본 (단순 임의 추출법) 20명의 몸무게(kg)를 측정한 자료이다.

51.0	38.1	44.4	46.4	37.6	46.4	38.3	51.0	38.1	45.1
22.9	40.8	34.9	50.7	68.0	58.0	60.2	38.5	50.7	52.1

A 고등학교 학생들의 몸무게에 대한 기초통계량을 구하고 학생들의 몸무게가 50kg 이상이라고 말할 수 있는지를 검정하시오?

class.sas

```
data one;
    input weight @@;
    cards;
51.0    38.1    44.4    46.4    37.6    46.4    38.3    51.0    38.1    45.1
22.9    40.8    34.9    50.7    68.0    58.0    60.2    38.5    50.7    52.1
;
run;

proc univariate data=one plot;
    var weight;
run;
```

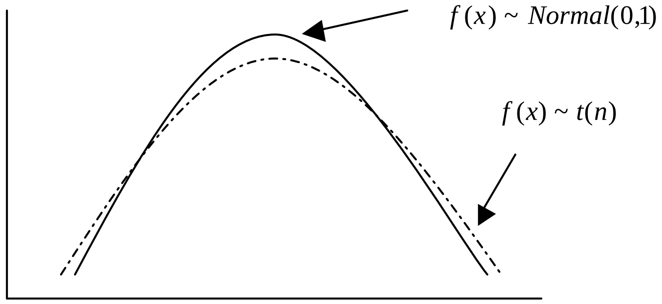
Output - (Untitled)

```
The UNIVARIATE Procedure
Variable:  weight

                          Moments

N                          20    Sum Weights                 20
Mean                    45.66    Sum Observations         913.2
Std Deviation       10.149379    Variance            103.009895
Skewness           0.09086768    Kurtosis            0.68320875
Uncorrected SS        43653.9    Corrected SS          1957.188
Coeff Variation    22.2281626    Std Error Mean      2.26947014


              Basic Statistical Measures

    Location                    Variability

Mean     45.66000     Std Deviation            10.14938
Median   45.75000     Variance                103.00989
Mode     38.10000     Range                    45.10000
                      Interquartile Range      12.80000

NOTE: The mode displayed is the smallest of 4 modes with a count of 2.


           Tests for Location: Mu0=0

Test           -Statistic-    -----p Value------

Student's t    t  20.11923    Pr > |t|    <.0001
Sign           M        10    Pr >= |M|   <.0001
Signed Rank    S       105    Pr >= |S|   <.0001
```

2.1.2. 소표본 (small sample)

1) 가정: 표본의 크기 n이 작은 경우 모집단 분포에 $f(x)$ 대한 가정이 있어야 모수적 방법을 사용할 수 있다.
2) t-분포: 모집단 분포가 $f(x)$ 정규분포를 따른다면 다음이 성립함을 W. S. Gosset이 실증적으로 보였다. 그는 이 분포를 발표할 때 그의 필명 Student를 사용했다.

$$\frac{\bar{x}-m}{s/\sqrt{n}} \sim t(n-1)$$; 평균=0, 분산=n/(n-2)



[표분 정규 분포와 t-분포의 관계]

3) 검정 통계량 및 신뢰 구간

$$T = \frac{\bar{x}-m}{s/\sqrt{n}} \sim t(n-1)$$

100(1-α)% 신뢰 구간: $\bar{x} \pm t(n-1; a/2)\frac{s}{\sqrt{n}}$

2.2. Nonparametric procedure I: Sign Test

표본의 크기 n이 작고 모집단 분포가 정규 분포를 따른다는 확신이 없다면 모수적 방법을 사용할 수 없다. 이런 경우 사용되는 비모수적 방법에서는 모수는 모집단의 중앙값이고 검정 통계량은 표본 중앙값을 사용하나 표본 중앙값의 분포는 알려져 있지 않다.
가장 오래된 비모수 검정 방법으로 측정치들의 값을 부호(sign)로 바꾼 후 그것을 이용하여 모집단의 중앙값에 대한 가설 검정을 하게 된다.

2.2.1. 가설 검정 (hypothesis testing)

1) 가정 (assumption): 관측치 $x_1, x_2, \ldots, x_n$ 는 모집단 분포에서 $f(x:M)$ 추출한 확률 표본 (random sample) 이다. 측정형 (연속형) 변수이어야 한다.

