개요
순서통계량

중앙값 1 모집단 - Sign Test, Wilcoxon Ranks Sum Test
Run 검정 : Randomness 검정, Trend 검정

독립인 두 모집단 평균 차이 - Median 검정, Mann-Whitney
산포 차이 검정 - Ansari -Bradley

짝이룬 t-검정 - McNemar 검정 포함

독립성/동질성 검정 - 카이제곱 검정 (범주형 교차표)

비모수 분산분석 - Median 검정, Kruskal-Wallis 검정
비모수 다중비교 (JONCKHEERE-TERPSTRA test)

적합도 검정 : Goodness of Fits 검정
Kolmogorov-Smirov 검정

상관계수 Spearman 순위 상관 계수, Kendall's Tau

1. Introduction

비모수 방법은 모수에 (parameter) 대한 어떠한 가정(assumption) 없이 통계 가설 검정하는 것이다. 모수에 대한 가정이란 일반적으로 모집단 분포에 대한 가정을 의미한다. 가설 검정을 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…

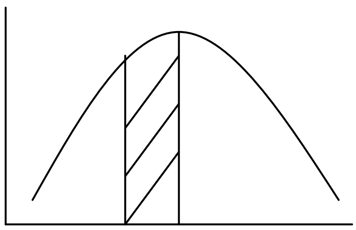
모집단 전체를 조사하는 것이 때로 시간, 비용면에서 비효율적이거나 불가능한 경우 적절한 표본 추출 방법을 이용하여 모집단 구성원의 일부만을 추출하여 분석한다. 이 모집단 일부를 표본(sample)이라 한다. 비모수 방법은 표본 자료의 크기가 작은 경우 (소표본, $n \le 20$) 주로 이용된다.

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.



70

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.

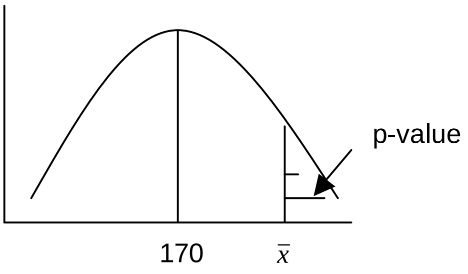
비모수 방법은 측정형 변수에 대한 분석 방법이고 분류형 변수의 경우는 비율에 대한 가설 검정에만 이용된다.

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)

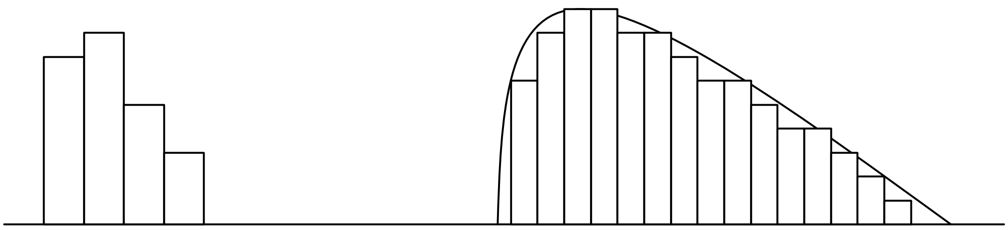
1.3.1. review

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: $X(c)=i$ 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: $f(x)=1/6$ 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) $X_{(1)}, X_{(2)}, ..., X_{(n)}$ 의 결합 밀도 함수 (joint distribution function)

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

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

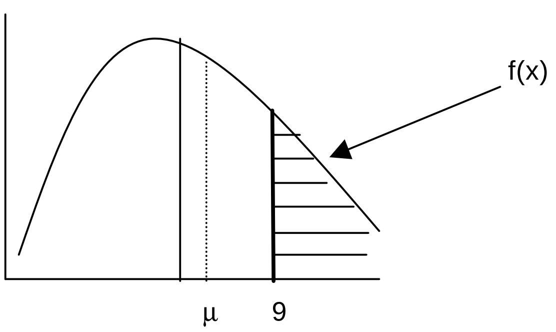
🕮 **모집단 분포 $f(x) = 2x,\ \ 0 < x < 1$로부터 크기 4의 확률 표본 x_1, x_2, x_3, x_4를 추출했을 경우 $x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}$의 joint pdf, 최소값 $x_{(1)}$, 최대값 $x_{(n)}$의 marginal pdf를 구하시오.**

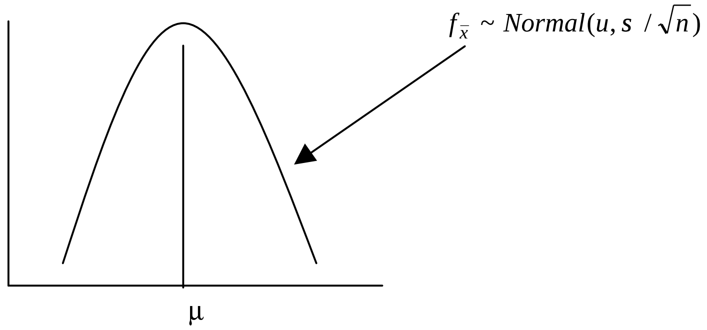
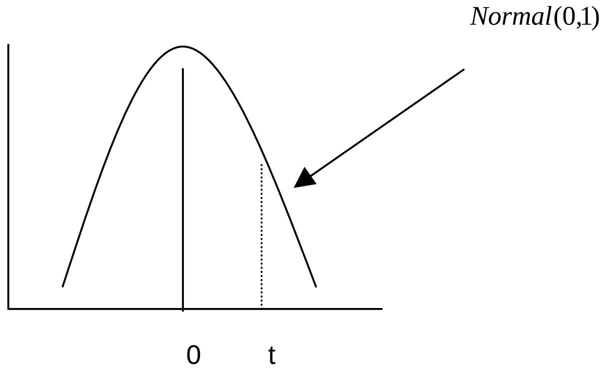
1.4. 언제 비모수 방법을 사용할 것인가? (예제)

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: $t = \dfrac{\bar{x}-u}{s/\sqrt{n}} \sim \mathrm{normal}(0,1)$ 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. $t = \dfrac{\bar{x}-u}{s/\sqrt{n}} \sim \mathrm{t(n-1)}$
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^2 분포라고 가정하자.
2) 다음 방법에 의해 지수 분포로부터 크기 20의 확률 표본을 추출하자.

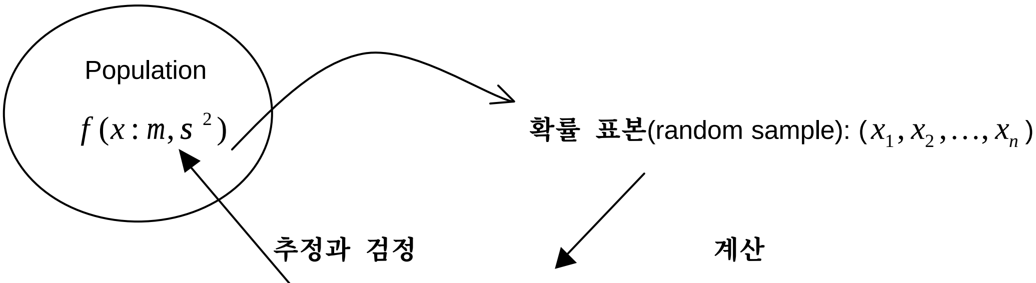
```
data one;
   do i=1 to 20;
      xc2=2*rangam(0, 0.5);
      xc6=2*rangam(0, 5);
      output;
   end;
run;
proc means data=one mean std;
        var xc2 xc6;
run;
```

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.



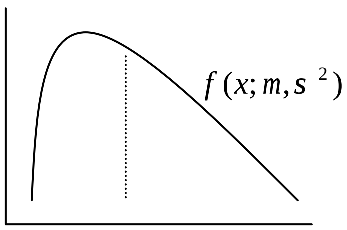
statistic: 평균 $\bar{x} = \frac{\sum x_i}{n}$, 중앙값 $m = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$ $or = x_{((n+1)/2)}$

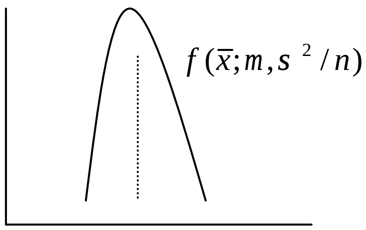
표준 편차 $s = \sum (x_i - \bar{x})^2 / (n-1)$

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)

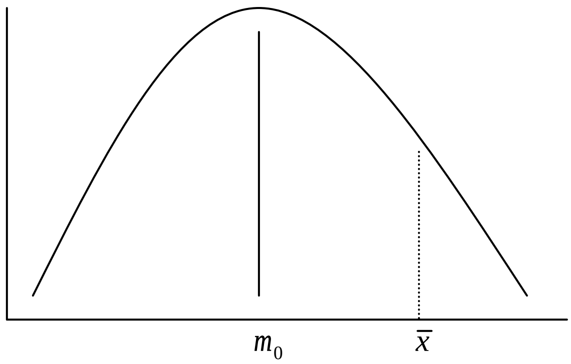
표본의 크기 n이 큰 경우 표본 평균 $\bar{x}$ 의 분포는 모집단 분포 $f(x)$ 관계없이 정규분포를 따른다.

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)

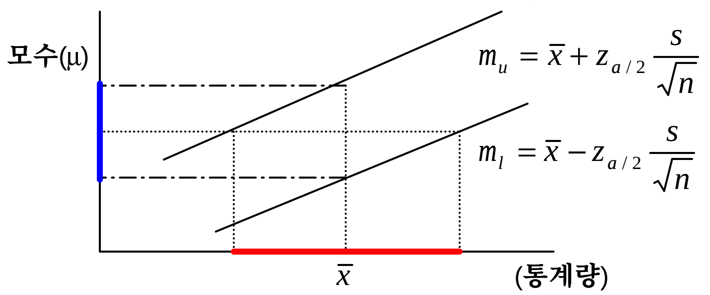
- 통계량의 분포를 이용하여 모수에 대한 100(1-α)% 신뢰 구간을 구한다.

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

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;
```

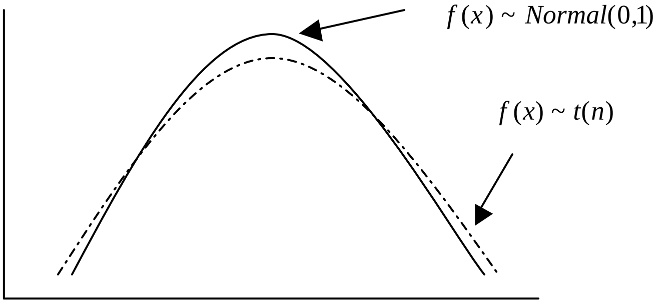
Output - (Untitled)

```
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.

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



[표분 정규 분포와 t-분포의 관계]

3) 검정 통계량 및 신뢰 구간

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

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

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.

II EXAMPLE II 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= $\Pr(K \leq 1 \mid n = 10, p = 0.5) = 0.0108$.
- 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 $\Pr(K \leq 1 \mid n = 10, p = 0.5) \approx 0.0136$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 $\Pr(K \leq K^*) \leq \alpha/2$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

Ⅱ EXAMPLE Ⅱ 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_0 은 귀무 가설에 설정된 모수 중앙값이다.
   - D_i 의 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_i 의 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.

ıı EXAMPLE ıı 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.

ıı EXAMPLE ıı 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 approximation**을 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?

TABLE 2.6

Weights of football players

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.

Ⅱ EXAMPLE Ⅱ 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.

$$T^* = \frac{r - \{[(2n_1n_2)/(n_1+n_2)]+1\}}{\sqrt{\dfrac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}} \sim Normal(0,1)$$

📖 [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

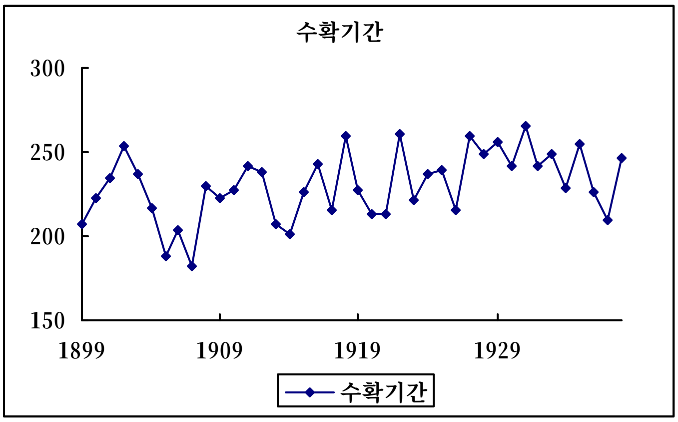
Test the null hypothesis that this sequence is random.

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** trend**가 존재하지 않는다**.

대립가설(alternative hypothesis)1: upward trend**가 존재한다**.

대립가설(alternative hypothesis)2: downward trend**가 존재한다**.

대립가설(alternative hypothesis)3: trend**가 존재한다**.

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

4) **결정 법칙** (decision rule):

부호가 0**인 것을 제외하고** +**의 부호 개수나** –**의 부호 개수에 의해 가설을 검정하므로** sign test**와 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

2.5. SAS 8.0를 이용하기

SAS에서 자료 분석을 하려면 외부 자료(Excel, ASC format, DB)는 SAS data로 만들어야 하므로 DATA step을 사용한다.

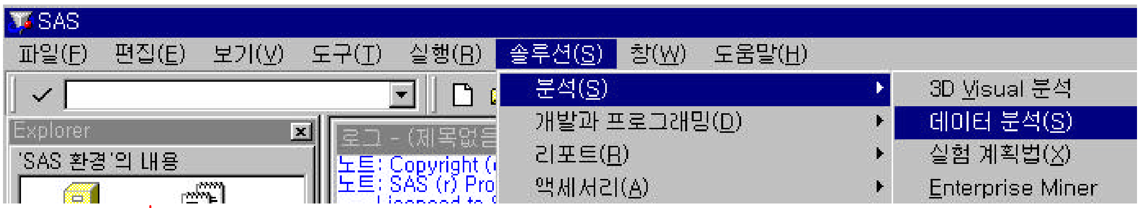
(예1) DATA one; input var1 var2; cards; ---- run;

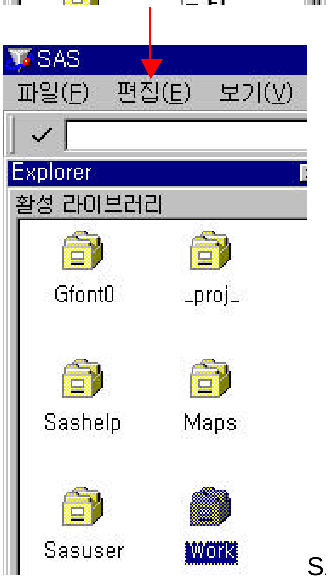
(예2) DATA one; infile “Text file 위치”; input var1 var2; run;

그러나 SAS version 8 이상에서는 spreadsheet 형태의 자료 입력 공간에서 자료를 직접 입력할 수 있다.

2.5.1. 자료 입력하기 (spreadsheet 형태)

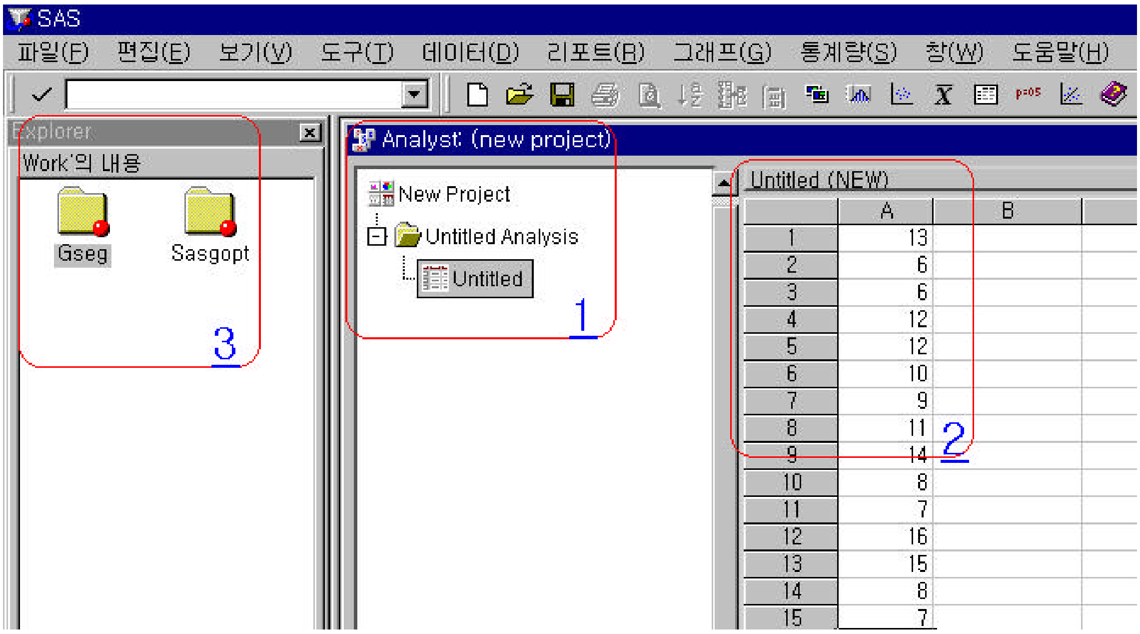
1) SAS를 실행한 후 Solution 메뉴에서 데이터 분석을 선택한다.





SAS 실행 중 만들어진 SAS data는 Work Library에 자장된다.

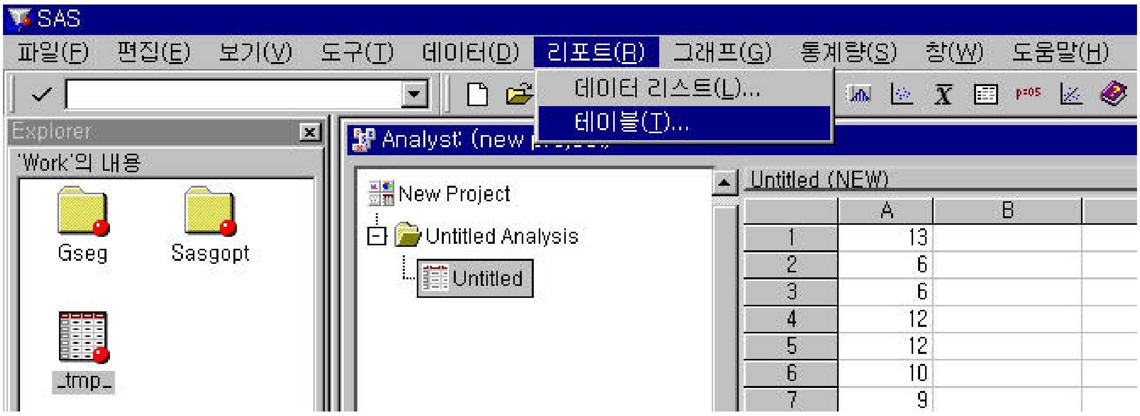
2) 분석(Analyst) 창(project)이 나타나고(1) 자료 입력을 위한 spreadsheet 형태(2)의 자료 입력 창이 나타난다. 그곳에 자료를 입력하면 된다. (숙제 #3, 2.7 예제 자료 몸무게) Explorer 창(3)은 SAS data가 만들어지는 과정을 보기 위하여 Work Library로 이동한 것이다.



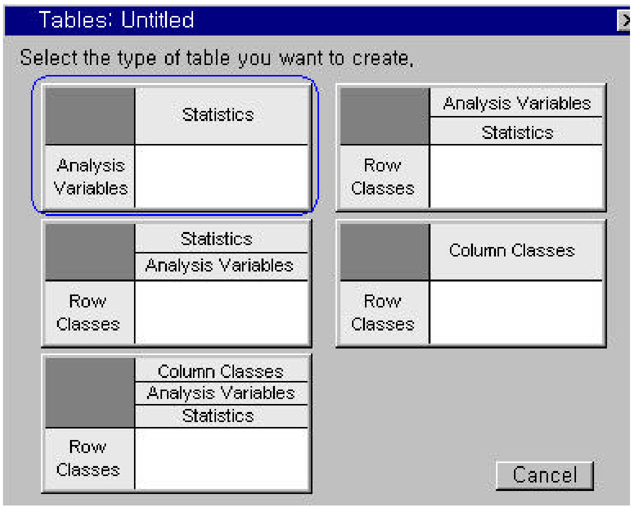
2.5.2. 기초 통계량 테이블 출력

다음은 몸무게 변수의 기초 통계량을 간단한 테이블 형태로 나타내기 위한 절차이다.

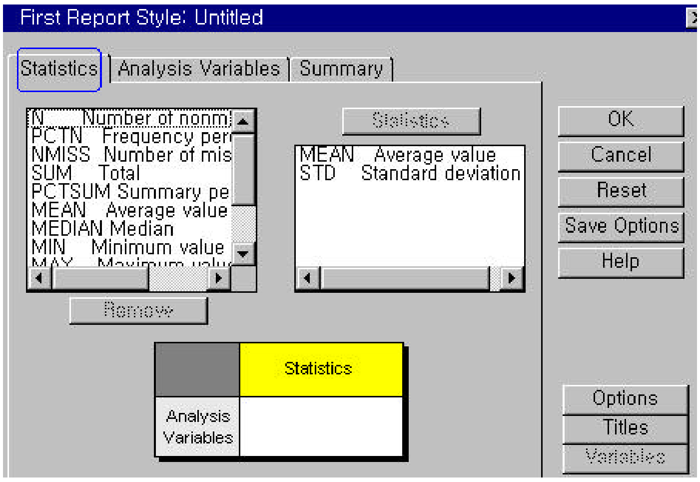
1) Report 메뉴에서 테이블을 선택한다.

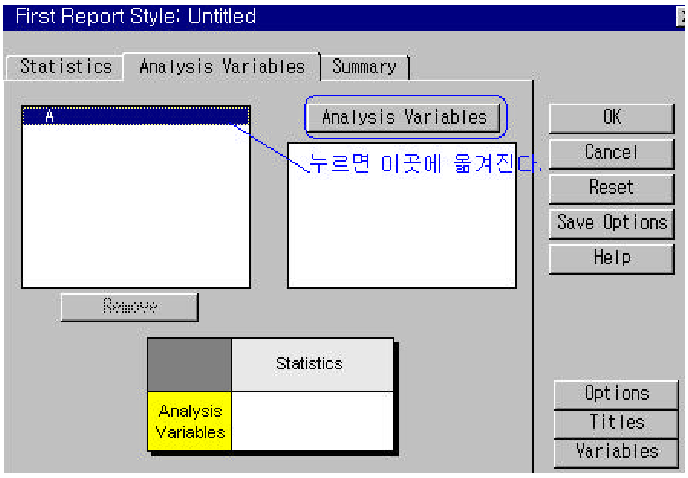


2) 테이블 형태를 지정한다. (예제에서는 블루 박스 테이블 선택)

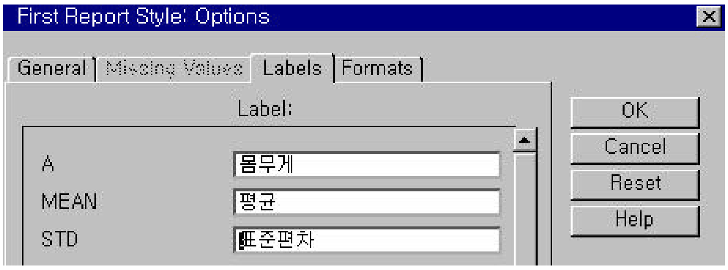


3) 환경 설정 팝업 메뉴가 나타나면 Statistics, Analysis Variables을 다음과 같이 설정한다.

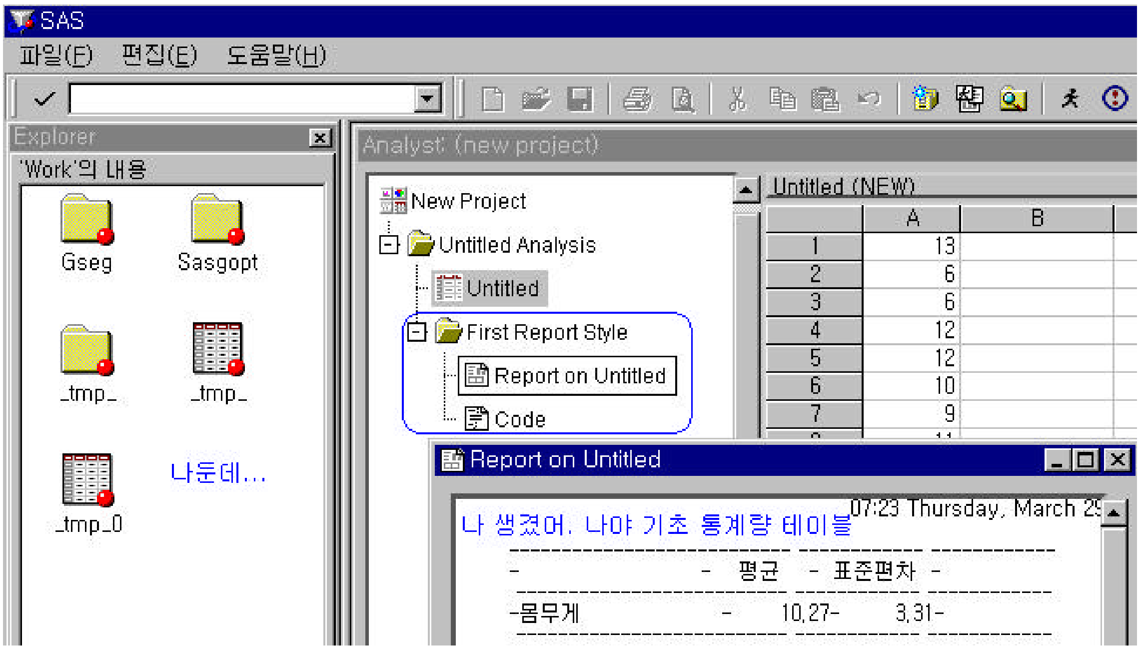




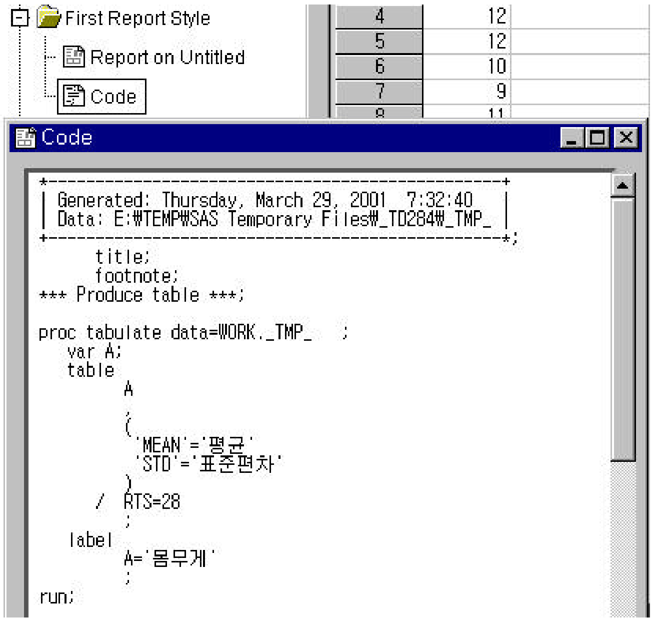
4) Options에서 Labels 폴더 환경을 다음과 같이 설정한다.



5) 환경 설정이 끝나면 (OK SK) 슥 아니다. 그냥 OK 버튼을 누르면 결과가 출력된다. 분석(Analyst) 창에 결과 보기를 위한 폴더가 생긴다. Default는 Report 창이다. 테이블 결과 출력은 물론 Temporary SAS data도 만들어진다.



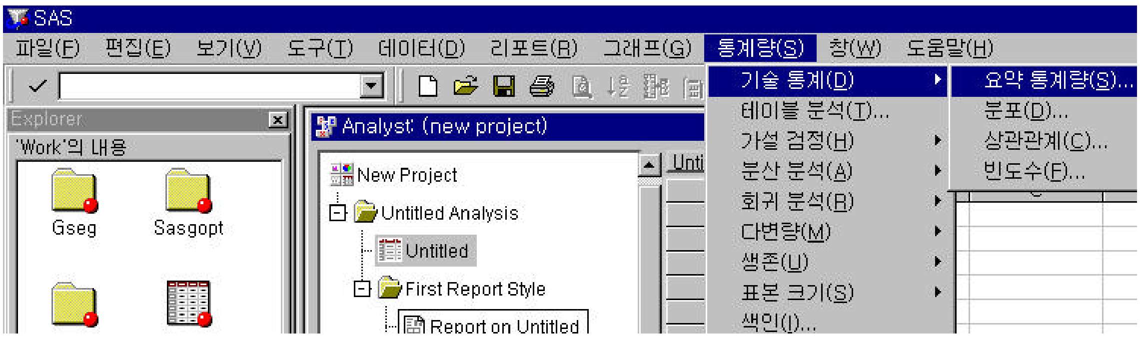
6) 앗 근데 Code 메뉴는 뭐지? 에이 한 번 눌러보자. 아니 SAS 코드가 열린다. 그럼 이것을 프로그램 에디터에서 실행하면 테이블을 만들 수 있나? 그렇구나. 그럼 내가 Tabulate procedure를 어떻게 사용하는 예제 프로그램을 얻었다는 것? 신난다.



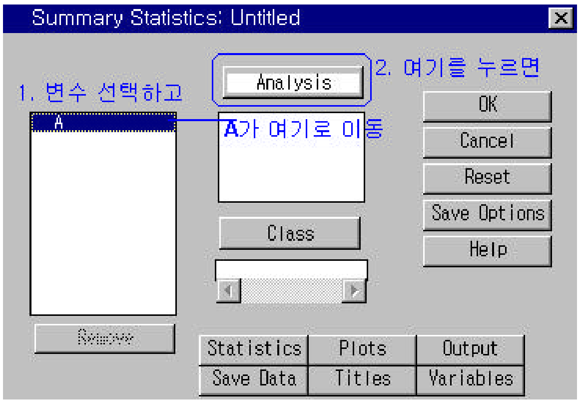
2.5.3. 기초 통계량 구하기

그럼 기초 통계량을 구하기 한 번 해 볼까? 자신이 생겼어요.

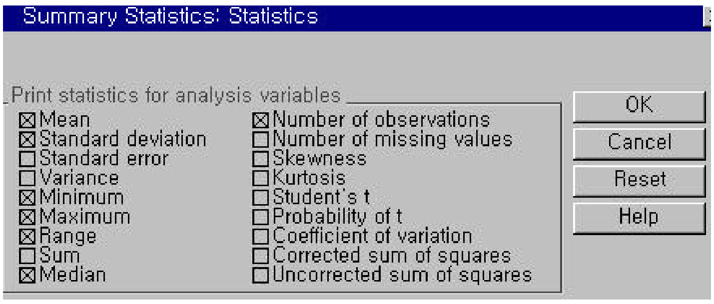
1) **메뉴를 선택하는 것이겠지? 그럼 당연히 통계량 메뉴에**...



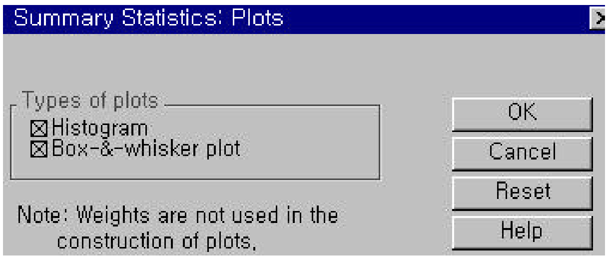
2) 예상대로 환경 설정 창이 pop-up되는구나. 이제 난 전문가야. 일단 기초 통계량을 구하기 위한 변수를 선택하자.



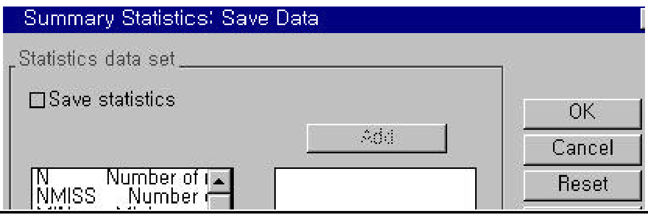
3) Statistics 설정 창을 열어 구하기 원하는 기초 통계량을 선택하고



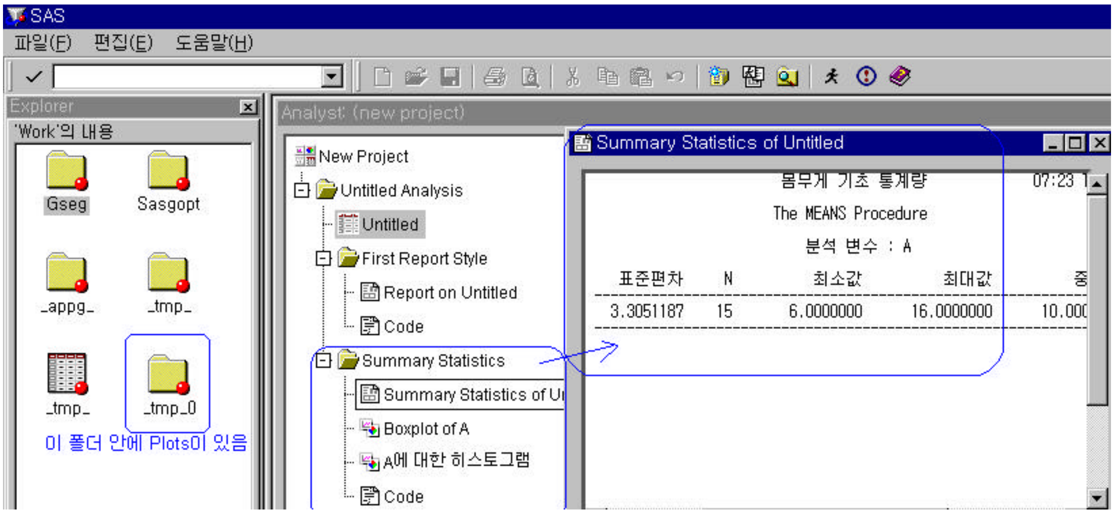
4) Plots 설정 창을 열어 원하는 그래프도 선택하자. 나머지는 쓸데없는데 시간 낭비말자.



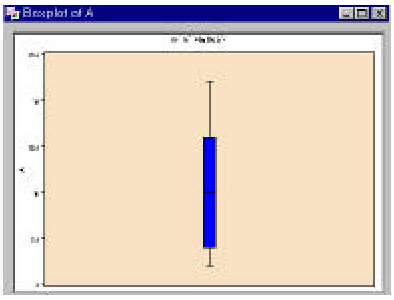
5) 그래도 다른 것은 궁금하지 않은데 Save Data 설정 창이 궁금하다. 아 기초통계량을 SAS data로 만들어 주는 곳이구나. 난 필요없는데...



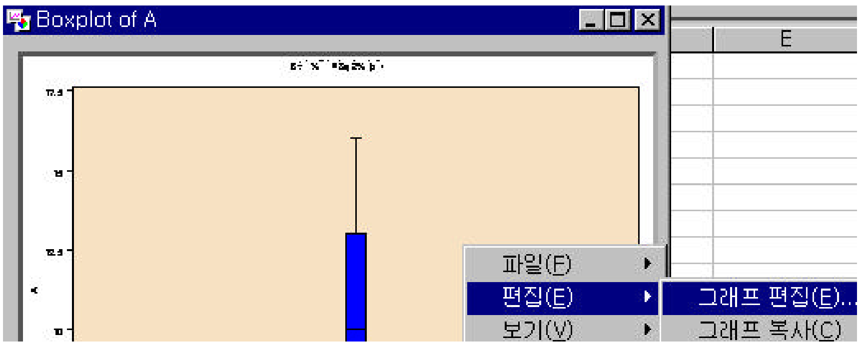
6) 이제 다 끝났나? 이번에는 실수말자. 그냥 OK 버튼만 누르면...



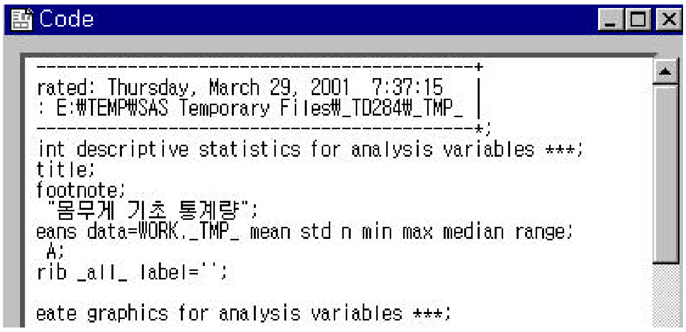
7) 와 대단하다. 결과가 나왔다. 그럼 정말 box plot (상자 그림)도 만들어졌나? 나 그냥 못 지나가지. Boxplot of A을 눌러보자.



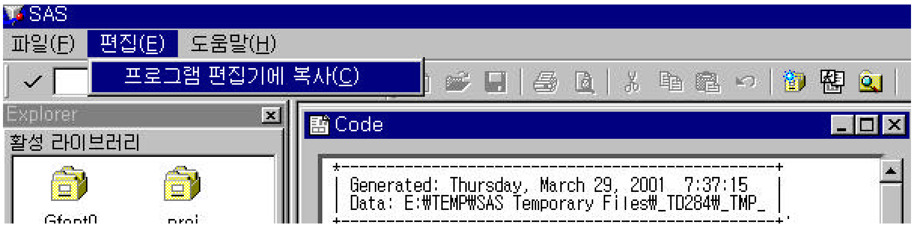
8) 근데 그림을 편집하고 싶은데 어쩌나? 메뉴가 없을까? 참 오른쪽 마우스 버튼이 있었지. 일단 오른쪽 마우스 버튼을 눌러보자.



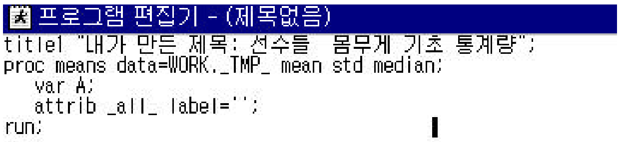
9) 앗 여기서도 Code 메뉴가 제공된다. 그럼 어떻게 프로그램이 되어 있는지 볼까?



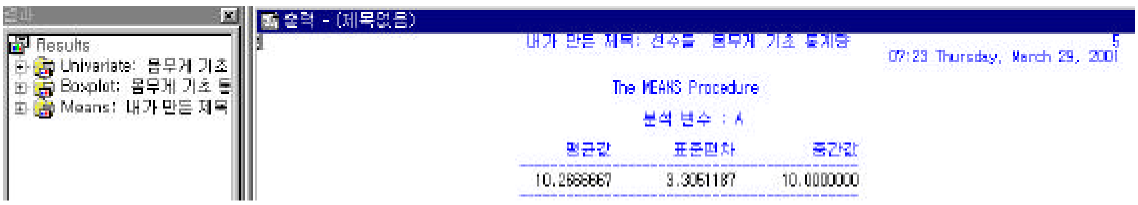
10) 음, 예상대로군. 그런데 나만의 프로그램을 원한다. 더 이상 SAS가 제공하는 프로그램은 싫다. 나는 나니까? 그럼 어떻게? 음... 전문가의 안목으로 돌아보니 편집이라는 메뉴가 보이는구나. 이 메뉴를 선택하면 프로그램이 프로그램 에디터(program editor) 창으로 옮겨지겠구나.



11) 다시 한 번 나의 천재성이 확인되는 순간이구나. 그럼 적당히 프로그램을 고치고 실행 버튼을 ( ) 누르거나 F8 키를 눌러 프로그램을 실행해 보자.



12) 정말 결과가 OUTPUT(출력) 창에 나타나네. 왼쪽에 있는 결과 창을 이용하여 이전 작업들의 결과를 볼 수 있습니다.