2) **가설** (statistical hypothesis):

- 귀무 가설: $H_0 : M = M_0$
- 대립 가설: $H_a : M \neq M_0$(양측 검정) $H_a : M > M_0$ 혹은 $H_a : M < M_0$ (단측)

3) **검정 통계량** (test statistic):

- 각 관측치에 대해 $(x_i - M_0)$의 부호를 계산한다. M_0은 귀무 가설에 설정된 모수 중앙값이다. $x_i = M_0$여서 부호가 결정되지 않는 것은 표본에서 제외한다.
- 만약 귀무 가설이 맞다면 표본 관측치로부터 얻은 +, - 부호의 수는 같을 것이다. 만약 + 부호 수가 상대적으로 많거나 (반대로 – 부호 수가 상대적으로 많다면) 귀무 가설을 기각할 것이다.

4) **결정 법칙** (decision rule):

- $(x_i - M_0)$의 부호는 "+", "-" 두 개만 있으므로(같으면 제외) 성공 확률이 p (사건이 일어날 확률; "+"의 비율)인 베르누이 (Bernoulli) 시행이다.
- "+" (혹은 "-") 부호 개수(K)는 이항 분포(n, p)을 (Binomial Dist.; $B(n, p)$) 따른다.
- 귀무 가설이 참이라면 $K \sim B(n, 0.5)$이다.
- 표본으로부터 구한 "+" 부호의 개수를 k라 관측된 p-값은 다음과 같다.
  - i) k가 $0.5n$보다 적은 경우: p-value= $\Pr(K \le k \mid n, 0.5)$
  - ii) k가 $0.5n$보다 큰 경우: p-value= $\Pr(K \ge k \mid n, 0.5)$
- 편의상 "+"부호의 개수와 "–"부호의 개수 중 적은 것을 k로 사용한다. 그러므로 sign test에서는 다음을 p-값으로 사용한다.

  p-value= $\Pr(K \le k \mid n, 0.5)$
- p-값이 유의 수준보다 크다면 귀무 가설 채택하고 반대의 경우 대립 가설을 채택한다.

▸▸참고 위의 Sign test 방법은 소표본인 경우 모비율(population ratio)에 대한 가설 검정에 $(H_0 : p = p_0)$ 이용한다. 모평균 가설 검정과는 달리 모비율 가설 검정에서 소표본의 의미는 $\min(np_0, (n-1)p_0) < 9$을 의미한다.

📖[Homework#3 Due 3월 20일] A회사가 만드는 제품의 불량률은 0.15이었다. 회사는 불량률을 줄이기 위하여 제품 생산 공정을 새롭게 바꾸었다. 제품 불량률이 줄었는가를 알아보기 위하여 생산된 제품 65개를 단순 임의 추출하여 불량 개수를 조사했더니 9개가 불량이었다. 불량률이 줄었는지 가설 검정하시오. 절차를 자세히 적으시오.

ıı EXAMPLE ıı 다음 자료는 11명 환자들의 심장 동맥 박동 시간(초)을 측정한 것이다. 환자들의 중앙값은 3.5초라 할 수 있는지 가설 검정하시오. (유의수준 α=5%) (Applied Nonparametric Statistics, W. Daniel)

1.8 3.3 5.65 2.25 2.5 3.5 2.75 3.25 3.1 2.7 3.0

- 가설: $H_0 : M = 3.5$, $H_a : M \neq 3.5$
- 검정 통계량: $(x_i - M_0)$의 부호를 구하면 "-"부호의 개수=9, "+"부호의 개수=1, 0인 개수=1이므로 검정 통계량 k는 1이다.
- p-값= $\Pr(K \le 1 \mid n = 10, p = 0.5) = 0.0108$.
- p<0.025 ($\alpha/2$)이므로 (양측 검정) 귀무 가설을 기각하고 환자들의 심장 박동 시간은 3.5초라고 할 수 없다. 특히 k는 "+"부호의 개수이므로 실제 심장 박동 시간은 3.5초 이하라고 할 수 있다.

▸▸참고 환자들의 심장 박동 시간의 중앙값이 3.5초 이하라고 할 수 있는가? 이런 단측 검정($H_a : M < 3.5$)에서는 p-값과 0.05을 비교한다.