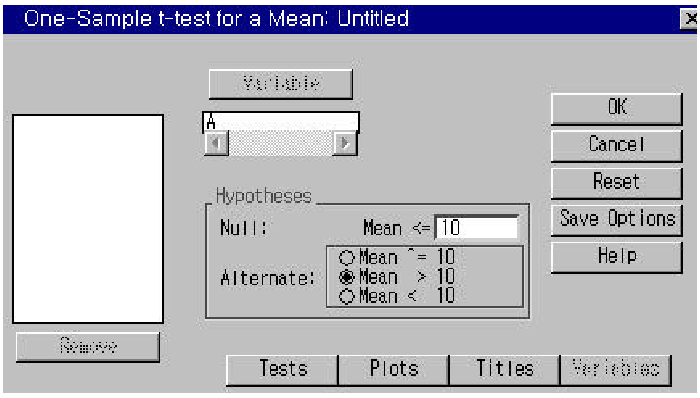


2.5.4. 일변량 분석하기(모 평균 검정)

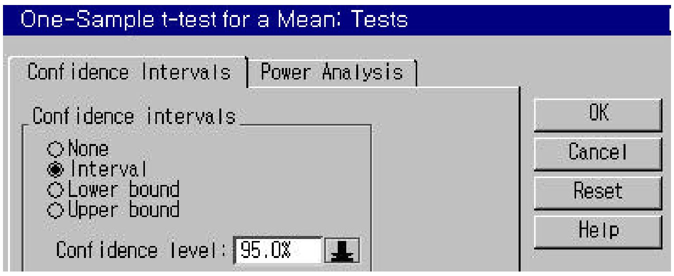
일변량 분석에서 모평균이 10 이상인가를 검정해볼까?

1) **어디를 선택하면 될까? 통계량 메뉴에 있겠지? 음 가설 검정, 그리고 t—검정**

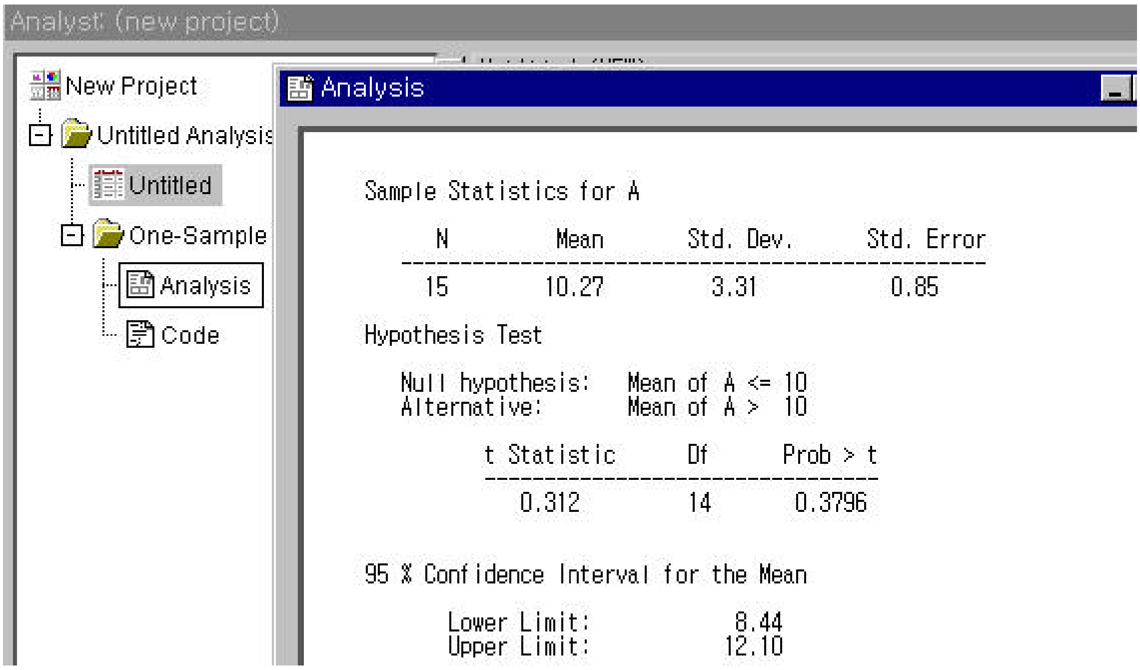
2) **변수**(variable)**를 선택하고 귀무가설 평균 값을 10을 입력하고 대립 가설을 설정한다.**



3) Tests **메뉴를 눌러 환경을 설정한다. 신뢰구간**(interval**은 양측 구간**)**을 선택하고 신뢰 수준**(significant level)**을 선택한다.**



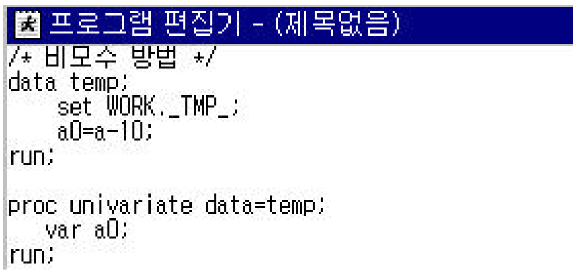
4) 이제 출력 결과를 해석하면 된다. p-value=0.3796이므로 유의수준 0.05 (5%)에서 귀무가설이 채택된다. 즉 몸무게는 10 이상이라고 말할 수 없다.



2.5.5. 비모수 방법은?

불행히도 메뉴에 없다. 아무리 찾아보아도 이제 절망이다. 내가 프로그램을 해야 하나? 그래도 이 정도는 가능하지. 우선 Code를 눌러 프로그램 에디터로 프로그램을 복사한다. (2.5.4.절 참조) 그리고 가볍게 수정하면 된다.

1) 중앙값이 10인가를 검정하는 것이므로 우선 원래 자료에서 10을 빼 주는 data step이 반드시 필요하다. 그리고 univariate procedure를 사용한다.



2) Sign Test(부호) 결과와 Wilcoxon Ranks-sum Test (부호 순위) 결과가 출력된다. p-value는 양측 검정을 사용할 경우 사용되는 값이다. $M = K - n/2$, $S = T_{+} - n(n+1)/4$ 이다.

위치모수 검정: Mu0=0

검정	--통계량---		-------p값--------	
스튜던드의 t	t	0.312484	Pr > |t|	0.7593
부호	M	0	Pr >= |M|	1.0000
부호 순위	S	4	Pr >= |S|	0.8250

3) Wilcoxon ranks-sum 검정의 경우 신뢰구간을 구하려면 각 관측치의 pairwise(쌍)의 평균이 필요하다. $n + {}_{n}C_{2}$ 의 수만큼 pairwise 평균을 계산하고 순서 통계량을 구해야 한다. 손으로 계산? 싫으면 다음 프로그램을...

확장 편집기 - 제목없음1 *

```
data one;
    input x1-x15;
    array xs{15} x1-x15;
    do i=1 to 15;
        do j=i to 15;
            x=(xs{i}+xs{j})/2; output;
        end;
    end;
    cards;
    13 6 6 12 12 10 9 11 14 8 7 16 15 8 7
;
run;

proc sort data=one;
    by x;
run;

proc print data=one;
    var x;
run;
```

```
19   7.5        90   12.0
20   7.5        91   12.0
21   8.0        92   12.0
22   8.0        93   12.0
23   8.0        94   12.0
24   8.0        95   12.0
25   8.0        96   12.0
26   8.0        97   12.5
27   8.0        98   12.5
28   8.5        99   12.5
29   8.5       100   12.5
30   8.5       101   12.5
31   8.5
32   8.5
33   8.5
34   9.0
35   9.0
36   9.0
```

▸▸**Point Estimator (점 추정치):** sign test에서는 자료의 중앙값이고 Wilcoxon ranks-sum test에서는 쌍 평균의 중앙값이다.

📖[Extra Homework]. CLASS 자료를 SAS version 8의 분석 tool을 이용하여 분석하는 절차를 강의 노트처럼 보이시오

1) 키는 몸무게에 영향을 미치나?

[이경화]

2) 성별, 성별에 따른 몸무게 평균, 몸무게 자료에 대한 적절한 그래프?

[손상록]

3. 독립인 두 모집단에 대한 비모수 추론

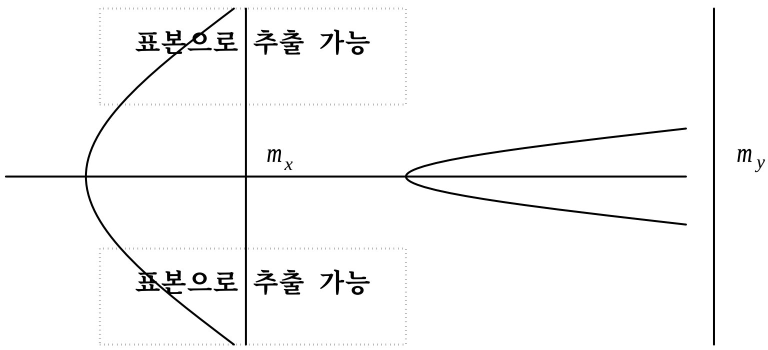
모집단의 중앙위치에 관한 추론은 두 모집단의 관측치가 서로 독립인 경우(independent samples)와 각 관측치가 쌍을 이룬 경우(paired samples)로 나누어진다. 쌍 이룬 두 모집단 중앙값 추론은 4장에서 다루기로 한다.

3.0. 모수적 방법

모수적 추론에서 중앙위치 통계량으로 (location parameter) 분포가 알려진 평균을 사용한다. 표본의 크기가 큰 경우(대표본) 표본 평균은 정규 분포를 따르고 (중심극한정리) 소표본의 경우는 모집단이 정규분포라는 가정이 만족되면 t-분포를 따른다. 독립인 두 모집단 중앙위치 차이에 대한 모수적 방법은 두 모집단의 평균의 차이가 있는지 검정도 이를 이용한다.

3.0.1. 등분산 가정

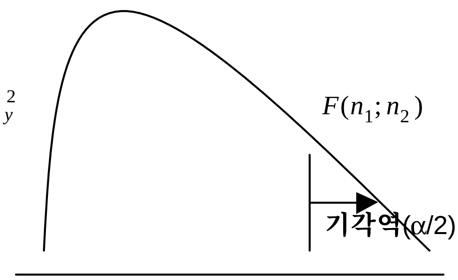
두 모집단 평균 차이에서 분산이 같다는 가정을 한다. 그 이유는 분산이 다를 경우 모집단의 평균이 같더라도 표본 평균이 같지 않을 가능성이 높게 된다.



1) 귀무가설: 두 모집단 분산은 같다. $s_x^2 = s_y^2$

2) 대립가설: 두 모집단 평균은 같지 않다. $s_x^2 \neq s_y^2$

3) 검정통계량: $T = \dfrac{\max(s_x^2, s_y^2)}{\min(s_x^2, s_y^2)} \sim F(n_1-1; n_2-1)$



n_1=분자 표본 크기, n_2=분모 표본 크기,

4) 검정방법: 검정통계량 T의 값이 기각역(critical region)에 속하면 귀무가설을 기각하고 그렇지 않으면 귀무가설을 채택한다. (cf) Hartley Test (모집단 3개 이상)

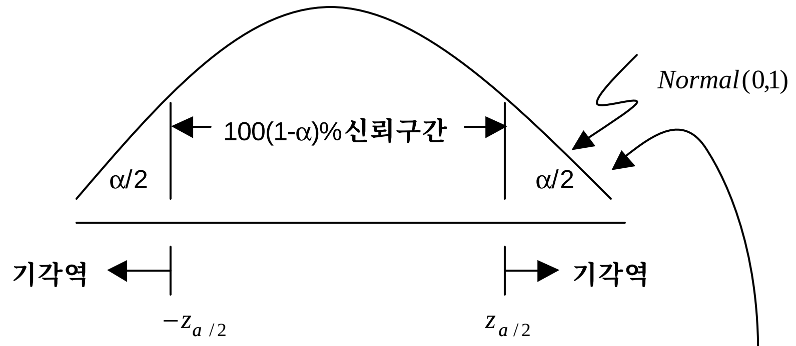
3.0.2. 대표본

1) **귀무가설**: **두 모집단 평균은 같다**. $m_x = m_y$
2) **대립가설**: **두 모집단 평균은 같지 않다**. $m_x \neq m_y$ **(양측검정)**
3) **검정통계량**:

$$T = \frac{(\bar{x}-\bar{y})-(m_x - m_y)}{s_p\sqrt{1/n_x + 1/n_y}} \sim N(0,1) \quad \text{where} \quad s_p = \sqrt{\frac{(n_x-1)s_x + (n_y-1)s_y}{(n_x+n_y-2)}}$$

Why? **중심극한정리**

4) **검정방법**: **검정통계량** T**의 값이 기각역**(critical region)**에 속하면 귀무가설을 기각하고 그렇지 않으면 귀무가설을 채택한다**.



5) 100(1-α)% **신뢰구간**: $(\bar{x}-\bar{y}) \pm z_{a/2} s_p \sqrt{1/n_x + 1/n_y}$

3.0.2. 소표본

1) **가정**: **각 모집단의 분포는 정규분포를 따른다**.
2) **귀무가설**: **두 모집단 평균은 같다**. $m_x = m_y$
3) **대립가설**: **두 모집단 평균은 같지 않다**. $m_x \neq m_y$ **(양측검정)**
4) **검정통계량**:

$$T = \frac{(\bar{x}-\bar{y})-(m_x - m_y)}{s_p\sqrt{1/n_x + 1/n_y}} \sim t(n_x + n_y - 2)$$

5) **검정방법**: **검정통계량** T**의 값이 기각역**(critical region)**에 속하면 귀무가설을 기각하고 그렇지 않으면 귀무가설을 채택한다**. $t(a/2; n_x + n_y - 2)$

6) 100(1-α)% **신뢰구간**: $(\bar{x}-\bar{y}) \pm t(a/2; n_x + n_y - 2) s_p \sqrt{1/n_x + 1/n_y}$

3.0.3. SAS 이용하기

19명 학생의 몸무게(pound) 자료이다. (SAS JMP class 자료)

확장 편집기 - 제목없음1 *

```
data one;
    input gender $ weight @@;
    cards;
    F   98   M   112.5   F   84   F   102.5   M   102.5
    M   83   F   84.5    F   112.5   M   84   M   99.5
    F   50.5   F   90   F   77   F   112   M   150
    M   128  M   133  M   85   M   112
run;
proc ttest data=one;
    class gender;
    var weight;
run;
```

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
weight	Pooled	~~Equal~~	~~17~~	~~-1.93~~	~~0.0702~~
weight	Satterthwaite	Unequal	17	-1.95	0.0680

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
weight	Folded F	9	8	1.37	$\boxed{0.6645}$

*)등분산 가정이 만족하지 않는 경우 근사 t-검정이나 근사 자유도를 이용한 t-검정 (Satterthwaite 방법)을 이용하여 두 모집단 평균 차이 검정을 해야 한다.

HOMEWORK #6 [due 4월 10일]

3.1 절의 homework #6(3.2) 자료에서 각 모집단이 정규분포를 따른다고 가정하고 모수 방법으로 알약의 두께 평균의 차이가 있는지 검정하시오. (유의수준=0.05)

3.1. 독립인 두 모집단 중앙 위치 차이에 대한 비모수 추론

3.1. Median Test (중앙값 검정)

1) 가정 (assumption)

- $(x_1, x_2, \ldots, x_{n1})$는 중앙값 M_x인 모집단으로부터의 확률 표본(random sample)이다.
- $(y_1, y_2, \ldots, y_{n2})$는 중앙값 M_y인 모집단으로부터의 확률 표본(random sample)이다.
- 두 확률 표본은 서로 독립이다.
- 두 모집단이 같은 중앙값을 갖는다면 각 표본 관측치가 총 중앙값(grand median)보다 클 확률은 동일하다. 이를 이용하여 검정통계량이 유도된다.

2) 가설 (hypothesis)

- 귀무 가설: $M = M_x = M_y$
- 대립 가설: $M_x \neq M_y$ (median 검정의 경우 단측 검정도 가능하지만 매우 복잡)

3) 검정 통계량 (test statistic)

① 귀무가설이 맞다면 각 표본에서 측정치의 반은 총 중앙값(M)보다 클 것이다.

② 총 중앙값(M)의 추정치는 두 표본 관측치 모두를 합하여 계산한 중앙값이다. 즉 $(n_1 + n_2)$개의 관측치 중 중앙값을 계산한다.

③ 위에서 계산된 중앙값에 의해 다음과 같이 교차표를 계산하다.

	표본		
계산된 중앙값보다	1 (X)	2 (Y)	total
큰 관측치 수	A	B	A+B
작은 관측치 수	C	D	C+D
Total	A+C=n_1	A+C=n_2	N=n_1+n_2

④ 만약 귀무가설이 $A = C = n_1/2$, $B = D = n_2/2$가 만족될 것이다.

⑤ $n_1 + n_2$이 홀수인 경우 계산된 중앙값과 관측치가 같아지는 경우가 발생하는 경우 그 관측치를 버리고 검정 통계량을 계산하거나 중앙값보다 작은 관측치에 포함한다. (귀무가설 기각할 확률을 적게 만든다)

⑥ Hyper-geometric distribution

주머니에 동일한 구슬이 N개가 들어 있는데 그 중 파랑 구슬이 M개이고 나머지는 빨강 구슬(N-M)이 들어 있다. 구슬을 K개를 (without replacement: 비복원) 무작위로 추출했을 경우 파랑 구슬의 수를 x라 할 경우 x는 초기하 분포를 따른다고 한다.

$$H(x \mid N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, x = 0,1,2,\ldots,K, \quad E(x) = \frac{KM}{N}, \quad V(x) = \frac{KM}{N}\left(\frac{(N-M)(N-K)}{N(N-1)}\right)$$

⑦ 초기하 분포의 Median Test 이용: N은 구슬 수, n_1는 주머니 파랑 구슬 수, n_2는 주머니 빨강 구슬 수, A는 파랑 구슬 수 x, (A+B)는 무작위로 뽑은 구슬 K이다.

$$H(A|N,M,K)=\frac{\binom{n_1}{A}\binom{n_2}{B}}{\binom{N}{A+B}},\ A=0,1,2,\ldots,(A+B)$$

⑧ Binomial approximation to Hyper-geometric distribution

- 대표본(N이 클 경우) 의미는 $\min(np,npq)>5$이어야 한다.
- 대표본이 성립하면 $H(x|N,M,K)\sim Binomial(x|n=K,\ p=M/N)\sim Nomal(np,npq)$
- 검정 통계량 분포

$$T=\frac{(A/n_1)-(B/n_2)}{\sqrt{\hat{p}(1-\hat{p})(1/n_1+1/n_2)}}\sim Normal(0,1):\ \text{중심극한정리}$$

⑨ 초기하 분포표를 이용하거나 정규분포 근사를 이용하여 가설을 검정한다.

II EXAMPLE II 심근 경색 환자들의 두 그룹(X, Y) 심장 발작 지수의 차이가 있는지 비모수 검정하시오. (유의수준=0.05)

그룹 X	25 25 17 26 18 30 24 21 13 30 20 23 26 12 20 37 9 17 37 20 11 32 16 31 46 20 25 17 36 54 8 26
그룹 Y	31 21 38 19 38 41 68 28 43 42 30 20 29 13 32 30

1) 귀무가설: 두 모집단 중앙값은 같다. $M_x=M_y$
2) 대립가설: 두 모집단 중앙값은 같지 않다. $M_x\neq M_y$ (양측검정)
3) 검정통계량:

	표본		
계산된 중앙값보다	X	Y	total
큰 관측치 수	12	12	24
작은 관측치 수	20	4	24
Total	32	16	48

$$T=\frac{(12/32)-(12/16)}{\sqrt{0.5(1-0.5)(1/32+1/16)}}=-2.45\quad \hat{p}=(12+12)/48=0.5$$

4) 결론: 검정통계량 값이 기각역에 속하므로 (<-1.96) 귀무가설을 기각한다. 즉 X 그룹과 Y 그룹 환자들의 발작 지수가 서로 다르다고 말할 수 있다. 위의 교차표를 살펴보면 X 그룹의 환자들의 지수가 더 높다고 할 수 있다.(why?)