▸▸**Large-sample approximation:** 표본의 크기가 12 이상이면 경우 p-value 편리하게 계산하기 위하여 이항 분포표를 쉽게 접하기 힘든 경우 이항분포의 정규분포 근사(normal approximation to the binomial) 이론을 이용하여 계산 가능하다.

$$z = \frac{(K + 0.5) - 0.5n}{\sqrt{0.5^2 n}} \sim Normal(0,1)$$

위의 예제의 경우 정규 분포 근사에 의하여 p-값을 계산하면, z=-2.21이므로 표준 정규 분포표에 의해 $\Pr(K \le 1 \mid n = 10, p = 0.5) \approx 0.0136$이다. 여전히 귀무 가설은 기각된다. 그러나 n=20 정도까지는 이항 분포표를 이용할 수 있고 n이 그 이상이면 중심극한 정리를 이용한 z-검정을 사용하면 된다.

▸▸**SAS 사용 방법:** 모수적 방법에 사용한 UNIVARIATE procedure를 사용하면 된다.

2.2.2. 신뢰 구간 (confidence intervals)

모집단의 중앙값은 표본의 중앙값으로 추정될 수 (점 추정: point estimate) 있으나 점 추정치는 모수와 같을 가능성(확률)이 거의 없다. 그러므로 추정의 경우 우리의 주된 관심은 모수에 대한 신뢰 구간에 (구간 추정: interval estimate) 있다.

다음은 Sign Test 검정 방법에 대응하는 100(1-α)% 신뢰구간을 구하는 방법을 정리한 것이다.

- 양측 검정의 경우 검정통계량의 값이 $\Pr(K \le K^*) \le \alpha/2$이면 귀무가설을 기각한다.

- $M_l = X_{(k^*+1)}$이 lower limit이고 $M_u =$ 끝에서 $(k^* + 1)$번째 순서 통계량이 upper limit이다.

▪ 100(1-α)% 신뢰구간을 구할 수 없으므로 α값을 조정할 필요가 있다. 예제 참조

॥ EXAMPLE ॥ 다음 자료는 16명의 잠에 드는 시간(분)을 측정한 것이다. 모집단 중앙값에 대한 95% 신뢰 구간을 구하시오. (Applied Nonparametric Statistics, W. Daniel)

1.90 3.08 9.1 3.53 1.99 3.1 10.16 0.69

1.74 2.41 4.01 3.71 8.11 8.23 0.07 3.07

▪ 이항 분포표로부터 $\Pr(K \le 3 \mid n=16, p=0.5)=0.0105$, $\Pr(K \le 4 \mid n=16, p=0.5)=0.0383$

▪ (4+1)을 사용하면 α=0.0383*2=0.0766이 되므로 신뢰 수준은 92.34%이다.

▪ 그러므로 92.34% 신뢰 구간은 ($X_{(5)}=1.99, X_{(12)}=4.01$)이다.

⏩ **Large-sample approximation** 이용하여 신뢰 구간 구하기: 표본의 크기가 12 이상이면 경우 이항분포의 정규분포 근사(normal approximation to the binomial) 이론을 이용하여 신뢰구간을 구할 수 있다.

$$(K^*+1) \approx (n/2) + z_{a/2}\sqrt{n/4}$$

위 예제에서 $(K^*+1) \approx (16/2)+1.96\sqrt{16/4} \approx 4$ 이므로

95% 신뢰 구간은 ($X_{(4)}=1.9, X_{(13)}=8.11$)이다.

근사 이론을 이용하면 마치 95% 신뢰 구간을 얻은 것처럼 보이지만 실제로는 그렇지 않다. α=0.0105*2=0.021이므로 95%가 아니라 97.9% 신뢰구간이 된다.

📖 [Homework#3 Due 3월 20일] (Applied Nonparametric Statistics, W. Daniel).

2.1 Lenzer et al. (E2) reported the endurance scores of animals during a 48-hour session of discrimination responding. The median score for animals with electrodes implanted in the hypothalamus was 97.5. Suppose that the experiment was duplicated in another laboratory, except that electrodes were implanted in the forebrain of 12 animals. Assume that investigators observed the endurance scores shown in Table 2.2.

Use the one-sample sign test to see whether the investigators may conclude at the 0.05 level of significance that the median endurance score of animals with electrodes implanted in the forebrain is less than 97.5. What is the *P* value for this test?

TABLE 2.2

Endurance scores of animals with electrodes implanted in forebrain

93.6	89.1	97.7	84.4	97.8	94.5	88.3	97.5	83.7	94.6	85.5	82.6

2.2 Iwamoto (E3) found that the mean weight of a sample of a particular species of adult female monkey from a certain locality was 8.41 kg. Suppose that a sample of adult females of the same species from another locality yielded the weights shown in Table 2.3.

Can we conclude that the median weight of the population from which this second sample was drawn is greater than 8.41 kg? Use the one-sample sign test and a 0.05 level of significance. What is the P value for this test?

Weights of female monkeys, kilograms ***TABLE 2.3***

8.30	9.50	9.60	8.75	8.40	9.10	9.25	9.80	10.05	8.15	10.00	9.60	9.80	9.20	9.30

2.6 Armstrong (E9) studied the daily exposure, in minutes, of 10 North Hawaiian families to the risk of a motor vehicle accident. Suppose that a similar survey in another area yielded the exposure times shown in Table 2.10. Find the point estimate and construct an approximate 95% confidence interval for the population median.

TABLE 2.10

Exposure times per day, in minutes, of individuals to motor-vehicle accidents

18.3	45.2	19.1	57.0	63.9	10.3	12.1	35.5	36.6	74.6
10.5	27.8	44.9	40.9	63.7	40.8	59.1	31.5	40.1	8.1

2.7 Abu-Ayyash (E10) found that the median education of heads of households living in mobile homes in a certain area was 11.6 years. Suppose that a similar survey conducted in another area revealed the educational levels of heads of households shown in Table 2.11. Find the point estimate, and construct the approximate 95% confidence interval for the population median.

TABLE 2.11

Educational levels (years of school completed) of heads of households residing in mobile homes

13	6	6	12	12	10	9	11	14	8	7	16	15	8	7

2.3. Nonparametric procedure II: Wilcoxon Signed-ranks test

sign test의 경우는 관측치와 귀무 가설의 간의 $(x_i - M_0)$ 부호만을 사용하지만 차이의 크기를 이용하여 분석하는 방법이다.

2.3.1. 가설 검정 (hypothesis testing)

1) 가정 (assumption): 관측치 $x_1, x_2, \ldots, x_n$ 는 모집단 분포에서 $f(x:M)$ 추출한 확률 표본 (random sample) 이다. 측정형 (연속형) 변수이어야 한다. 그리고 sign test와는 달리 표본 분포는 좌우 대칭이어야 한다는 가정이 하나 더 필요하다.

2) 가설 (statistical hypothesis):
   - 귀무 가설: $H_0 : M = M_0$(양측), $H_0 : M \le M_0$ 혹은 $H_0 : M \ge M_0$(단측)
   - 대립 가설: $H_a : M \ne M_0$(양측), $H_a : M > M_0$ 혹은 $H_a : M < M_0$(단측)

3) 검정 통계량 (test statistic):
   - $D_i = (x_i - M_0)$의 부호를 계산한다. 만약 $D_i = 0$이면 그 관측치는 표본에서 제외한다. M_0은 귀무 가설에 설정된 모수 중앙값이다.
   - D_i의 절대값 $|D_i|$의 크기에 따라 작은 것부터 순위를 부여한다. $|D_i|$ 크기가 같은 것이 있다면 그 순위들의 평균을 각각 부여한다. 예를 들어 관측치 중 3개의 값이 동일하다면 $|D_i|$의 크기는 같다. 그리고 그 값이 최소값이면 순위는 1, 2, 3을 부여할 수 있는데 $|D_i|$값이 모두 같으므로 (1=2+3)/3=2, 즉 3개 관측치에 대해 순위는 모두 2를 부여한다.
   - 각 관측치의 순위 값에 D_i의 부호를 붙인다.
   - 부호가 –인 순위의 합을 계산하여 T_-, 부호가 +인 순위의 합을 계산하여 T_+라 한다. 두 통계량에는 $T_+ = [n(n+1)/2] - T_-$의 관계가 있다.
   - 귀무 가설이 참이라면 T_-의 기대치와 T_+의 기대치는 같을 것이다. 그러므로 T_-나 T_+가 충분히 적다면 귀무 가설을 기각하면 된다.
   - 대립가설이 $H_a : M \ne M_0$이면 검정 통계량으로 $T = \min(T_+, T_-)$ 사용한다.
   - 대립가설이 $H_a : M > M_0$이면 검정 통계량으로 $T = T_-$ 사용한다.
   - 대립가설이 $H_a : M < M_0$이면 검정 통계량으로 $T = T_+$ 사용한다.