⏭ 참고: $n(=n_1+n_2)$이 큰 경우 Median test는 동질성(homogeneity) 검정을 위한 c^2-검정을 (교차표 검정) 이용할 수 있다. How?

📖 **HOMEWORK #6 계속** [due 4월 10일]

The quality control manager with a drug manufacturer wishes to know whether two methods of producing a particular tablets result in a difference between the median thick-nesses. A random sample of tablets is drawn from batches produced by the two methods. The following table shows the results, which have been coded for computational convenience. Do these data provide sufficient evidence to indicate that the two population medians are different? Let $a=0.05$

Method	Thickness
A	51 42 45 48 52 44 58 41 52 44 45 52 61 60 41
B	40 47 36 39 37 46 43 55 53 56

3.2. Mann-Whitney Test

1) 가정 (assumption)
   - $(x_1, x_2, \ldots, x_{n1})$ 는 중앙값 M_x 인 모집단으로부터의 확률 표본이고 $(y_1, y_2, \ldots, y_{n2})$ 는 중앙값 M_y 인 모집단으로부터의 확률 표본이다.
   - 두 확률 표본은 서로 독립이다.
   - 두 모집단이 다르다면 중앙의 위치에만 차이가 있다.
2) 가설 (hypothesis)
   - 귀무 가설: $M = M_x = M_y$
   - 대립 가설: $M_x \neq M_y$ (양측 검정) $M_x < M_y$ 혹은 $M_x > M_y$ (단측 검정)
3) 검정 통계량 (test statistic)

   ① 두 표본 $(x_1, x_2, \ldots, x_{n1})$ 과 $(y_1, y_2, \ldots, y_{n2})$ 을 크기 순으로 정렬한 후 순위(rank)를 매긴다. 관측치 크기가 같으면 평균으로 순위를 사용한다.

   ② 만약 모집단 X가 Y보다 중앙값이 크다면 $(x_1, x_2, \ldots, x_{n1})$ 의 순위 합이 $(y_1, y_2, \ldots, y_{n2})$ 순위 합보다 클 것이다. 반대의 경우도 같은 개념이 성립한다.

   ③ 검정통계량 $T = S - \dfrac{n_1(n_1+1)}{2}$, S=표본1 관측치의 순위 합.
4) 결정 이론

   ① 양측 검정일 경우 T가 매우 크거나 작은 경우 귀무가설을 기각한다. 표C에서 (Mann-Whitney 표: $w_{a/2}$ 기각역) 계산된 검정 통계량이 $w_{a/2}$ 보다 적거나 $w_{1-a/2}$ 보다 큰 경우 귀무 가설을 기각한다. $w_{1-a/2} = n_1 n_2 - w_{a/2}$

   ② 대립 가설이 $M_x < M_y$ 단측 검정은 T가 w_a 보다 적은 경우 귀무 가설 기각

   ③ 대립 가설이 $M_x > M_y$ 단측 검정은 T가 w_{1-a} 보다 큰 경우 귀무 가설 기각

⏭ **Large sample Approximation:** n_1 이나 n_2 가 20 이상이어서 표C를 사용할 수 없는 경우 다음 정규분포 근사를 이용하여 가설을 검정하면 된다.

$$z = \frac{T - n_1 n_2 / 2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

Ⅱ EXAMPLE Ⅱ 두 그룹(X, Y) 사이의 점수의 차이가 있는지 비모수 검정하시오. (유의수준 =0.05)

그룹 X	11.9	11.7	9.5	9.4	8.7	8.2	7.7	7.4	7.4
	7.1	6.9	6.8	6.3	5	4.2	4.1	2.2	
그룹 Y	6.6	5.8	5.4	5.1	5	4.3	3.9	3.3	2.4	1.7

1) 귀무가설: 두 모집단 중앙값은 같다. $M_x = M_y$
2) 대립가설: 두 모집단 중앙값은 같지 않다. $M_x \neq M_y$ (양측검정)
3) 검정통계량:

$$T = 296.5 - 17(17+1)/2 = 143.5$$

4) 결론: 표C에서 $n_1 = 17, n_2 = 10$ 경우 $w_{a/2} = w_{0.025}$ =46 & $w_{1-a/2} = w_{0.975} = 17*10-46$ =124 이므로 검정통계량은 기각역에 속한다. 두 그룹의 중앙값은 차이가 있고 그룹 X의 중앙값이 크다고 말할 수 있다. (Why?)

⏭ **p-value를 구할 수 있나?:** Yes, but approximation. $n_1 = 17, n_2 = 10$.인 경우 $T = 143.5$ 는 $(17)(10) - 26 = 144$ 와 $(17)(10) - 35 = 135$ 사이 값이므로 $2 \times 0.005 > p > 2 \times 0.001$ 사이의 값이다.

SAS 이용하기

Editor - Untitled1 *

```
data example;
    do i=1 to 17;
        input y @@; group=1;output;
    end;
    do i=1 to 17;
        input y @@; group=2;output;
    end;
    cards;
    11.9    11.7    9.5    9.4    8.7    8.2    7.7    7.4    7.4
    7.1     6.9     6.8    6.3    5      4.2    4.1    2.2
    6.6     5.8     5.4    5.1    5      4.3    3.9    3.3    2.4    1.7
;
run;

proc npar1way median wilcoxon data=example;
    class group;
    var y;
run;
```

```
            Wilcoxon Scores (Rank Sums) for Variable y
                  Classified by Variable group

                       Sum of      Expected       Std Dev          Mean
group          N       Scores      Under H0      Under H0         Score
---------------------------------------------------------------------------
1             17       296.50         238.0     19.910412     17.441176
2             10        81.50         140.0     19.910412      8.150000

               Average scores were used for ties.


   Wilcoxon Two-Sample Test

Statistic              81.5000

Normal Approximation   우리가 사용한 검정 통계량 값이랑 다르네?
Z                      -2.9130
One-Sided Pr <  Z       0.0018
Two-Sided Pr > |Z|      0.0036

t Approximation
One-Sided Pr <  Z       0.0036
Two-Sided Pr > |Z|      0.0073

Z includes a continuity correction of 0.5.

                           표본이 3개 이상인 경우 사용하는 검정방법
     Kruskal-Wallis Test

Chi-Square              8.6328
DF                           1
Pr > Chi-Square         0.0033


    Median Scores (Number of Points Above Median) for Variable y
                  Classified by Variable group

                       Sum of      Expected       Std Dev          Mean
group          N       Scores      Under H0      Under H0         Score
---------------------------------------------------------------------------
1             17         12.0      8.185185      1.277644      0.705882
2             10          1.0      4.814815      1.277644      0.100000

               Average scores were used for ties.


   Median Two-Sample Test

Statistic               1.0000
Z                      -2.9858
One-Sided Pr <  Z       0.0014
Two-Sided Pr > |Z|      0.0028


   Median One-Way Analysis

Chi-Square              8.9151
DF                           1
Pr > Chi-Square         0.0028
```

📖 HOMEWORK #6 계속 [due 4월 17일]

EXERCISES

3.3 West (E4) conducted an experiment with adult aphasic subjects, in which each was required to respond to one of 62 commands. Five subjects received an experimental treatment program, and five controls received conventional speech therapy. Table 3.8 shows the percentage of correct responses of each subject in the two groups following treatment. Do these data provide sufficient evidence to indicate that the experimental treatment improves the proportion of correct responses? Let $\alpha = 0.05$. What is the P value?

TABLE 3.8 Percentage of correct responses to 62 commands by aphasic subjects in two treatment groups

Experimental (X)	73	42	90	58	62
Control (Y)	50	23	68	40	45

Source: Joyce A. West, "Auditory Comprehension in Aphasic Adults: Improvement through Training," *Arch. Phys. Med. Rehabil.*, 54 (1973), 78–86

3.4 Table 3.9 shows the tidal volume of 37 adults suffering from atrial septal defect. In 26 of these, pulmonary hypertension was absent, and in 11 it was present. The data were reported by Ressl et al. (E5). Do these data provide sufficient evidence to indicate a lower tidal volume in subjects without pulmonary hypertension? Let $\alpha = 0.05$. What is the P value?

TABLE 3.9 Tidal volume, in milliliters, in two groups of subjects

Pulmonary hypertension absent		*Pulmonary hypertension absent*	
Case	**(X)**	**Case**	**(Y)**
1	652	1	676
2	556	2	556
3	618	3	493
4	500	4	348
5	500	5	530
6	526	6	780
7	511	7	569
8	538	8	546
9	440	9	766
10	547	10	819
11	605	11	710
12	500		
13	437		
14	481		
15	572		
16	589		
17	605		
18	436		
19	724		
20	515		
21	552		
22	722		
23	778		
24	677		
25	680		
26	428		

Source: J. Ressl, M. Kubis, P. Lukl, J. Vykydal, and J. Weinberg, "Resting Hyperventilation in Adults with Atrial Septal Defect," *Br. Heart. J.*, 31 (1969), 118–121; used by permission of the authors and the editor.

3.3. 신뢰구간

일변량 신뢰구간과 같이 귀무가설(두 모집단 중앙값은 같다 $M_x = M_y$)을 기각하지 않는 모수들의 영역이 가장 많이 사용된다. 유의수준 α 하에서 귀무가설을 기각하지 않는 모수의 영역들을 100(1-α)% 신뢰 구간이라 한다.

Median Test를 이용한 신뢰 구간

Median 검정의 경우 Approximation 이론에 의해 다음을 검정 통계량으로 사용하였다.

$$T = \frac{(A/n_1) - (B/n_2)}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} \sim Normal(0,1)$$

그러므로 두 모집단 중앙값 차이($M_x - M_y$)의 100(1-α)% 신뢰구간은

$$(A/n_1) - (B/n_2) \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}, \quad \hat{p} = \frac{A+B}{n_1+n_2}$$

ΙΙ **EXAMPLE** ΙΙ Median Test에 있는 심근 경색 환자들의 두 그룹(X, Y) 심장 발작 지수 중앙값 차이에 대한 95% 신뢰수준을 구하시오. (43 page)

$$(12/32) - (12/16) \pm 1.96\sqrt{(24/48)(1-24/48)(1/32+1/16)}$$

Mann-Whitney Test를 이용한 신뢰 구간

M-W 방법은 순서를 고려한 방법이므로 다음 절차에 의해 신뢰구간을 구한다.

1) 우선 각 표본 관측치를 크기 순으로 나열한 후 교차 표를 만든다. (X를 열, Y를 행)
2) 우선 두 모집단의 차이($x_i - y_i$)를 구하여 (개수=$n_1 n_2$) 이를 크기 순으로 정렬한다.
3) Mann-Whitney 표로부터 (n_1, n_2)에 대응하는 $w_{\alpha/2}$를 구한다.
4) (100-α)% 신뢰구간의 하한(lower bound)은 $w_{\alpha/2}$번째 작은 차이이고, 상한(upper bound)은 $w_{\alpha/2}$번째 큰 차이이다.

II EXAMPLE II 다음 표본 자료를 이용하여 두 모집단 중앙값 차이에 대한 95% 신뢰구간을 구하시오.

X	77	82	85	86	86	86	89	91	92	93	100
Y	65	65	73	75	77	78	83	85	90	97	

1) 우선 각 표본 관측치를 크기 순으로 나열한 후 교차 표를 만든다. (X를 열, Y를 행)

TABLE 3.11 **Array of differences between *X* and *Y* values in Example 3.4**

	X										
Y	77	82	85	86	86	86	89	91	92	93	100
65	12	17	20	21	21	21	24	26	27	28	35
65	12	17	20	21	21	21	24	26	27	28	35
73	4	9	12	13	13	13	16	18	19	20	27
75	2	7	10	11	11	11	14	16	17	18	25
77	0	5	8	9	9	9	12	14	15	16	23
78	−1	4	7	8	8	8	11	13	14	15	22
83	−6	−1	2	3	3	3	6	8	9	10	17
85	−8	−3	0	1	1	1	4	6	7	8	15
90	−13	−8	−5	−4	−4	−4	−1	1	2	3	10
97	−20	−15	−12	−11	−11	−11	−8	−6	−5	−4	3

2) 우선 두 모집단의 차이$(x_i - y_i)$를 구하여 (개수$=n_1 n_2$) 이를 크기 순으로 정렬한다.

3) Mann-Whitney 표로부터 $(11,10)$에 대응하는 $w_{a/2}$ 값은 27이다.

4) 95% 신뢰구간의 하한(lower bound)은 27번째 작은 차이 값 1이고 상한(upper bound)은 27번째 큰 차이 값인 17이다.

3.4. 산포도 차이 검정

3.4.1. 모수적 방법

독립인 두 모집단의 산포도(dispersion) 차이가 있는지 검정하는 방법

- 모수적 방법: F-검정
  1) 귀무가설: 두 모집단 산포도는 같다. $s_x^2 = s_y^2$
  2) 대립가설: 두 모집단 산포도는 같지 않다. $s_x^2 \neq s_y^2$ (양측검정)
  3) 검정통계량: $F = \dfrac{\max(s_x^2, s_y^2)}{\min(s_x^2, s_y^2)} \sim F(n_1, n_2)$
  4) 결론: 표에 의해 검정 통계량 값이 기각역에 속하면 산포도가 같지 않다.

HOMEWORK #7 [due 5월 8일]

미국과 한국의 소득 분포의 산포도 차이가 있는지 알아보기 위하여 자료 조사하여 다음을 얻었다. (단위: 천$) 산포도가 같은지 검정하시오. (유의수준=0.05)

미국 $n_x = 50 / \bar{x} = 50 / s_x = 3.12$ 한국 $n_y = 30 / \bar{x} = 10 / s_x = 2.87$

3.4.2. 비모수 방법: Ansari –Bradley 방법

1) 귀무가설: 두 모집단 산포도는 같다. $s_x^2 = s_y^2$
2) 대립가설: 두 모집단 산포도는 같지 않다. $s_x^2 \neq s_y^2$ (양측검정)
3) 검정통계량: 두 모집단 표본 자료를 하나로 하여 크기 순으로 정렬한다. 양 끝에서 순위를 부여한다. X 집단의 순위 합을 검정 통계량으로 사용한다. Ansari –Bradley 표를(n_1=X집단 표본 수, n_2=Y집단 표본 수) 이용하여 가설 검정한다.

EXAMPLE 그룹(X, Y) 사이의 산포도의 차이가 있는지 검정하시오. (유의수준=0.05)

그룹 X	3.84	2.6	1.19	2	
그룹 Y	3.97	2.5	2.7	3.36	2.3

1) 귀무가설: 두 모집단 산포도는 같다.
2) 대립가설: 두 모집단 산포도는 같지 않다. (양측검정)
3) 검정통계량:

관측치 정렬	1.19	2	2.3	2.5	2.6	2.7	3.36	3.84	3.97
집단 (group)	X	X	Y	Y	X	Y	Y	X	Y
순위	1	2	3	4	5	4	3	2	1

검정통계량 T=1+2+5+2=10 (n_1=4, n_2=5) 유의수준 0.05이므로 0.025와 0.975에 가장 근사한 값의 T를 찾으면 T=16(0.0159)과 8(0.9603)이다. 그러므로 계산된 검정 통계량 값이 8이하이거나 16이상인 경우 기각한다.

4) 결론: 계산된 검정 통계량 값이 10이므로 기각역($T \ge 16$ 혹은 $T \le 8$)에 속하지 않으므로 유의수준 5%하에서 귀무가설을 기각할 수 없다. 도 모집단의 산포도는 같다고 할 수 있다.

⏭ **Large sample Approximation:** Ansari-Bradley 표는 $n_1 + n_2 \le 20$ 인 경우 제한적인 값에 대해 표를 제공하고 있다. 표본 수가 큰 경우 다음 근사 분포를 이용하면 된다.

$$T^* = \frac{T - [n_1(n_1+n_2+2)/4]}{\sqrt{n_1 n_2 (n_1+n_2+2)(n_1+n_2-2)/[48(n_1+n_2-1)]}} \sim N(0,1) \text{ if } n_1+n_2 \text{이 짝수}$$

$$T^* = \frac{T - [n_1(n_1+n_2+1)^2/4(n_1+n_2)]}{\sqrt{n_1 n_2 (n_1+n_2+1)(3+(n_1+n_2)^2)/48(n_1+n_2)^2}} \sim N(0,1) \text{ if } n_1+n_2 \text{이 홀수}$$

HOMEWORK #7 [계속]

다음은 자폐증 증상이 있는 그룹과 그렇지 않은 그룹의 뇌 Dopamine 흡수 정도를 측정한 것이다. 산포도가 같은지 비모수 검정하시오. (유의수준=0.1)

자폐증	433	347	328	607	478
증상 없음	428	372	434	425	336

SAS 이용하기

Editor - Untitled1 *

```
data nonpa1;
    input group response @@;
    cards;
    1 3.84 1 2.6 1 1.19 1 2
    2 3.97 2 2.5 2 2.7 2 3.36 2 2.3
run;
proc npar1way ab;
    class group;
    var response;
run;
```

```
           Ansari-Bradley Scores for Variable response
                 Classified by Variable group

                      Sum of      Expected       Std Dev          Mean
group          N      Scores      Under H0       Under H0        Score
-------------------------------------------------------------------------
1              4        10.0     11.111111       2.078699         2.50
2              5        15.0     13.888889       2.078699         3.00

Ansari-Bradley Two-Sample Test

Statistic              10.0000
Z                      -0.5345
One-Sided Pr <  Z       0.2965
Two-Sided Pr > |Z|      0.5930
```

4. 쌍 이룬 (paired) 두 모집단

"before and after", "pre and post" 자료로부터 처리 효과에 (treatment effect) 유무에 대한 차이 검정을 실시할 경우 두 집단은 서로 독립이 아니다. 이런 형태 자료를 짝진(paired) 표본이라 한다. 3장의 독립인 두 모집단에 대한 추론과는 달리 짝진 표본의 경우 짝 이룬 두 표본의 차이가 유의한지(0인지)를 추론하게 된다.

독립인 두 모집단 처리 효과 초론 ⇔ 짝진 표본의 처리 효과 검정

$$u_x = u_y ??? \quad \bar{x} - \bar{y} \quad \Leftrightarrow \quad \overline{X - Y} = 0 ??? \quad \overline{d_i} = \overline{x_i - y_i}$$

짝진 집단의 자료는 다음과 같이 표현된다. $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

4.0. 모수적 방법 (paired t-test)

관측치의 차이 $d_i = (x_i - y_i)$ 이용하여 차이를 검정하게 된다.

1) 가설 (hypothesis)
   - 귀무가설: $m_d = 0$ (처리 효과의 차이가 없다, 집단의 평균 차이가 없다)
   - 대립가설: $m_d \neq 0$ (양측 검정) $m_d > 0$ 혹은 $m_d < 0$ (단측 검정)

2) 검정 통계량 (test statistic)
   ① 쌍을 이룬 관측치에 대해 $d_i = (x_i - y_i)$ 계산한다.
   ② d_i 의 평균($\bar{d}$)을 구한다. 이제 이를 이용하여 일변량 모집단 평균에 대한 가설 검정을 시행하면 된다.
   ③ 검정통계량 $T = \dfrac{\overline{D}}{s_d / \sqrt{n}} \sim Normal(0,1)$ (대표본) $\sim t(n-1)$ (소표본: 모집단이 정규분포라는 가정이 필요)

3) 결정 (decision rule): 정규분포나 t-분포를 이용하여 p-value를 구하거나 기각역 을 이용하여 가설을 검정하면 된다.

📖 **HOMEWORK #8** [due 5월 10일] 다음은 10명에 대해 광고 전후 상품에 대한 선호도를 조사한 자료이다. 광고 전과 공고 후의 상품 선호도가 차이가 있는지 검정하시오. 만족도 점수의 분포는 정규분포를 따른다. (유의수준=0.05) SAS를 이용하시오.

사람	1	2	3	4	5	6	7	8	9	10
광고 전	50	25	30	50	60	80	45	30	65	70
광고 후	53	27	38	55	61	85	45	31	72	18

4.1. Sign Test

대응하는 관측치의 차이의 부호를 이용하여 차이의 유의성을 검정하는 방법이다. + 부호와 – 부호의 수가 비슷하면 차이가 없는 것이고 한쪽 부호의 수가 월등히 많으면 귀무가설(차이가 없다)을 기각하게 된다. 이제는 이 방법 사용에 익숙해졌으리라 믿는다.

1) 가정 (assumption)
   - 자료는 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 서로 대응하고 $d_i = y_i - x_i$가 계산된다.
   - d_i은 서로 독립이다.

2) 가설 (hypothesis)
   - 귀무가설: $M_d = 0$ (짝진 표본의 중앙값의 차이는 0이다)
   - 대립가설: $M_d \neq 0$ (양측 검정) $M_d > 0$혹은 $M_d < 0$ (단측 검정)

3) 검정 통계량 (test statistic)

   ① $d_i = y_i - x_i$의 부호를 정한다. 만약 $d_i = 0$이면 분석에서 제외한다. + 부호 개수와 – 부호 개수 중 작은 것은 검정 통계량 K로 한다.

   ② p-value를 계산: 단측검정 경우 $\Pr(K \le k \mid n, 0.5)$, 양측검정 경우 $2 \times \Pr(K \le k \mid n, 0.5)$

4) 결정 이론

   계산된 p-값과 유의수준 α보다 크면 귀무가설을 채택하고 작으면 기각한다.

॥ EXAMPLE॥ 10마리의 쥐를 임의 추출하여 혼자 있을 때와 다른 쥐와 함께 있을 때의 심장 박동을 측정한 것이다. 함께 있을 때 심장 박동이 혼자 있을 때보다 큰지 유의성을 검정하시오. (유의수준=0.05) W. Daniel 'Applied Nonparametric Statistics'

쥐	1	2	3	4	5	6	7	8	9	10
혼자	463	462	462	456	450	426	418	415	409	402
함께	523	494	461	535	476	454	448	408	470	437
부호	-	-	+	-	-	-	-	+	-	-

1) 귀무가설: 심장 박동의 차이가 없다. $M_d = 0$
2) 대립가설: 심장 박동의 차이가 있다. $M_d < 0$ (단측 검정)
3) 검정통계량: $K = 2$
4) 결론: p-값= $\Pr(K \le 2 \mid n = 10,\ p = 0.5) = 0.0547$ 이므로 귀무가설을 기각하지 못한다. 즉 심장 박동에는 차이가 없다.

📖 **HOMEWORK #8** 계속 [due 5월 10일] Solve the following problems.

EXERCISES

4.1 Van Duijn (E2) studied the effect of clonazepam on cobalt-induced focal seizures in alert cats. Before and after administering clonazepam, he recorded focal paroxysmal activity as mean seconds with spikes/10 seconds recording. Table 4.3 shows the results of this part of the experiment. Can we conclude on the basis of these data that clonazepam decreases focal spiking? Let $\alpha = 0.05$. What is the P value?

TABLE 4.3

Effect of clonazepam on focal spikes produced by cobalt

Cat	1	2	3	4	5	6	7	8	9	10
Before (X)	2.7	4.2	3.3	5.3	4.2	3.5	6.5	4.6	3.7	7.1
After (Y)	4.5	2.6	1.4	2.5	2.5	2.3	3.0	2.6	1.9	0.4

Source: H. Van Duijn, "Superiority of Clonazepam over Diazepam in Experimental Epilepsy," *Epilepsia*, 14 (1973), 195–202.

4.2 Shani et al. (E3) studied the effect of phenobarbital on liver functions in patients with Dubin–Johnson syndrome. Table 4.4 shows the total bilirubin in the sera of these patients before and after treatment with phenobarbital. Can we conclude on the basis of these data that phenobarbital reduces total bilirubin level? Let $\alpha = 0.05$. What is the P value?

TABLE 4.4

Total bilirubin, milligrams per 100 ml, in sera of patients with Dubin–Johnson syndrome, before and after treatment with phenobarbital

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13
Before (X)	4.0	3.2	3.8	1.8	3.0	5.3	5.7	3.0	2.7	2.9	2.8	1.8	2.6
After (Y)	3.1	3.0	3.5	1.0	1.8	3.9	2.2	2.1	1.4	2.9	2.6	1.4	2.5

Source: Mordechai Shani, Uri Seligsohn, and Judith Ben-Ezzer, "Effect of Phenobarbital on Liver Functions in Patients with Dubin–Johnson Syndrome," *Gastroenterology*, 67 (1974), 303–308; copyright 1974, Williams & Wilkins, Baltimore.

4.2. Wilcoxon Matched pairs Singed Ranks Test

짝진 표본의 차이를 검정하는 방법으로 대응하는 관측치의 차이 부호를 이용하여 Sign test 방법을 사용할 수 있듯이 차이의 크기(순위)까지 고려한다면 WSR 검정을 사용할 수 있다. 방법은 2장의 WRS 방법과 동일하다.

1) 가정 (assumption)
  - 자료는 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 서로 대응하고 $d_i = y_i - x_i$가 계산된다.
  - d_i은 서로 독립이고 분포는 좌우 대칭이다.

2) 가설 (hypothesis)
  - 귀무가설: $M_d = 0$ (짝진 표본의 차이의 중앙값은 0이다)
  - 대립가설: $M_d \neq 0$ (양측 검정) $M_d > 0$혹은 $M_d < 0$ (단측 검정)

3) 검정 통계량 (test statistic)
  ① $d_i = y_i - x_i$을 계산한다. 물론 $d_i = 0$이면 그 자료는 제외한다.
  ② 절대값의 크기에 따라 $|d_i| = |y_i - x_i|$ 정렬하여 순위를 구한다.
  ③ $d_i = y_i - x_i$의 부호에 따라 순위에 부호를 붙이고 + 부호인 순위 합(T_+)과 – 부호인 순위 합(T_-)을 구한다.
  ④ 양측 검정이면 검정 통계량으로 $T = \min(T_+, T_-)$ 사용한다.
  ⑤ 단측 검정 $M_d > 0$이면 검정 통계량으로 $T = T_-$ 사용한다.
  ⑥ 단측 검정 $M_d < 0$이면 검정 통계량으로 $T = T_+$ 사용한다.

4) 결정 이론
  - Wilcoxon 표를 (WRS 표) 참조하여 가설을 검정한다. 표에는 표본의 크기 n에 따라 단측 검정의 경우 기각역 (critical region)과 거기에 대응하는 α값이 나타나 있다.
  - 그러므로 양측 검정인 경우는 α/2에 근사한 값을 찾아 가설 검정하면 된다.
  - 검정 통계량이 기각역에 포함되어 있으면 귀무가설을 기각하고 반대의 경우 귀무가설을 채택한다.

II EXAMPLE II 폐병 치료제 효과를 알아보기 위하여 9명의 환자를 대상으로 치료제 복용 전과 후의 동맥 압력을 측정하였다. 동맥 압력이 떨어졌는지 검정하시오 (유의수준 =0.05) W. Daniel 'Applied Nonparametric Statistics'

사람	1	2	3	4	5	6	7	8	9
복용 전	33	17	30	25	36	25	31	20	18
복용 후	21	17	22	13	33	20	19	13	9

1) 귀무가설: 동맥 압력 차이가 없다. $M_d = 0$

2) 대립가설: 치료 후 동맥 압력이 떨어졌다. $M_d < 0$ (단측 검정)

3) 검정통계량:

Calculation of test statistic for Example 4.2

Before therapy (X)	After 24 hours (Y)	$D_i = Y_i - X_i$	Signed rank of $\lvert D_i \rvert$
33	21	−12	−7
17	17	0	omit
30	22	−8	−4
25	13	−12	−7
36	33	−3	−1
25	20	−5	−2
31	19	−12	−7
20	13	−7	−3
18	9	−9	−5
			$T_+ = 0$

4) 결론: n=8이고 $T_+ = 0$ 이므로 WRS 표에 의하면 p-값은 0.0039이다. (단측 검정) 그러므로 유의 수준 0.05하에서 귀무가설은 기각되고 치료제가 동맥 압력을 낮춘다고 말할 수 있다.

HOMEWORK #9 [due 5월 15일]

4.5 Bhatia et al. (E7) reported the data shown in Table 4.9. Can we conclude on the basis of these data that treatment lowers the stroke index in patients of this type? Let $\alpha = 0.05$.

Stroke index (ml/beat/m²) in pre- and post-treatment studies of coronary circulation in chronic severe anaemia ***TABLE 4.9***

Case	1	2	3	4	5	6	7	8
Before treatment (*X*)	109	57	53	57	68	72	51	65
After treatment (*Y*)	56	44	55	40	62	46	49	41

Source: M. L. Bhatia, S. C. Manchanda, and Sujoy B. Roy, "Coronary Haemodynamic Studies in Chronic Severe Anaemia," *Br. Heart. J.*, 31 (1969), 365–374. Reprinted by permission of the authors and the editor.

4.6 Hall et al. (E8) reported the Lintner Soluble Starch (LSS) to Amylase Azure (AA) ratios shown in Table 4.10. Do these data provide sufficient evidence to indicate that saliva and duodenal fluid LSS/AA ratios are, on the average, different? Let $\alpha = 0.05$. Find the P value for this test.

TABLE 4.10
LSS/AA ratios for saliva and duodenal fluid in 11 patients

Subject	A	B	C	D	E	F	G	H	I	J	K
Saliva (*X*)	0.78	0.86	0.82	1.10	0.71	1.00	0.65	0.85	0.64	0.86	0.70
Duodenal fluid (*Y*)	0.46	0.37	0.41	0.46	0.45	0.67	0.47	0.37	0.25	0.33	0.42

4.3. 신뢰 구간

신뢰 구간을 구하는 방법도 동일하다. 다른 점이 있다면 우선 $d_i = (x_i - y_i)$ 을 구한 후 이을 가지고 순서 통계량을 구하거나 쌍들의 평균을 구하는 것만 다를 뿐 다른 것은 동일하다. Sign test 방법은 2.2.2절, WRS 방법은 2.3.2절을 참조하여 다음 숙제를 함으로써 기억을 되살리기 바란다.

📖 **HOMEWORK #9** 계속 [due 5월 15일] 동맥 압력을 떨어뜨리는 치료제의 효과가 있는지 알아보기 위하여 11명을 임의 추출하여 다음 자료를 얻었다. 동맥 압력의 차이에 대한 95% 신뢰구간을 Sign 검정과 WRS 방법을 이용하여 구하시오.

Mean arterial blood pressure in 11 anesthetized subjects before and after receiving oxytocin

Subject	1	2	3	4	5	6	7	8	9	10	11
Before	95	173	94	97	81	100	97	104	72	101	83
After	55	90	38	59	46	48	49	92	23	55	49

Source: F. Robert Weis, Jr., and Jerome Peak, "Effects of Oxytocin on Blood Pressure during Anesthesia," *Anesthesiology*, 40 (1974), 189–190.

4.4. 교차표의 수준이 짝을 이룬 경우 (McNemar 검정)

수준이 짝을 이루었다는 것은 무슨 뜻인가? 예를 들어 보자. 새로운 이슈가 발생한 경우 A 대통령 후보에 대한 지지 여부가 바뀌었는지 알아보거나 (물론 이 경우 동일 응답자, 즉 패널(panel) 구성), 안전 벨트 착용 거부여부가 교육 전후에 바뀌었는지 알아보고자 할 때 사용되는 방법이다. 전후 사이에 응답 대상이나 실험 대상이 바뀌는 경우나, 전현 다른 것을 측정하는 경우(즉 수준이 달라지는 경우)는 McNemar 방법을 사용할 수 없다.

4.4.1. 실험 상황

다음 교차표는 McNemar 검정 방법을 사용할 경우

		After		
		Yes	No	Total
Before	Yes	**A**	**B**	A+B
	No	**C**	**D**	C+D
	Total	A+C	B+D	N

N=응답자 총수
A=실험 전후 모두 YES인 응답자
D=실험 전후 모두 NO인 응답자

B=실험 전 YES, 실험 후 NO
D=실험 전 No, 실험 후 YES

4.4.2. McNemar 방법

반드시 동일 실험 대상이 전후에 사용되어야 하고 같은 개념을 묻거나 실험해야 한다. McNemar는 이 방법을 수준이 2개(Yes, No)인 경우만 제안했으나 Bennett & Underwood가 3개 이상인 경우로 확대하였다.

편의를 위하여 수준이 2개인 경우를 가설 검정 순서를 살펴보기로 하자.

1) 가설 (hypothesis)

- 귀무가설: $p_1 = p_2$ (실험 전의 yes 비율과 실험 후의 yes 비율이 같다)
- 대립가설: $p_1 \neq p_2$ (양측 검정) $p_1 > p_2$ 혹은 $p_1 < p_2$ (단측 검정)

2) 검정 통계량 (test statistic)

표본 추정치: $\hat{p}_1 = \dfrac{A+B}{N}$, $\hat{p}_2 = \dfrac{A+C}{N}$ 표본 추정치 차이: $\hat{p}_1 - \hat{p}_2 = \dfrac{B-C}{N}$

귀무가설이 맞다면 $(B-C)/N = 0$ 이므로 이를 이용하여 McNemar는 검정 통계량으로 다음을 제안하였고 이가 성립하기 위해서는 (B+C)가 적어도 10이상이어야 한다.

$$z = \frac{B-C}{\sqrt{B+C}} \sim Normal(0,1)$$

‖ EXAMPLE‖ 안전벨트 교육 효과를 알아보기 위하여 85명을 임의로 선택하여 교육 전 벨트 작용 여부와 교육 후 벨트 작용 여부를 조사하여 아래 표를 만들었다. 교육 효과가 있는지 검정하시오 (유의수준=0.05)

		교육 후		
		Yes	No	Total
교육 전	Yes	**7**	**37**	41
	No	**26**	**15**	44
	Total	33	52	85

1) 가설 (hypothesis)

- 귀무가설: $p_1 = p_2$ (교육 전후 벨트 착용 비율의 같다)
- 대립가설: $p_1 < p_2$ (교육 후 벨트 착용 비율이 높아졌다)

2) 검정 통계량 (test statistic): $z = \dfrac{37-26}{\sqrt{37+26}} = 1.385$ 그러므로 p-value= $pr(z \geq 1.38) = 0.084$

이므로 귀무가설을 기각하지 못한다.

4.2.3. SAS 이용하기 (McNemar 검정)

Editor - Untitled1 *

```
data mcnemar;
    input pre $ post $ f @@;
    cards;
    y y 7 y n 37 n y 26 n n 15
run;
proc freq data=mcnemar order=data;
    weight f;
    table pre*post/nopercent norow;
    exact mcnem;
run;
```

- order=data 옵션은 자료 입력 순서대로 수준이 출력된다.
- nopercent 옵션은 각 셀의 비율을 출력하지 않는다.
- norow는 셀의 향 비율을 출력하지 않는다. McNemar 검정에는 이 비율이 필요하지 않기 때문이다.

```
Table of pre by post

pre       post

Frequency-
Col Pct  -y        -n        -  Total
--------- -------- --------
y        -       7 -      37 -     44
         -   21.21 -   71.15 -
--------- -------- --------
n        -      26 -      15 -     41
         -   78.79 -   28.85 -
--------- -------- --------
Total           33        52       85

Statistics for Table of pre by post

      McNemar's Test
----------------------------
Statistic (S)         1.9206
DF                         1
Asymptotic Pr >  S    0.1658
Exact      Pr >= S    0.2074
```

- SAS가 사용하는 검정통계량은 정규분포 $z=\dfrac{B-C}{\sqrt{B+C}}$ 이 아니라 $c^2=\dfrac{(B-C)^2}{B+C}$ Chi-square 통계량 값이다. 같은 결론에 도달하는가? 그렇다. 이유는?

📖 **HOMEWORK #9** [due 5월 15일] McNemar 검정이 가능한 자료를 수집하고 분석하시오..

5. 독립성, 동질성 검정

두 변수가 서론 연관(associate)을 맺고 있는지 관심을 갖게 될 때 변수들간의 독립성을 알아본다고 한다. 예를 들면 학생들의 영양 상태와 성적간의 관계가 있는지, 취득 학위와 소득 수준간의 관계가 있는지 알아보는 것을 변수간의 독립성(independence) 검정이다.

자료(변수)들의 분포가 동일(likeness)한지를 살펴보는 것을 동질성(homogeneity) 검정이라 한다. 예를 들어 보자. 독립인 두 모 비율을 검정하는 경우(남자의 성 경험 비율과 여자의 성 경험 비율의 차이가 있는가) 귀무가설은 $H_0 : p_1 = p_2$ 설정하여 z-검정이나 t-검정을 시행할 수 있다. 만약 이 귀무가설을 "두 모집단이 동질하다" 한다면 동질성 검정을 통해 비율의 차이가 있는 검정하게 된다.

독립성 검정이나 동질성 검정은 c^2-검정(Chi-Square)을 실시한다. 또한 각 변수들은 분류형이므로 두 변수 간의 관계를 알아보기 위하여 빈도표 2개를 겹치게 되는 이를 교차표(cross-tabulation) 혹은 분할표(contingency table)라 부르기도 한다.

5.1. Chi-Square 분포

정규 분포(평균= m , 표준편차= s)를 따르는 변수 X를 다음과 같이 표준화하면 Z는 표준정규분포(평균=0, 표준편차=1)를 따른다.

$$Z = \frac{X - m}{s} \sim Normal(0,1)$$

표준 정규분포를 따르는 확률변수 Z의 제곱은 자유도 1인 c^2분포를 따른다.

$$Z^2 = (\frac{X - m}{s})^2 \sim c^2(df = 1)$$

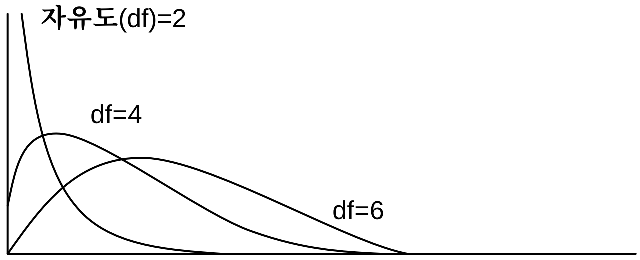
⏭ 알아두기

- 자유도 k 인 Chi-Square 분포 함수는 다음과 같다.

$$f(x) = \frac{1}{\Gamma(k/2)2^{k/2}} x^{(k/2)-1} e^{-x/2}, \quad 0 \le x < \infty$$

- 자유도 k 인 Chi-Square의 평균은 k 이고 분산은 $2k$ 이다.
- 각각 $c^2(df = 1)$ 분포를 따르고 서로 독립인 변수 $W_1, W_2, \ldots, W_k$ 의 합은 자유도가 k인 Chi-square 분포를 따른다. $Y = \sum W_i \sim c(df = k)$
- $W \sim c^2(df = m1)$ 이고 $V \sim c^2(df = m2)$ 분포를 따르고 서로 독립인 경우 두 변수의 비율은 F-분포를 따른다. $H = \frac{W}{m1} / \frac{V}{m2} \sim F(m1, m2)$
- 자유도 $k = 2$ Chi-Square 분포는 지수분포(exponential dist.: $b = 2$) 따른다.

- 자유도에 따른 Chi-Square 확률 밀도 함수를 그리면 다음과 같다.



5.2. Contingency Table

독립성과 동질성 검정은 분류형 변수인 두 변수간의 관계를 살펴보는 것으로 자료가 가진 정보를 분할표로부터 얻어지고 관련 검정통계량도 계산된다.

변수1의 위치에 설명 변수, 변수 2의 위치에 종속 변수(반응 변수)를 두는 것이 좋다. 예를 들어 영양 상태와 학점의 관계를 보는 경우 영양 상태가 설명 변수의 역할을 하게 되므로 변수1의 위치에 영양 상태를 두는 것이 좋다. 변수1의 수준이 r이고, 변수2의 수준이 c인 경우 분할표는 다음과 같다.