4) 결정 법칙 (decision rule):
   - Wilcoxon이 제공한 표를 참조하여 가설을 검정한다. 표에는 표본의 크기 n에 따라 기각역 (critical region)과 거기에 대응하는 α값이 나타나 있다.
   - 검정 통계량이 기각역에 포함되어 있으면 귀무 가설을 기각하고 반대의 경우 귀무 가설을 채택한다.

▸▸참고 위의 Sign test 방법은 소표본인 경우 모비율(population ratio)에 대한 가설 검정에 $(H_0 : p = p_0)$ 이용한다. 모평균 가설 검정과는 달리 모비율 가설 검정에서 소표본

의 의미는 $\min(np_0, (n-1)p_0) < 9$ 을 의미한다.

▮ EXAMPLE ▮ 마약에 중독된 학생들의 IQ 중앙값이 107이다. 마약이 학생들의 IQ에 영향을 미치는지 알아보기 위하여 마약을 하지 않는 학생들 중 15명을 단순 임의 추출하여 조사하였다. (유의수준 α=5%) (Applied Nonparametric Statistics, W. Daniel)

99 100 90 94 135 108 107 111 119 104 127 109 117 105 125

- 가설: $H_0: M = 107$, $H_a: M \neq 107$
- 검정 통계량: $T_+ = 64.5$, $T_- = 40.5$ 이므로 검정통계량 $T = \min(64.5, 40.5) = 40.5$.

IQ	$D_i = X_i - M_0$	Rank of $\|D_i\|$	Signed rank of $\|D_i\|$
99	−8	7	−7
100	−7	6	−6
90	−17	11	−11
94	−13	10	−10
135	+28	14	+14
108	+1	1	+1
107	0		Eliminate from analysis
111	+4	5	+5
119	+12	9	+9
104	−3	4	−4
127	+20	13	+13
109	+2	2.5	+2.5
117	+10	8	+8
105	−2	2.5	−2.5
125	+18	12	+12
			$T_+ = 64.5$
			$T_- = 40.5$

- Wilcoxon signed-ranks Table에 의하면 기각역 { $T \le 21$ }이다. 검정 통계량 $T = 40.5$ 이므로 기각역에 포함되지 않는다. 그러므로 귀무 가설을 기각할 수 없다. 마약은 학생들의 IQ에 영향을 주지 못한다.
- 위의 가설 검정의 경우 유의수준 α=0.0247*2=0.494이다.
- $H_0: M \ge 107$ vs. $H_a: M < 107$ 단측 검정에서는 기각역이 { $T \le 25$ }이다. 검정 통계량 $T = T_- = 40.5$ 가 기각역에 포함되지 않으므로 귀무가설을 기각할 수 없다. 마약을 하지 않은 학생들의 IQ가 높다고 할 수 없다.

⏩Large-sample approximation: 표본의 크기가 20 이상이면 경우 Wilcoxon 표를 사용할 수 없다. 이 경우 다음의 정규분포 근사를 이용하여 계산 가능하다.

$$T^* = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim Normal(0,1)$$

T^* 는 양측 검정의 경우 T , 단측 검정의 경우는 T_- 나 T_+ 중 적당한 것을 사용하면 된다. 대표본의 경우는 그냥 중심극한 정리를 사용하여 z-검정하는 것이 더 편리하다.

▶▶ **sampling distribution of** T_+**:표본의 크기가 n=4인 경우 예를 들어** T_+ **의 표본 분포를 Wilcoxon이 어떻게 만들었는지 살펴보자.**

+ 순위	T_+ 값	+ 순위	T_+ 값	+ 순위	T_+ 값	+ 순위	T_+ 값
없음	0	4	4	2,3	5	1,2,4	7
1	1	1,2	3	2,4	6	1,3,4	8
2	2	1,3	4	3,4	7	2,3,4	9
3	3	1,4	5	1,2,3	6	1,2,3,4	10

위의 결과를 이용하여 Wilcoxon 표가 (n=4) 작성되었다.

2.3.2. 신뢰 구간 (confidence intervals)

Wilcoxon 검정 방법에 대응하는 신뢰 구간을 다음 방법을 통하여 얻을 수 있다.