변수 1	변수 2				
	1	2	…	c	total
1	O_{11}	O_{12}	…	O_{1c}	$n_{1.}$
2	O_{21}	O_{22}	…	O_{2c}	$n_{2.}$
:	:	:	:	:	:
r	O_{r1}	O_{r2}	…	O_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$		$n_{.c}$	$n_{..}$

독립성이나 동질성 Chi-Square 검정의 기본 개념은???

만약 독립이라면 (귀무가설이 성립한다면) $\Pr(AB) = pr(A)\Pr(B)$

그러므로 귀무가설 하에서 기대 빈도 $E_{ij} = (\frac{n_{i.}}{n_{..}})(\frac{n_{.j}}{n_{..}})n_{..}$ 이다.

검정 통계량 $T = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ 의 의미는???

5.3. 독립성 Chi-Square 검정

분류형 변수들간의 관계(associate)를 검정한다. A학부 1학생 230명을 대상으로 남녀별 전공 선택(3개 전공)의 차이가 있는지 알아보고자 하여 자료를 조사하여 다음 교차표를 얻었다고 하자.

성별	전공			
	A 전공	B전공	C전공	Total
남자	75	46	23	144
여자	30	32	24	86
Total	105	78	47	230

가정

관심의 모집단으로부터 n명(개)을 추출하여 그 표본에서 2개의 변수를 측정하는데 이는 모두 분류형 변수이다. 교차표를 작성하는 경우 설명 변수에 해당하는 변수가 존재하면 행으로 둔다. 위의 예제에서는 성별이 설명 변수가 된다. 성별에 따라...

가설

귀무가설: 두 분류형 변수간의 관계가 없다. ←→ 두 변수는 서로 독립이다.

- 남녀 변수와 전공 선택 변수간에의 관계가 없다.
- 남녀별 전공 선택 변수의 차이는 없다.

대립가설: 관계가 없다. 서로 독립이 아니다.

- 남녀별 전공 선택의 차이는 있다

검정통계량

만약 두 변수가 서로 독립이라면

$$\text{Pr(표본 n명 중 ij 셀에 속하는 빈도)} = E_{ij} = \left(\frac{n_{i.}}{n_{..}}\right)\left(\frac{n_{.j}}{n_{..}}\right)n_{..}$$

위의 예에서 두 변수가 독립이라면 Pr(남자∩A전공) = Pr(남자) Pr(A전공)이다. 그러므로

- 1행 1열의 기대 빈도는 $E_{11} = \left(\frac{n_{1.}}{n_{..}}\right)\left(\frac{n_{.2}}{n_{..}}\right)n_{..} = 105\times144/230 = 65.7$
- 다른 셀도 같은 방법... 2행 3열 기대 빈도 $E_{23} = 86\times47/230 = 17.6$

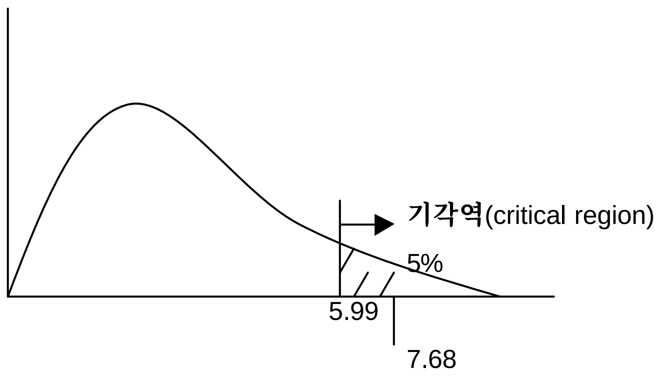
만약 귀무가설이 성립하면 (독립이라면) 각 셀의 $(O_{ij}-E_{ij})$**는 0에 가까운 값일 것이다.**

검정통계량 $T=\sum\sum\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$**는 자유도** $(r-1)(c-1)$ c^2**-분포에 근사(approximate) 한다.**

- $T=\frac{(75-65.7)^2}{65.7}+...+\frac{(24-17.6)^2}{17.6}=7.68\sim c^2(df=(2-1)(3-1)=2)$

결론

계산된 검정통계량 값이 c^2**-분포표로부터 구한 기각역에 속하면 귀무가설을 기각하고 그렇지 않으면 귀무가설을 채택한다.**



계산된 검정 통계량이 7.68이므로 기각역에 $\{>5.99\}$ **속하므로 귀무가설이 기각되고 성별과 전공 선택에는 관계가 있다고 할 수 있다. 그러면 어떤 관계가 있는가? 이것에 대한 대답은 행 퍼센트를 참고하면 된다. 설명 변수가 행에 있으므로... SAS 출력을 미리 살펴보면 남자는 A 전공을, 여자는 B 전공을 선호한다고 말할 수 있다. 여자는 각 전공을 골고루 택하지만 남자는 A 전공을 선호하고 있다.**

대립 가설이 양측 검정의 형태인데 기각역이 왜 한 쪽 방향만 고려되지라고 묻고 싶은 사람이 있다면 묻기 전에 생각해 보라. 검정 통계량의 값이 어떻게 계산되었는지를... 그래도 이해가 되지 않으면 머리를 벽에다 세 번 박으면 알게 될 것이다.

SAS 이용하기

```
a.sas *
data one;
    input gender $ major $ w @@;
    cards;
    m a 75 f a 30 m b 46 f b 32 m c 23 f c 24
run;

proc freq data=one;
    weight w;
    table gender*major/nocol nopercent chisq expected;
run;
```

- NOCOL: 열 퍼센트를 출력하지 않는 옵션.
- NOPERCENT: 각 셀의 퍼센트 출력을 하지 않게 하는 옵션.
- CHISQ: 변수들간의 관계를 검정하는 검정통계량을 출력한다.
- EXPECTED: 기대 빈도를 출력한다.

```
gender     major

Frequency-
Expected -
Row Pct  -a        -b        -c        -  Total
--------- --------- --------- ---------
f        -     30 -     32 -     24 -     86
         - 39.261 - 29.165 - 17.574 -
         -  34.88 -  37.21 -  27.91 -
--------- --------- --------- ---------
m        -     75 -     46 -     23 -    144
         - 65.739 - 48.835 - 29.426 -
         -  52.08 -  31.94 -  15.97 -
--------- --------- --------- ---------
Total         105        78        47     230

Statistics for Table of gender by major

Statistic                    DF       Value      Prob
------------------------------------------------------
Chi-Square                    2      7.6823    0.0215
Likelihood Ratio Chi-Square   2      7.6869    0.0214
Mantel-Haenszel Chi-Square    1      7.6186    0.0058
Phi Coefficient                      0.1828
Contingency Coefficient              0.1798
Cramer's V                           0.1828
```

- Chi-Square = 독립성 검정을 위한 c^2 검정 통계량
- Likelihood Ratio = $2\sum\sum O_{ij}\ln(O_{ij}/E_{ij}) \sim c^2(df=(r-1)(c-1))$
- M-H $Q_{MH}=(n-1)r^2 \sim c^2(df=1)$ r은 두 변수간의 Person 상관 계수다. M-H 검정 통계량은 두 분류형 변수간의 선형 상관 관계를 검정한다. 그러므로 두 변수 모두 순서형(ordinal)이어야 한다.
- Phi, Contingency, Cramer's coefficient는 변수의 상관계수 값.

‖ EXAMPLE ‖

소득 수준과 지난 병원 방문 후 다음 병원 방문 기간의 관계가 있는지 알아보기 위하여 2764명을 대상으로 자료를 조사하여 다음 분할표를 작성하였다. 관계가 있는지 유의수준 5% 하에서 검정하시오.

Income	Within previous 6 months	Within 7 months to one year	More than one year	Total
Less than $3000	186	38	35	259
$3000–$4999	227	54	45	326
$5000–$6999	219	78	78	375
$7000–$9999	355	112	140	607
$10,000 plus	653	285	259	1197
Total	1640	567	557	2764

1) 귀무가설: 소득 수준과 병원 찾는 기간은 관계가 없다. (소득 수준에 따라 병원을 찾는 기간은 차이가 없다)
2) 대립가설: 차이가 있다.
3) 기대빈도: $E_{11} = 256 \times 1640 / 2764 = 153.68 \ldots$ 을 계산한다.

Income	Within previous 6 months	Within 7 months to one year	More than one year	Total
Less than $3000	186 (153.68)	38 (53.13)	35 (52.19)	259
$3000–$4999	227 (193.43)	54 (66.87)	45 (65.70)	326
$5000–$6999	219 (222.50)	78 (76.93)	78 (75.57)	375
$7000–$9999	355 (360.16)	112 (124.52)	140 (122.32)	607
$10,000 plus	653 (710.23)	285 (245.56)	259 (241.22)	1197
Total	1640	567	557	2764

4) 검정통계량: $T = \dfrac{(186-153.68)^2}{153.68} + \dfrac{(227-193.43)^2}{193.43} + \ldots + \dfrac{(259-241.22)^2}{241.22} = 47.9$, 자유도가 8이므로 c^2-분포표에 의하면 기각역은 $\{>15.507\}$ 이다. 귀무가설을 기각한다. 즉 소득 수준에 따라 소득 수준이 높을수록 찾는 기간이 다소 길어짐을 알 수 있다.

‖ HOMEWORK #10 ‖ (due May 17)

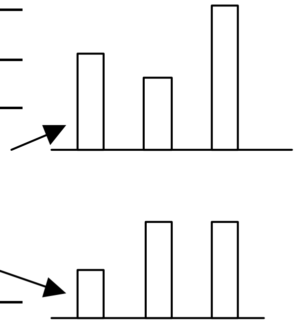
1) 위의 예제를 SAS를 이용하여 가설 검정하시오. (유의수준=0.1)
2) 다음 자료를 이용하여 안경 착용 나이가 성별에 따라 차이가 있는지 검정하시오. 수작업으로 계산하고 SAS로도 계산하시오.

성별	안경 착용 나이			
	15세 미만	15-19	20세 이상	합계
남자	2	38	22	62
여자	8	93	15	116
합계	10	131	37	178

5.4. 동질성 검정을 위한 Chi-Square 검정

분류형 변수의 동질성이란 한 변수의 각 수준의 빈도(히스토그램, 확률 밀도 함수)가 동일한 분포의 형태를 갖고 있는지 검정하는 방법이다. 이것을 무엇을 의미하는가?

변수1	변수2				
	1	2	…	c	합계
1	O_{11}	O_{12}	…	O_{1c}	$n_{1.}$
2	O_{21}	O_{22}	…	O_{2c}	$n_{2.}$
:	:	:	:	:	:
r	O_{r1}	O_{r2}	…	O_{rc}	$n_{r.}$
합계	$n_{.1}$	$n_{.2}$		$n_{.c}$	$n_{..}$

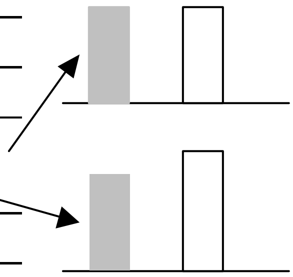


각 수준의 히스토그램들이 같은 형태를 갖는지를 검정하는 것이 동질성 검정이다. 예를 들어 단과 대학별로 애인 여부가 동일한 분포를 갖는가 알아보고자 할 때 동질성 검정을 한다.

실제 응용에서는 독립성 검정과 동질성 검정의 결과 해석에는 큰 차이가 없다. 위의 동질성 검정 결과 학년별로 동질하지 않다는 것은 학년별 애인 여부의 차이가 있다는 것을 의미한다.

어떻게 동질성 검정을 하는지 2x2 분할표(이를 fourfold 분할표라 부르기도 한다) 다음 예를 생각해 보자.

성별	결혼 여부		
	예	아니오	Total
남자	A	B	A+B
여자	C	D	C+D
Total	A+C	B+D	N



동질성 검정이란 성별 변수 수준의 각각에서 (남자 결혼 비율)=(여자 결혼 비율), (남자 미혼 비율)=(여자 미혼 비율)을 검정하는 것이다. 즉 남자 집단과 여자 집단이 결혼 여부면에서 동질인가를 검정하는 것이다. 2x2 분할표의 경우 동질성 검정 통계량을 다음과 같이 계산한다.

$$\text{검정 통계량 } T = \frac{N(AD-BC)^2}{(A+C)(B+D)(C+D)(A+B)} \sim c^{2}(df=1)$$

▸ 참고

위의 교차 분석은 독립인 두 모집단 비율의 차이 검정하는 것과 동일하므로 귀무가설: 남자의 폐암 발생 비율(p_1)과 여자의 폐암 발생 비율이 (p_2) 동일한가를 검정하는 모수적 방법을 사용할 수 있다.

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} \xrightarrow{appro} z(0,1)$$

$$\hat{p}_1 = \frac{B}{A+B}, \quad \hat{p}_2 = \frac{D}{C+D}, \quad \text{그리고} \quad \hat{p} = \frac{B+D}{N}$$

Ⅱ EXAMPLE Ⅱ

AIDS 감염 여부면에서 볼 때 A 지역 거주자와 B 거주자는 동질인가 검정하시오. (유의 수준=0.05)

지역별	AIDS 감염 여부		
	yes	no	합계
A지역	20	6	26
B지역	16	14	30
합계	36	20	56

검정 통계량 $T = \frac{56(20\times14-16\times6)^2}{(20+6)(16+14)(20+14)(6+16)} = 3.38 \sim c^2(df=1)$ $c^2(df=1)$ 분포표의 기각역 $\{> 3.84\}$에 속하지 않으므로 귀무가설을 기각하지 못한다. AIDS 감염 여부면에서 두 지역 집단은 차이가 없다고 할 수 있다.

Ⅱ HOMEWORK #11 Ⅱ (due May 22)

학점 수준면에서 남자 집단과 여자 집단이 동질인지 검정하시오. Fourfold Chi-square 수작업(hand calculation) 방법으로 알아보고 모수적 방법으로 검정하시오. (유의수준=0.05)

성별	학점 수준		
	3.0 미만	3.0 이상	합계
남자	31	19	50
여자	17	33	50
합계	48	52	100

RXC 분할표인 경우 동질성 검정

각 수준이 동질이라면(귀무가설 하에서) 셀의 기대도수는 다음과 같다.

$$E_{ij}=(\frac{n_{.j}}{n_{..}})\times n_{i.}=(\text{전체 열 퍼센트})\text{x}(\text{행 총 빈도})$$

위의 기개 빈도는 독립성 검정 시 사용되는 기대 빈도와 동일함을 알 수 있다. 그러므로 검정 통계량은 독립성 검정과 동일하다. 차이가 있다면 기대 빈도 계산 규칙인데 이는 귀무가설이 다르기 때문이다. 그러나 검정 통계량의 계산 값은 두 방법 모두 동일함을 알 수 있다.

॥ HOMEWORK #11 ॥ (계속 due May 22)

결혼 여부면에서 각 전공별 학생들이 동질인지 검정하시오. 수 작업과 SAS 이용하여 검정하시오. (유의수준=0.05)

전공	결혼 여부		
	미혼	결혼	합계
A 전공	50	20	70
B 전공	12	25	37
C 전공	6	8	14
D 전공	1	21	22
합계	69	74	143

5.5. 기대 빈도가 적은 셀 (Fisher exact test)

독립성이나 동질성 검정하는 검정 통계량은 근사적으로 Chi-square 분포에 따른다. 근사 조건으로는 각 셀의 기대 빈도가 5이상 (어떤 이는 10이상)이어야 한다. Cochran 은 자유도 2 이상인 경우 기대 빈도 5 이상인 셀이 전체 20%만 넘으면 분할표에서 구한 검정 통계량은 Chi-square 분포에 근사한다고 했다. 이처럼 근사 조건은 일치하지는 않지만 일반적으로 Cochran의 이론을 받아들인다.

SAS 출력 결과의 경우 기대 빈도가 5이하인 셀이 존재하는 경우 몇 개 인지 출력하게 된다. 출력의 의미는 총 6개의 셀 중 기대 빈도가 5이하인 셀이 33%가 있다는 것이다.

```
Chi-Square                      2      4.9267    0.0851
Likelihood Ratio Chi-Square     2      5.1193    0.0773
Mantel-Haenszel Chi-Square      1      3.1207    0.0773
Phi Coefficient                        0.2993
Contingency Coefficient                0.2867
Cramer's V                             0.2993

WARNING: 33% of the cells have expected counts less
         than 5. Chi-Square may not be a valid test.
```

이런 경우 계산된 검정 통계량은 Chi-Square 분포에 근사하지 않는다. 이런 경우 해결책은 무엇인가?

독립성 검정의 경우 변수의 수준을 합쳐 셀의 수를 줄이는 방법이다. 위의 예에서 변수 X의 수준 중 0과 1을 합쳐 하나의 수준으로 하면 이 문제는 해결된다. 셀을 합칠 경우 그룹으로 할 수 있는 것을 합친다. 예를 들어 수준이 (상, 중, 하)인 경우 '상'과 '하'를 합치는 것은 정말 어리석은 일이다. 수준의 의미가 상실되기 때문이다. 다음은 X 변수의 0과 1을 1 수준으로 합친 후 교차 분석을 실시한 출력 결과이다.

```
x         y

Frequency-
Row Pct  -       0-       1-  Total
--------- -------- --------
       1 -      17 -     13 -     30
         -   56.67 -  43.33 -
--------- -------- --------
       2 -      21 -      4 -     25
         -   84.00 -  16.00 -
--------- -------- --------
Total           38       17       55
Statistics for Table of x by y

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      4.7706    0.0289
Likelihood Ratio Chi-Square    1      4.9834    0.0256
Continuity Adj. Chi-Square     1      3.5766    0.0586
Mantel-Haenszel Chi-Square     1      4.6839    0.0304
Phi Coefficient                      -0.2945
Contingency Coefficient               0.2825
Cramer's V                           -0.2945
```

동질성 검정의 경우 Exact test를 시행하는 것이다. 물론 이 방법은 독립성 검정에도 적용될 수 있다. 이는 근사 통계량을 이용하는 것이 아니다. 처음 이 방법을 제안한 사람은 Fisher인데 그는 2x2 분할표의 경우 제안하였고 후에 RxC 분할표로 확대되었다. Fisher's Exact test (2x2 분할표) 방법을 예제 중심으로 살펴보기로 하자. 다음은 성별 변수와 애인 유무 변수의 관계(애인 유무면에서 남자와 여자는 동질이다) 알아보기 위하여 자료를 조사하여 다음 분할표가 작성되었다고 하자. 귀무가설: 남자 여자 집단의 분포는 같다.

성별	애인 유무		
	있음	없음	합계
남자	5	2	7
여자	3	3	6
합계	8	5	13

이 분할표보다 극단의 분할표가 나오는 확률을 구하여 합하면 설정한 귀무가설 검정을 위한 p-값이 된다.

성별	애인 유무		
	있음	없음	합계
남자	6	1	7
여자	2	4	6
합계	8	5	13

성별	애인 유무		
	있음	없음	합계
남자	7	0	7
여자	1	5	6
합계	8	5	13

위의 3개 분할표의 나올 확률을 모두 더하면 p-값이 된다.

N_{11}	N_{12}	$N_{1.}$
N_{21}	N_{22}	$N_{2.}$
$N_{.1}$	$N_{.2}$	$N_{..}$

위의 분할표가 나올 확률은? $p = \dfrac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n_{11}!n_{12}!n_{21}!n_{22}!n_{..}!}$ 이다. 그러므로 위의 세 분할표의 확률을 각각 구하면 다음과 같다.

$$p = \frac{7!6!8!5!}{5!2!3!3!13!} = 0.3263 \ , \quad p = \frac{7!6!8!5!}{6!1!2!4!13!} = 0.0816 \ , \quad p = \frac{7!6!8!5!}{7!0!1!5!13!} = 0.047$$

그러므로 p-값은 0.4126이다. 즉 두 변수간의 관계가 존재하지 않는다(혹은 남자집단과 여자 집단은 동질이다)라고 결론지을 수 있다.
위에서 계산된 p-값은 단측 검정의 경우 p-값에 해당한다. 검정 통계량의 분포가 좌우 대칭이 아니므로 양측검정의 경우 p-값 계산은 다소 복잡하다. 그냥 단측 검정을 실시하거나 SAS 출력 결과를 이용하는 것이 편리하다.

SAS 이용하기

```
a.sas *
data one;
     input gender $ major $ w @@;
     cards;
     m y 5 m n 2 f y 3 f n 3
run;

proc freq data=one;
     weight w;
     exact fisher;
     table gender*major/nocol nopercent chisq;
run;
```

```
gender     major

Frequency-
Row Pct  -n        -y        -  Total
--------- -------- --------
f        -      3 -      3 -      6
         -  50.00 -  50.00 -
--------- -------- --------
m        -      2 -      5 -      7
         -  28.57 -  71.43 -
--------- -------- --------
Total           5        8       13
Statistics for Table of gender by major
생략
WARNING: 100% of the cells have expected counts less
         than 5. Chi-Square may not be a valid test.

       Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)         3
Left-sided Pr <= F          0.9138
Right-sided Pr >= F         0.4126

Table Probability (P)       0.3263
Two-sided Pr <= P           0.5921
```

- left sided는 1행 1열이 5, 4, 3, 2, 1, 0가 되는 분할표의 확률을 더한 것이다.

॥ HOMEWORK #11 ॥ (due May 22)

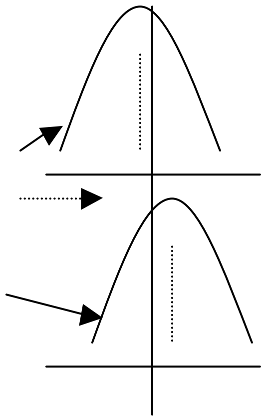
면접 방법에 따른 혈압 상승 정도의 차이를 보기 위해 자료를 수집하였다. Fisher의 Exact test를 시행하시오. (유의수준=0.05) 수작업 계산, SAS 방법 모두 시행하시오.

면접 방법	혈압 상승		
	높음	낮음	합계
그룹	1	5	6
개인	6	0	6
합계	7	5	12

6. One-way layout

서로 독립인 모집단이 3개 이상인 경우 중앙 위치를 비교 분석은 이는 일원 배치(one-way layout) 실험 계획에 대한 분석 방법과 같다. 요인의 수준이 k인 일원 배치 실험과 k 개의 모집단 중앙 위치에 대한 비교를 위한 자료 수집을 비교하면 다음과 같다.

	일원 배치 설계				모집단 중앙
요인	1	2	…	반복 수	위치 비교
1	y_{11}	y_{12}	$...y_{1n1}$	n_1	y_{11} y_{12} … y_{1n1}
2	y_{21}	y_{22}	$...y_{2n2}$	n_1	y_{21} y_{22} … y_{2n2}
…	…	…		…	…
k	y_{k1}	y_{k2}	$...y_{knk}$	n_k	y_{k1} y_{k2} … y_{1nk}



총 평균 (grand mean)

모수적 방법의 경우 표본 평균을 이용하여 모집단의 평균의 차이를 검정한다. 총 평균과 각 모집단의 평균의 차이가 크면 평균의 차이가 있을 것이다. 분산 분석은 자료의 총 변동(SST)을 처리 변동과 (SStr: 총 평균과 각 모집단 평균차이 제곱합) 오차 변동(SSE= SST-SStr)으로 나누고 처리 변동이 오차 변동에 비해 많이 크면 모집단 평균의 차이가 있다, 즉 처리 효과의 차이가 있다고 결론 내린다. 이 분석 방법이 분산 분석이고 가정은 모집단이 2개인 경우처럼 (페이지 39 참조) 각 모집단의 분산은 등분산이어야 한다는 가정이 필요하다. 분산 분석 방법의 경우 반복 수는 크지 않으므로 대표본 성질을 이용할 수 없고 분산(변동)의 비(ratio)가 F-검정을 따른다는 것을 이용하여 분석하므로 가정을 하게 되는데 이것이 정규성 가정(normality assumption)이다.,

등분산 가정이나 정규성 가정이 만족하지 않는 경우 사용되는 비모수 검정 방법으로는 Median 검정 일반, Kruskal-Wallis 분석 방법이 있다. 정규성이나 등분산성 가정 성립에 확신이 없으면 비모수적 방법을 사용해야 한다. 실험 계획에서 분산 분석에서 반복 수의 크기가 크지 않으므로(표본의 크기) 정규성이나 등분산 가정을 검정하는 것을 현실적으로 불가능하고 실험 계획에서는 측정치 수집 절차를 분석자가 관리하므로 실험 계획에서의 분산분석은 일반적으로 두 가정이 만족하다고 가정한다.

6.1. 모수적 방법 (분산 분석: Analysis of Variance: ANOVA)

실험 설계

요인의 수준이 t개이고 각 수준의 반복 수를 n_t라고 한다면 다음과 같은 실험 설계 표를 만들고 각 셀에 1-n까지 번호를 차례로 매긴다. 난수표에 의해 실험 순서를 정하고 실험을 실험 단위(unit)에 실시하면 된다. 그러므로 일원 분산의 실험 설계는 각 셀에 번호를 부여한 후 난수를 이용하여 실험 순서를 결정하므로 이를 완전 임의 실험 설계(CRD: Completely Randomized Design)라 부르기도 한다.

비료 종류(A, B, C)에 따른 벼 수확량의 차이를 알아보기 위하여 실험을 한다고 하자. 반복 수는 각 비료에 대해 4번을 한다면 실험을 위해 총 12곳의 경작지가 필요하다.

①	②	③
④	⑤	⑥
⑦	⑧	⑨
⑩	⑪	⑫

12 경작지가 땅의 비옥도가 동일하다면 CRD 방법을 이용하여 실험하면 된다. 난수표를 이용하여 임의로 비료를 배정할 수도 있지만 총 실험 수가 많지 않으므로 12 장의 종이에 각 번호를 적고 던진 후 하나씩 선택하여 실험을 배정하면 된다. 만약 1, 5, 2, 7, 12, 10 …, 3 순으로 선택하였다면

① A	② A	③ C
④ B	⑤ A	⑥ B
⑦ A	⑧ C	⑨ C
⑩ B	⑪ C	⑫ B

만약 경작지 아래 개천이 흐른다면 물로부터 거리에 따라 땅의 비옥도의 차이는 있을 것이므로 CRD 방법은 적합하지 않다. 물의 위치에 따라 경작지를 나누고(블록화:block) 각 블록 내에서 각 비료를 하나씩 임의(randomized)로 배정하는 실험 계획을 실시하면 된다. 이를 Randomized Block Design이라 한다.

A	B	C
A	C	B
C	A	B
B	C	A

예제 자료

다음은 3개 호수의 산소량의 차이가 있는지 알아보기 위하여 각 호수의 중앙에서 깊이 1m 의 물로부터 산소량(ppm)을 측정한 자료이다. 호수에서 위치에 따라 산소량의 차이가 있을 것이므로 10곳을 선택하여 각 산소량을 측정한 것이다.

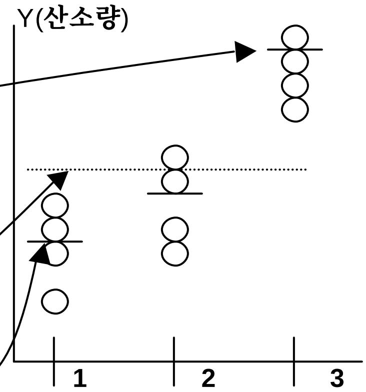
Lake	Observation
1	0 2 1 3 1 2 3 4 1 5
2	1 3 4 6 8 7 5 3 4 5
3	14 26 25 18 19 22 21 16 20 30

모형 $y_{ij} = u + t_i + e_{ij} = m_i + e_{ij} \quad i = 1,2,\dots k \quad j = 1,2,\dots,n_i$

추정치: $(y_{ij} - \bar{y}) = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})$

변동: $\sum\sum(y_{ij} - \bar{y})^2 = \sum\sum(y_{ij} - \bar{y}_i)^2 + \sum\sum(\bar{y}_i - \bar{y})^2$

$SST = SSE + SStr$



가정 $e_{ij} \sim Normal(0, s^2)$

⏭ **등분산 가정 검정**: **분산분석에서 등분산 가정은** Hartley's test**를 이용하면 된다.**

귀무가설: $H_0 : s_1^2 = s_2^2 = \dots = s_t^2$

검정통계량: $F_{\max} = \dfrac{\max(s_i^2)}{\min(s_i^2)} \sim F(\text{자유도 : 분자표본크기} - 1, \text{분모표본크기} - 1)$

⏭ **등분산**(Homoscadicity)dl **만족하지 않아 이분산**(Heteroscadicity)**이면 다음과 같이 자료 변환을 시행할 수 있다.**

$s^2 = ku$ ➔ $y^* = \sqrt{y}$

$s^2 = ku^2$ ➔ $y^* = \log(y)$

모형 유의성 검정 (분산 분석표)

Source	DF	SS	MS	
Treatment	t-1	$SStr = \sum\sum(\bar{y}_i - \bar{y})^2$	$MStr = SStr/(t-1)$	$F = \dfrac{MStr}{MSE}$
Error	n-t	$SSE = SST - SStr$	$MSE = SSE/(n-t)$	
Total	n-1	$SST = \sum\sum(y_{ij} - \bar{y})^2$		$\sim\ F(t-1, n-t)$

사후 검정 (Post-hoc test) 혹은 다중 비교 (multiple comparison)

분산 분석의 F-검정은 단지 귀무가설 $H_0: u_1 = u_2 = ... = u_t$ 즉 전체적인 차이를 검정하는 것이다. 그러므로 수준별 차이(pairwise: 예: $H_0: u_1 = u_3$)가 있는지 혹은 수준의 선형 결합 대비(contrast: 예: $H_0: u_1 = \frac{u_2 + u_3}{2}$)의 차이가 있는지 검정할 필요가 있는데 이를 사후 검정 혹은 다중 비교(대비 포함)라 한다. 사후 검정이므로 비록 F-검정 결과와 관계없이 (귀무가설을 채택하더라도) 시행하게 된다. 다중 비교에서는 여러 개의 가설을 동시에 검정하므로 유의수준을 조정해야 한다. 이를 조정된 실험 유의수준 (controlled experimental error rate)이라 하고 $1-(1-a)^c$ 이다. 여기서 c는 가설 수를 의미한다. pairwise 다중 비교의 경우 $c = t(t-1)/2$ 가 된다.

- Fisher's Least Significant Difference
  - pairwise (두 수준별 평균 비교) 검정에 사용
  - LSD를 구하고 평균의 차이가 그보다 크면 귀무가설을 기각한다.
- Tukey W procedure
  - studentized range distribution: $W = \max(y_i) - \min(y_i)$ 이용하여 $q = w/s$ 표 작성
  - 가장 보수적인 방법이다. ➔ 자연 과학에서 가장 많이 이용한다.
- Student-Newman-Keuls procedure
  - Tukey 방법의 변형한 것으로 표만 다르다. 즉 임계치(critical value)만 차이가 있고 방법은 Tukey와 동일하다.
- Duncan Multiple range test
  - Tukey 방법과 매우 유사하나 수준별 표본 평균을 크기 순으로 나열하여 차이가 가장 큰 것을 비교해 가면서 유의 수준을 $1-(1-a)^r$ 으로 조정해 가면서 검정한다. r은 검정 단계 순서이다.
  - 귀무가설을 기각할 확률이 매우 높아 자주 사용하지 않는다.

- Scheffe's S method
  - o 대비(contrast)까지 고려한 다중 비교 방법이다. ➔ 사회 과학 분야
- Dunnett's procedure
  - o 처리 효과의 수준 하나가 control (실험 집단)인 경우 (예: placebo 집단, 교육을 하지 않는 집단, 이전 약 투여 집단) 이 집단과 다른 집단들을 pairwise 비교할 경우 사용된다.

대비 (contrast)

처리 효과 수준별 선형 결합의 차이가 있는지 알아보기 위한 분석 방법이다.

- **대비 설정 방법**

o 호수 1이 동쪽에 있고 호수 2, 3이 서쪽에 있는 경우 동쪽 호수와 서쪽 호수간의 산소량의 차이가 있는가?

대비: $Q = u_1 - (u_2 + u_3)/2$ ➔ 추정치? $Q = \bar{y}_1 - (\bar{y}_2 + \bar{y}_3)/2$

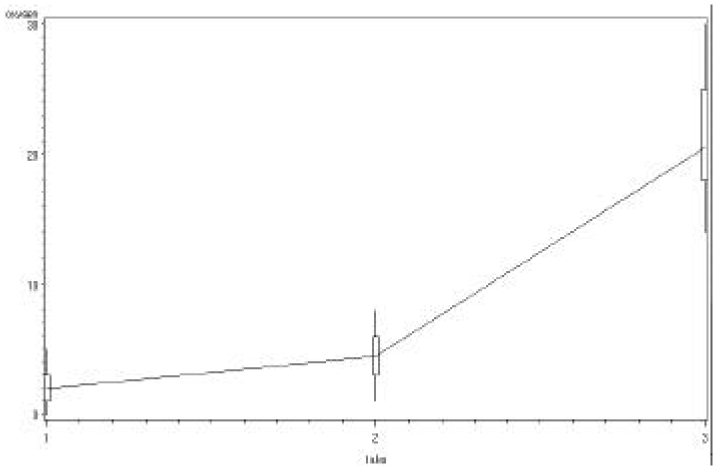
검정 통계량: $F = \dfrac{(Q)^2}{MSE(\sum \frac{c_i^2}{n_i})} \sim F(1, n-t)$ where $c = \sum c_i$ 이고 c_i 는 수준 앞 계수

o 동쪽의 호수를 하나 더 조사하여 호수 4의 자료를 얻은 경우 동쪽 호수와 서쪽 호수간의 산소량의 차이가 있는가?

$$Q = (u_1 + u_4) - (u_2 + u_3) \;\; ➔ \;\; c_1 = 1, c_2 = 1, c_3 = -1, c_4 = -1$$

SAS 이용하기

```
data lake;
    do i=1 to 10;
        input oxygen @@; lake=1; output;
    end;
    do i=1 to 10;
        input oxygen @@; lake=2; output;
    end;
    do i=1 to 10;
        input oxygen @@; lake=3; output;
    end;
    cards;
    0  2  1  3  1  2  3  4  1  5
    1  3  4  6  8  7  5  3  4  5
    14  26  25  18  19  22  21 16  20  30
    ;
    keep oxygen lake;
run;

symbol i=boxj;
proc gplot data=lake;
    plot oxygen*lake;
run;

proc glm data=lake;
    class lake;
    model oxygen=lake;
    means lake / tukey lines;
    contrast 'by region ' lake 2 -1 -1;
run;
```



Dependent Variable: oxygen

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2117.400000	1058.700000	105.52	<.0001
Error	27	270.900000	10.033333		
Corrected Total	29	2388.300000			

R-Square	Coeff Var	Root MSE	oxygen Mean
0.886572	34.05961	3.167544	9.300000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
lake	2	2117.400000	1058.700000	105.52	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
lake	2	2117.400000	1058.700000	105.52	<.0001

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	lake
A	21.100	10	3
B	4.600	10	2
B			
B	2.200	10	1

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
by region	1	756.1500000	756.1500000	75.36	<.0001

II HOMEWORK#12 II (due May 24)

다음은 연구비 (Research and Development) 투자액에 (High, Moderate, Low) 따른 생산성 향상 정도(10점 만점)를 나타낸 것이다. 다음 물음에 답하시오. (SAS 이용 가능)

분류	관측치
Low	7.6 8.2 6.8 5.8 6.9 6.6 7.7 6
Moderate	6.7 8.1 9.4 8.6 7.8 7.7 8.9 7.9 8.3 8.7 7.1 8.4
High	8.5 9.7 10.1 7.8 9.6 9.5

- 투자액에 따른 생산성 향상 점수에 대한 Box-Plot을 그리고 해석하시오.
- 투자액에 따른 생산성 향상 점수의 평균에 대한 Bar 차트를 그리시오.
- 분산 분석표를 작성하고 연구비 투자액에 따른 생산성 향상의 차이가 있는지 검정하시오.
- 각 투자액의 pairwise 다중 비교를 시행하고 해석하시오 (Tukey 방법 이용)
- 투자액을 High, (Low + Moderate) 나누었을 경우 두 그룹 간에 차이가 있는지 검정하시오.

6.2. Median 검정 이용

독립인 두 모집단 중앙값 비교에 사용한 Median 검정을 기억하나요? 모집단으로부터 추출한 두 표본 측정치들의 총 중앙값을 이용하여 1집단의 +부호(혹은 –부호) 개수와 2집단 +부호의 개수의 차이에 의해 두 모집단 중앙값의 차이를 검정한 방법이다. 이를 확장하여 모집단이 3개 이상인 경우 모집단 중앙값의 차이를 검정할 수 있다.

가정

- k개 모집단의 중앙값은 각각 $M_1, M_2, \ldots M_k$ 이다.
- 중앙값이 모두 같다면 표본 자료들 중 총 중앙값보다 큰 측정치의 비율은 모두 같다.
- 모집단으로부터 추출된 표본 자료는 서로 독립이다.
- 모집단에 대한 정규성, 등분산성의 가정 성립 여부에 대한 확신이 없다.

가설

귀무가설: 모집단의 중앙값은 동일하다. $H_0 : M_1 = M_2 = \ldots = M_k$

대립가설: 모든 모집단 중앙값은 같지 않다. 즉 적어도 하나의 모집단 중앙값은 다른 모집단 중앙값들과는 다르다.

검정통계량

1. 표본 자료 전체를 합하여 총 중앙값을 계산한다.
2. 각 표본의 측정치를 총 중앙값보다 큰 그룹과 이하인 그룹으로 나누어 아래와 같은 표를 만든다.

	표본				
	1	2	…	k	합계
>총 중앙값	O_{11}	O_{12}	…	O_{1k}	A
≤총 중앙값	O_{21}	O_{22}	…	O_{2k}	B
합계	n_1	n_2	…	n_k	N

3. 귀무가설이 성립한다면 {>총 중앙값} 그룹과 {≤총 중앙값} 그룹이 표본에 대해 동질성이 만족하게 된다.
4. $T = \sum\sum \frac{(O_{ij} - E_{ij})}{E_{ij}} \sim c^2(df = k-1)$ (예) $E_{11} = ???$

결론

표본으로부터 계산된 검정통계량 값과 $c^2(df = k-1)$ 분포표로부터 구한 기각 값(critical value: 기각 값보다 큰 영역을 기각역이라 한다)을 비교하여 검정통계량 값이 크다면 귀무가설을 기각한다.

II EXAMPLE II

신장 절제를 한 그룹 2개와 정상이 그룹 1개에 속한 쥐를 임의 추출하여 심실의 근육의 세포질 공간을 측정한 것이다. 이를 이용하여 세포질 공간의 중앙값이 그룹에 따라 차이가 있는지 비모수 검정하시오. (유의수준 $a = 0.05$)

절제 그룹1	185	187	209	194	175	197	188	185
절제 그룹2	189	193	176	195	169	183	185	179
정상 그룹	219	204	219	234	233	194	209	195

1) 귀무가설: 그룹별 세포질 공간의 중앙값은 같다. $H_0 : M_1 = M_2 = M_3$
2) 대립가설: 모두 같지는 않다.
3) 검정통계량: 총 중앙값은 193.5이다.

	표본			
	절제1	절제2	정상	합계
>총 중앙값	3	1	8	12
≤총 중앙값	5	7	0	12
합계	8	8	8	24

$$T = \frac{(3-4)^2}{4} + \frac{(5-4)^2}{4} + \ldots + \frac{(0-4)^2}{4} = 13$$

4) 결론: 계산된 검정 통계량 값(13)이 $c^2(df = 2, a = 0.05) = 5.99$ 이므로 귀무가설은 기각되고 그룹별 세포질 공간 중앙값은 차이가 있다고 결론 지울 수 있다.