1) $u_{ij}=\dfrac{x_i+x_j}{2},\ 1\le i\le j\le n$ 평균은 몇 개나 있나? 같은 것까지 포함 ${}_{10}C_2+10=55$
2) u_{ij} 을 크기 순서대로 나열한 순서통계량을 구하고 중앙값을 구하면 그것이 모집단 중앙값의 추정치가 된다.
3) Wilcoxon 표에서 n와 P를 이용하여 귀무가설을 기각하는 T값을 찾는다. K(=T+1)번째 u_{ij} 의 순서 통계량이 하한 값(lower limit) 끝에서 K번째 순서 통계량dl 상한 값(upper limit)이 된다.

ll EXAMPLE ll 10명 학생들의 신뢰성 정도를 측정한 것이다. 모집단 중앙값에 대한 95% 신뢰구간을 구하시오. (Applied Nonparametric Statistics, W. Daniel)

28.5 25.2 28.7 41 29.1 21.3 37.7 39.9 26.8 28.8

1) $u_{ij}=\dfrac{x_i+x_j}{2},\ 1\le i\le j\le n$ 을 구한다.
2) u_{ij} 의 순서통계량을 구하고 중앙값을 구하면 $u_{ij(28)}=31.45$ 이다.
3) 표에 의하면 n=10이고 P=0.0244일 때 T=8이다. ($u_{ij(9)}=27.75$, $u_{ij(47)}=35.05$)이 95.12% 신뢰 구간이 된다.

▶▶ **Large-sample approximation을 이용한 신뢰 구간: 표본의 크기가 20 이상이면 경우 Wilcoxon 표를 사용할 수 없다. 다음의 근사 이론을 이용하여 신뢰구간을 구할 수 있다.**

$$K\approx\frac{n(n+1)}{4}-z_{\alpha/2}\sqrt{\frac{n(n+1)(2n+1)}{24}}$$

📖[Homework#4 Due 3월 27일] (Applied Nonparametric Statistics, W. Daniel).

2.3 Malina (E5) reported the results of a study of weights of football players at the University of Texas at Austin between 1899 and 1970. Suppose that the weights of a random sample of 15 football players during the past 10 years at another large state university are those shown in Table 2.6.

Can we conclude that the median weight of the population from which this sample was drawn is greater than 163.5 pounds? Let $\alpha = 0.05$. What is the P value for this test?

Weights of football players ***TABLE 2.6***

Player	Weight	Player	Weight
1	188.0	9	214.4
2	211.2	10	221.0
3	170.8	11	162.0
4	212.4	12	222.8
5	156.9	13	174.1
6	223.1	14	210.3
7	235.9	15	195.2
8	183.9		

2.4 Moore and Ogletree (E6) investigated the readiness of pupils at the beginning of the first grade. They compared scores on a readiness test of pupils who had attended a Head Start program for a full year with the scores of those who had not. Suppose that a random sample of 20 pupils who had not attended a Head Start program achieved the scores on the readiness test shown in Table 2.7.

Can we conclude that the median score of the population represented by this sample is less than 45.32? What is the P value for this test?

TABLE 2.7

Readiness test scores of pupils who did not attend a Head Start program

Pupil	Readiness score	Pupil	Readiness score
1	33	11	41
2	19	12	31
3	40	13	46
4	35	14	51
5	51	15	34
6	41	16	37
7	27	17	36
8	23	18	55
9	39	19	52
10	21	20	32

[homework #4 계속]Sign Test의 숙제 중 2.7에 대해 Wilcoxon 방법에 의해 모집단 중앙값에 대한 approximation 95% confidence interval을 구하시오. approximation의 의미는 정확하게 95% 신뢰 수준을 맞출 수 없다는 것을 의미하는 것이지, large-sample이론을 적용하라는 것은 아님.

2.4. One-sample runs test for randomness

통계 자료분석에서 가장 중요한 가정은 확률 표본(random sample) 가정이다. 이 가정은 성립된다고 가정하지만 randomness가 의심이 되는 경우 이 가정을 검정하는 방법이 Run 검정이다.

2.4.0. Example

1) p에 대한 control chart: 불량률에 대한 관리도의 경우 표본을 정기적으로 추출하여 불량률이 control limit을 벗어나는 경우 생산 공정을 중단하고 재조사한다. 이런 경우 표본의 randomness가 만족하지 않고 일정한 pattern을 갖는다면 공정의 control에 문제가 발생한다.
2) 회귀분석 잔차의 독립성

2.4.1. 과정

0) 기본 개념:

randomness는 runs의 수(부호가 바뀐 회수)와 그 run의 길이에 의해 관측될 수 있다.

MFMFMFMFMF ➔ runs의 수는 10이다. pattern이 존재

MMMMMFFFFF ➔ runs의 수는 2이고 길이가 5인 경우이다.

1) 가정

관측치를 두 개의 군으로 나눌 수 있다. 표본의 크기가 n, 한쪽 그룹의 관측치 수가 n_1, 다른 그룹의 관측치 수가 n_2인 경우 $n=n_1+n_2$이다.

2) 가설

귀무가설(null hypothesis): 관측치는 randomness를 만족한다.

대립가설(alternative hypothesis): 관측치는 randomness를 만족하지 않는다.

3) 검정통계량: runs의 총 회수(r)이다. 부호가 바뀌는 회수이다.
4) 결정 법칙 (decision rule):
   - 검정통계량 r이 (n_1, n_2)에 적절한 표2 (critical values of r in the runs test)의 하한 기각역보다 작거나* 같은 경우 포함) 상한 기각역보다 큰(같은 경우 포함) 경우 귀무가설을 기각한다.

|| EXAMPLE || 다음은 A 도시의 기온은 평년 기온으로부터 벗어난 자료이다. 이 자료를 이용하여 평년 기온을 벗어난 +, - 패턴이 random인가를 알아보려고 한다.

TABLE 2.16 **Departures from normal of daily temperatures recorded at Atlanta, Georgia, during November, 1974**

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Departure from normal	12	13	12	11	5	2	−1	2	−1	3	2	−6	−7	−7	−12
Day	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Departure from normal	−9	6	7	10	6	1	1	3	7	−2	−6	−6	−5	−2	−1

Source: Local Climatological Data, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, Environmental Data Service, National Climatic Center, Federal Building, Asheville, N.C., November 1974.

- 귀무가설: 평년 기온으로부터 벗어나는 형태가 random이다.
- 대립가설: 평년 기온으로부터 벗어나는 형태가 random이 아니다.
- 검정 통계량: $n_1 = 17$ (+ 방향), $n_2 = 13$ (- 방향) $n = 30$ 이고 r=8이다.
- 표2에 의하면 기각역은 $\{r \le 10\}$ 이거나 $\{r \ge 22\}$ 이므로 $r = 8$ 은 기각역에 속한다. 그러므로 귀무가설은 유의수준 5%에서 기각되고 평년 기온과의 차이는 랜덤이 아니다.

▸▸Large-sample approximation: n_1, n_2 가 20 이상이면 경우 다음의 정규분포 근사를 이용하여 계산 가능하다.

$$T^* = \frac{r - \{[(2n_1 n_2)/(n_1 + n_2)] + 1\}}{\sqrt{\dfrac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}} \sim Normal(0,1)$$

📖 [Homework#5 Due 3월 29일] (Applied Nonparametric Statistics, W. Daniel).

2.19 Table 2.19 shows the actual daily occurrence of sunshine in Atlanta during November 1974, as a percentage of the possible time the sun could have shone if it had not been for cloudy skies. The data are from the U.S. Department of Commerce (E28). Dichotomize the observations according to whether the amount of sunshine was more than 50% of possible or 50% or less, and test the null hypothesis that the pattern of occurrence of the two types of day is random.

TABLE 2.19 **Percentage of day during which sunshine occurred in Atlanta, November, 1974**

Day	Percentage	Day	Percentage	Day	Percentage
1	85	11	31	21	87
2	85	12	86	22	100
3	99	13	100	23	100
4	70	14	0	24	88
5	17	15	100	25	50
6	74	16	100	26	100
7	100	17	46	27	100
8	28	18	7	28	100
9	100	19	12	29	48
10	100	20	54	30	0

Source: *Local Climatological Data*, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, Environmental Data Service, National Climatic Center, Federal Building, Asheville, N.C., November 1974.

2.22 Littler et al. (E31) studied the blood flow in lung capillaries in 16 patients with scoliosis or neuromuscular weakness. They reported the sex of the patients in the following order:

F F F M F F M M M F F F F F F M

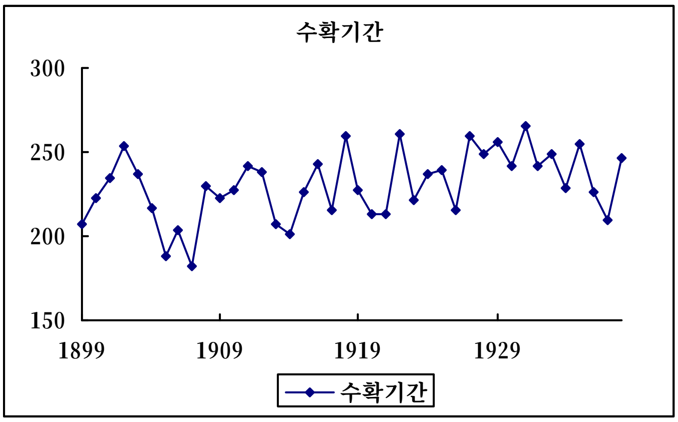
Test the null hypothesis that this sequence is random.

2.5. Cox-Stuart Test for Trend

시계열 자료의 trend가 있는지를 검정하는 방법으로 Sign test의 변형이다.

Cox, D. R. and A. Stuart, “Some Quick Tests for trend in Location and dispersion”, Biometrika, 42 (1955), 80-95

2.5.0. Example (수확기간)



년도	수확기간	년도	수확기간
1899	207	1919	227
1900	223	1920	213
1901	235	1921	213
1902	254	1922	261
1903	237	1923	222
1904	217	1924	237
1905	188	1925	239
1906	204	1926	216
1907	182	1927	260
1908	230	1928	249
1909	223	1929	256
1910	227	1930	242
1911	242	1931	266
1912	238	1932	242
1913	207	1933	249
1914	201	1934	228
1915	226	1935	255
1916	243	1936	226
1917	215	1937	209
1918	259	1938	247

2.5.1. 과정

1) 가정

시계열 자료 관측치는 독립이고 측정형 변수이어야 한다.

2) 가설

귀무가설(null hypothesis): 시계열 자료에 trend가 존재하지 않는다.

대립가설(alternative hypothesis)1: upward trend가 존재한다.

대립가설(alternative hypothesis)2: downward trend가 존재한다.

대립가설(alternative hypothesis)3: trend가 존재한다.

3) 검정통계량: (x_i, x_{c+i}) 의 쌍을 이룬 후 $(x_{c+i} - x_i)$ 차이의 부호(+, -)에 의해 trend의 유무를 검정한다. n이 짝수이면 $c = n/2$ 이고 n이 홀수이면 $c = (n+1)/2$ 이므로 홀수인 경우는 제일 가운데 하나를 제외하고 쌍을 이루게 된다.

4) 결정 법칙 (decision rule):

부호가 0인 것을 제외하고 +의 부호 개수나 –의 부호 개수에 의해 가설을 검정하므로 sign test와 동일하다.

■ EXAMPLE ■ 미국 A 도시의 수확 기간을 조사한 것이다.

- 귀무가설: 수확기간 시계열 자료에 trend가 존재하지 않는다.
- 대립가설: 수확기간 시계열 자료에 trend가 존재한다. (양측검정)
- 검정 통계량: (207, 227), (223,213), … ($n' = 40$) ➔ + 개수=6 – 개수는=14, 그러므로 검정통계량은 6이다. (sign test와 동일)
- 이항 분포표로부터 p-value $P(K \le 6 \mid n = 20, p = 0.5)$ =0.0577이므로 0.025보다 크다. 그러므로 귀무가설을 채택하고 이 시계열 자료에는 trend가 없다고 결론을 맺는다.

Large-sample approximation: sign test와 동일하다.

[Homework#5 계속 Due 3월 29일] (Applied Nonparametric Statistics, W. Daniel).

2.24 The 1972 edition of the *FAA Statistical Handbook of Aviation* (E34) gives the information on annual United States exports of aircraft, aircraft parts, and accessories shown in Table 2.24. Do these data reflect an upward trend in exports? Let $\alpha = 0.05$. What is the P value?

TABLE 2.24 Number of U.S. aircraft exports, 1947–1971

Year	Aircraft exports*	Year	Aircraft exports*
1947	3125	1960	2336
1948	2259	1961	2459
1949	881	1962	2131
1950	756	1963	2251
1951	894	1964	2577
1952	1180	1965	3129
1953	1377	1966	3611
1954	1053	1967	3881
1955	1714	1968	3682
1956	1711	1969	3322
1957	2025	1970	3383
1958	1689	1971	2904
1959	1628		

Source: FAA Statistical Handbook of Aviation, Department of Transportation, Federal Aviation Administration, for sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1972.

* 1949–1954, civil only