II HOMEWORK#13 II (due June 5)

다음은 4개 식품에 철 함유량을 조사한 것이다. 그룹별 철 함유량 중앙값의 차이가 있는지 비모수 방법인 Median 방법을 사용하시오. (수작업: hand calculation) 유의수준=0.05

식품A	27	16	19	4	2	16	30	9	16	
식품B	44	34	43	47	22	35	51	22	37	29
식품C	17	45	28	13	36	3	42	41	15	
식품D	51	29	30	50	47	40	43	44	54	

6.3. Kruskal-Wallis 검정

3개 이상의 모집단 중앙값 차이 검정(비모수 분산 분석)에 가장 많이 사용되는 방법으로 모집단이 2개인 경우는 Mann-Whitney 방법과 동일하다.

가정

- k개 모집단의 중앙값은 각각 $M_1, M_2, \ldots M_k$ 이다.
- 적어도 하나 이상의 모집단 중앙값의 차이가 없다면 모든 중앙값은 동일하다.
- 모집단에 대한 정규성, 등분산성의 가정 성립 여부에 대한 확신이 없다.

가설

귀무가설: 모집단의 중앙값은 같다. $H_0 : M_1 = M_2 = \ldots = M_k$

대립가설: 모든 모집단 중앙값은 같지 않다. 즉 적어도 하나의 모집단 중앙값은 다른 모집단 중앙값들과는 다르다.

검정통계량

1. 표본 자료 전체를 고려하여 크기 순으로 정렬하여 순위를 매긴다.
2. 각 집단별 순위 합을 구한다.
3. 각 표본의 순위 합이 동일하다면 귀무가설이 채택될 것이고 순위 합이 서로 다르다면 귀무가설을 기각하게 될 것이다.
4. 검정 통계량: $T = \dfrac{12}{N(N+1)} \sum \dfrac{R_i^2}{n_i} - 3(N+1)$
    - N = 총 표본의 크기
    - R_i = i번째 표본의 순위합
    - n_i = i번째 표본의 크기

결론

Kruskal-Wallis 분포표를 이용하여 가설 검정을 한다. K-W 분포표는 각 표본의 크기가 5이하인 경우만 표가 나타나 있다. 표는 집단이 3개인 경우만 예제로 부록에 제시하였다.
표본의 크기가 5보다 큰 경우, 즉 표에 나타나지 않는 경우는 K-W 검정통계량이 $c^2(df = k-1)$ 분포에 근사하는 사실을 이용하여 가설을 검정하면 된다.

II EXAMPLE II

다음은 정맥 호르몬 운송 시간을 3그룹의 병원 환자들을 조사한 것이다. 이를 이용하여 환자 그룹별 운송 시간의 중앙값에 차이가 있는지 Kruskal-Wallis **검정을 시행하시오.** (**유의수준**=0.05)

환자 그룹 1	262	307	211	323	454	339	304	154	287	356
환자 그룹 2	465	501	455	355	468	362				
환자 그룹 3	343	772	207	1048	838	687				

1) **귀무가설: 환자 그룹별 호르몬 운송 시간 중앙값은 같다.** $H_0 : M_1 = M_2 = M_3$
2) **대립가설: 모두 같지는 않다.**
3) **검정통계량:**

환자 그룹	순위	순위 합계
1	4 7 3 8 14 9 6 1 5 12	69
2	16 18 15 11 17 13	90
3	10 20 2 22 21 19	94

$$T = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1) = \frac{12}{22(22+1)} [\frac{69^2}{10} + \frac{90^2}{6} + \frac{94^2}{6}] - 3(22+1) = 9.232$$

4) **결론: 계산된 검정 통계량 값**(13)**이** $c^2(df = 2, a = 0.05) = 5.99$ **이므로 귀무가설은 기각되고 환자 그룹별 운송 시간 중앙값은 차이가 있다고 결론 지울 수 있다.**

II HOMEWORK#13 II (계속 due June 5)

쥐 3 그룹에 대한 세로토닌(혈액: 뇌 속에 있는 혈관 수축 물질) 무게를 조사한 것이다. 그룹별 차이가 있는지 비모수 검정인 Kruskal-Wallis **검정하시오.** (**수작업**: hand calculation) **유의수준**=0.05**로 하시오.**

Control 그룹	340	340	356	386	386	402	402	417	433	495	557
LSD 그룹	294	325	325	340	356	371	385	402			
UML 그룹	263	309	340	356	371	371	402	417			

6.4. JONCKHEERE-TERPSTRA test

Median 확장 검정이나 Kruskal-Wallis 검정의 경우 대립 가설은 "모든 모집단의 중앙값이 같지는 않다" 이다. 예를 들면, 약의 복용량에 따라 약의 효과가 커지는지 알아보고자 하는 경우 대립가설은 $H_a: M_1 < M_2 < \ldots < M_k$ 이어야 한다. 이처럼 순서형(ordered) 대립가설을 검정할 때 사용하는 방법이 JONCKHEERE-TERPSTRA 검정이다.

가정

- k개 모집단의 중앙값은 각각 $M_1, M_2, \ldots M_k$ 이다.
- 관측치는 서로 독립이다.
- 모집단에 대한 정규성, 등분산성의 가정 성립 여부에 대한 확신이 없다.

가설

귀무가설: 모집단의 중앙값은 같다. $H_0: M_1 = M_2 = \ldots = M_k$

대립가설: 모집단 중앙값은 크기는 순서가 존재한다. $H_a: M_1 \le M_2 \le \ldots \le M_k$. 부호 중 적어도 하나는 < 이다.

검정통계량 : $J = \sum_{i<j} U_{ij}$

U_{ij} 는 표본 i와 j의 관측치를 쌍으로 비교하여 i표본의 관측치가 j표본의 관측치보다 적은 쌍의 수이다.

결론

JONCKHEERE-TERPSTRA 분포표를 이용하여 가설 검정을 한다. J-T 분포표는 집단이 3 개이고 각 표본의 크기가 $n_1 < n_2 < n_3$ 인 경우 기각역 값을 제시하고 있다. J-T 표에 나타나지 않는 경우 다음 근사 통계량을 이용하여 검정하면 된다.

$$T = \frac{J - [(N^2 - \sum_i^k n_i^2)/4]}{\sqrt{[N^2(2N+3) - \sum_i^k n_i^2(2n_i+3)]/72}} \sim N(0,1)$$

II EXAMPLE II

다음은 감기 환자 증상에 따른 정신 스트레스 수치이다. 감기 증상이 심해지면 스트레스 수치가 높아진다고 할 수 있는가? (유의수준=0.05)

약함	54	67	47	71	62	44	67	80
보통	79	82	88	79	85	81	88	
심함	98	99	95	93	98	91	94	

1) 귀무가설: 환자별 스트레스 수치 중앙값은 같다. $H_0 : M_1 = M_2 = M_3$
2) 대립가설: 증상에 따라 스트레스 수치는 커진다. $H_0 : M_1 \le M_2 \le M_3$
3) 검정통계량:

$$U_{12} = 54, U_{13} = 56, U_{23} = 49 \ \rightarrow \ J = \sum_{i<j} U_{ij} = 159$$

4) 결론: J-T 표 $(n_1 = 7, n_2 = 7, n_3 = 8)$, 유의수준 $a = 0.05$ 인 경우 기각역 값은 108이나 109이다. 159는 그 값보다는 크므로 귀무가설은 기각되고 감기 증상이 심해지면 스트레스 수치가 높아진다고 결론을 내릴 수 있다. 값이 134일 경우 유의확률이 0.00443이므로 p-값은 0.0443보다 적음을 알 수 있다.
5) 근사 통계량을 사용하면

$$J = 159, N = 8+7+7 = 22, \sum_{i=1}^{3} n_i^2 = 8^2 + 7^2 + 7^2 = 162, \sum_{i}^{3} n_i^2 (2n_i + 3) = 2882$$

$$T = \frac{159 - [(22^2 - 162)/4]}{\sqrt{[22^2 (2 \times 22 + 3) - 2882]/72}} = 4.73$$

표준정규분포표에 의해 p-값은 0.001 이하임을 알 수 있다. 비모수 검정 방법 J-T와 결과와 일치한다.

II HOMEWORK#14 II (due June 7)

다음은 독립인 세 그룹의 쥐들에 대한 성장 능력 지수이다. 이를 이용하여 성장 능력 지수가 그룹2 ≤그룹1≤그룹3 순인지 비모수 검정하시오 (유의수준=0.05). 그리고 p-값이 어느 정도인지 구하시오. J-T 분포표 이용 방법과 근사 분포인 표준 정규분포표를 이용한 방법 모두 사용하시오.

그룹 1	71	57	85	67	66	79		
그룹 2	76	94	61	36	42	49		
그룹 3	80	104	81	90	93	85	101	83

6.5. Multiple Comparison (다중 비교)

Median 확장 검정이나 Kruskal-Wallis 검정 방법에 의해 귀무가설이 기각된다면 우리는 단지 중앙값이 다른 집단이 적어도 하나 이상이다라는 결론만 내릴 수 있다. 그러므로 두 집단의 중앙값들의 차이가 있는지 (pairwise), 집단 그룹별 중앙값의 차이가 있는지 (대비, contrast) 알아보는 post-hoc test이 (혹은 multiple comparison) 필요하다. 대비 검정 방법은 계산 과정이 매우 복잡하므로 여기서는 pairwise 비교만 살펴보기로 하자.

모집단이 3개 있다고 가정하자. pairwise 검정의 경우 3개의 가설을 동시에 검정하게 된다. 이런 경우 적어도 3개의 귀무가설 중 적어도 하나 이상 잘못 기각할 확률은 $1-(1-a)^3$ 이다. 이 확률이 유의수준이 된다. 두 개 이상의 귀무가설을 동시에 검정하는 경우 유의수준(귀무가설이 참인데 기각할 확률)은 $1-(1-a)^c$ 이다. (c는 동시에 검정하는 귀무가설의 개수)

검정통계량

1) 각 표본의 관측치를 크기 순으로 정렬하여 각 표본의 순위 평균 $(\bar{R}_i)$ 을 구한다. (각 표본의 순위 계산 방법은 Kruskal-Wallis 검정과 동일)
2) 모집단의 수가 k인 경우 두 표본의 순위 합 평균 차이 $(|\bar{R}_i-\bar{R}_j|)$ 가 다음 값보다 크다면 귀무가설(두 모집단 중앙값은 같다)을 기각한다.

$$z_{(1-[a/k(k-1)])}\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

$\bar{R}_i$ = (i번째 표본 순위 합)/(i번째 표본의 크기) / N = 표본 전체 총 관측치 수

n_i = i번째 표본의 크기

순위 합 차이의 절대값이 검정 통계량으로 사용되므로 양측 검정과 같다. 그러므로 $k(k-1)/2$ 대신 $k(k-1)$ 이 사용된다.

결론

표준 정규분포표를 이용하여 기각역을 계산하고 순위 평균의 차이의 절대값이 기각역 값보다 크다면 귀무가설을 기각한다.

Ⅱ EXAMPLE Ⅱ

Kruskal-Wallis 검정의 예제(정맥 호르몬 운송 시간: page 83) 계속: (유의수준=0.15) 모집단의 개수 $k=3$이므로 표준 정규분포표로부터 $a/k(k-1)=0.15/3(2)=0.025$ 에 해당되는 값을 얻어야 한다. 그룹1과 그룹2의 pairwise 검정에 필요한 기각역을 구하면

$$1.96\sqrt{\frac{22(22+1)}{12}\left(\frac{1}{10}+\frac{1}{6}\right)}=6.57$$

$|\bar{R}_1-\bar{R}_2|=|69/10-90/6|=8.1$이 6.57보다 크므로 귀무가설(그룹1의 중앙값과 그룹2의 중앙값은 같다)을 기각한다. (SAS를 이용해서 하는 방법은 없다)

Ⅱ HOMEWORK#14 Ⅱ (계속 due June 7)

Kruskal-Wallis 검정의 숙제(page 83)에 대해 집단간 pairwise 검정을 실시하시오. (유의수준=0.1)

Ⅱ 기말고사에 대하여 Ⅱ

일시:	6월 22일(금) 오후 1시~
장소:	경상대 2층 통계실습실
시험방법:	컴퓨터 사용 가능, Open Book Exam.

기말 시험 점수는 70점 이상이어야 A 학점을 받을 수 있습니다. 물론 이전 성적과 합산한 점수도 A 학점 요건을 만족해야 합니다.

7. Goodness-of-fit test

우리의 관심이 모집단 모수(모 평균, 모 분산)에 대한 것이 아니라 분포 자체에 있다면 우리는 더 이상 모수적 방법을 사용할 수 없을 것이다. 왜? 모집단의 분포를 모르고 관심이 바로 그 분포이니까? 표본 평균처럼 모집단의 분포를 모르더라도 대표본인 경우 정규분포를 따른다는 정리(중심극한정리)와 같은 것을 사용할 수 없다.

모집단으로부터 추출한 표본이 어떤 분포에 적합한지(fit)를 알아보는 검정 방법을 적합성 검정 (Goodness-of-fit test)라 한다. 여기서는 다음 2가지 경우에 대해 검정하는 방법을 살펴보기로 한다.

1) 표본 자료가 귀무가설에 설정한 분포를 따르는가?
2) 독립인 두 표본 자료가 같은 분포를 따르는가? (본 강의에서는 다루지 않는다)

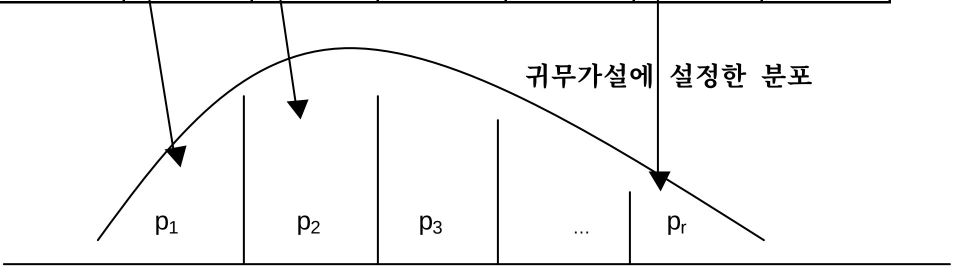
7.1. c^2 Goodness-of-fits test

독립성, 동질성 검정에 이용하였던 c^2(chi-square) 검정 방법을 이용하여 모집단 분포에 대한 검정을 하는 것이다. 방법은 매우 유사하다. 귀무가설에 설정한 분포로부터 기대(expected) 빈도를 얻고 표본으로부터 관측(observed) 빈도를 계산하여 이를 이용하여 귀무가설을 검정하면 된다.

가정

- 표본 관측치는 $(x_1, x_2, \ldots, x_n)$ 서로 독립이다. (확률 표본: random sample)
- 표본 자료로부터 분석자가 설정한 구간은 서로 겹치지 않고 (non-overlapping) 각 표본 관측치는 하나의 구간에 반드시 속한다.

구간 범주	1	2	3	…	r	합계
기대 빈도	E_1	E_2	E_3	…	E_r	n
관측 빈도	O_1	O_2	O_3	…	O_r	n



가설

귀무가설: 표본을 추출한 모집단은 XX 분포를 따른다.

대립가설: 표본을 추출한 모집단은 XX 분포를 따르지 않는다.

검정통계량 :

$$T=\sum_{i=1}^{r}\frac{(O_i-E_i)^2}{E_i}\sim c^2(df=r-1)$$

각 셀의 기대빈도가 5이상이면 근사 통계량이 c^2-분포를 사용할 수 있고 비록 5미만이 셀이 있더라도 전체 셀 중 20%이상만 되지 않으면 c^2-분포를 사용해도 무방하다. (Cochran) 만약 이 조건을 만족하지 않는 경우 범주를 합쳐 셀의 수를 줄이면 된다.

결론

c^2(자유도=r-1) 분포표로부터 기각역을 구하고 표본으로부터 계산된 검정 통계량 값과 비교하여 검정통계량이 기각역에 속하면 귀무가설을 기각한다.

분포에 대한 가설 검정의 경우 다른 검정과는 달리 우리의 관심은 귀무가설(어떤 분포에 따른다)에 있다.

모수 추정

귀무가설에 분포를 가정할 경우 모수(parameter) 값에 대한 가정이 필요하다. 예를 들어 정규분포를 따르는가? 알아보려고 하는 경우 구간 확률을 $(p_1, p_2, \ldots p_r)$ 계산하려면 평균(m), 표준 편차(s)를 알아야 구할 수 있다. 그러나 일반적으로 모수에 대한 사전 정보가 없으므로 표본 자료로부터 추정하게 된다. 모평균은 표본 평균으로, 모 분산은 표본 분산으로...

이렇게 모수가 추정되는 경우 c^2-분포의 자유도를 $(r-1)$에서 $(r-g-1)$로 조정해야 한다. g는 추정된 모수의 수이다.

II EXAMPLE1 II

상담하러 오는 사람이 상담원을 택할 때 선호하는 사람이 있는지 알아보기 위하여 상담하러 오는 사람 36명을 무작위로 선택하여 원하는 상담원을 고르라고 하여 다음 자료를 얻었다. (유의수준=0.05)

상담원	중년 여자	젊은 여자	중년 남자	젊음 남자	늙은 여자	늙은 남자
빈도	13	6	0	3	11	3

- 귀무가설: 상담하러 온 사람이 상담원을 선택에는 차별이 없다. (Equally distributed)
- 대립가설: 선택에 차별이 있다.

각 셀의 기대 빈도는 6이다. 왜냐하면 상담원의 선택에 차별이 없다면 각 셀에는 각가 6명이 기대된다.

상담원	중년 여자	젊은 여자	중년 남자	젊음 남자	늙은 여자	늙은 남자
빈도	13	6	0	3	11	3
기대	6	6	6	6	6	6

- 검정 통계량:

$$T = \frac{(13-6)^2}{6} + \frac{(6-6)^2}{6} + \ldots + \frac{(3-6)^2}{6} = 21.33$$

- 결론: 표로부터 c^2 (자유도=6-1=5, $a = 0.05$) =11.1이므로 귀무가설이 기각된다. 상담하러 온 사람들은 상담원 선택에 차이를 보이며 중년 여자나 늙은 여자를 선호한다.

II EXAMPLE2 II

다음은 실의 내구성을 시험을 위해 5개의 실을 묶어 만든 280 뭉치실의 내구성 테스트하여 다음 자료를 얻었다. 이 자료가 이항분포(binomial)로부터 왔는지 검정하시오. (유의수준=0.05)

끊어진 실 수	0	1	2	3	4	5
해당 실 뭉치	157	69	35	17	1	1

- 귀무가설: 위의 자료는 이항 분포를 따른다.
- 대립가설: 이항 분포를 따르지 않는다.

각 셀의 기대 빈도를 구하기 위해서는 이항 분포의 모수(p)를 알아야 한다.

$$p(x) = \binom{5}{x} p^x (1-p)^{5-x}, \quad x = 0,1,2,\ldots,5$$

표본 자료로부터 모수 p의 추정치$(\hat{p})$를 구하면

$$\hat{p} = (0\times157 + 1\times69 + 2\times35 + \ldots + 5\times1)/(280\times5) = 0.14$$

그러므로 기대빈도는 다음 이항 확률 분포에 의해 구한다.

$$p(x) = \binom{5}{x} 0.14^{x} (1-0.14)^{5-x}, \quad x = 0,1,2,\ldots,5$$

실험 장소	0	1	2	3	4	5
끊어진 실 수	157	69	35	17	1	1
기대 확률	0.4047	0.3829	0.1247	0.0203	0.0017	0.0001
기대 빈도	131.7	107.2	34.9	5.7	0.5	0.03

○ 검정 통계량:

$$T = \frac{(157-131.7)^2}{131.7} + \frac{(69-107.2)^2}{107.2} + \ldots + \frac{(1-0.03)^2}{0.03} = 44.99$$

○ 결론: 표로부터 c^2(자유도=5-1-1=3, $a = 0.05$) =7.81이므로 귀무가설이 기각되고 이 자료는 이항 분포를 따른다고 할 수 없다. 자유도가 3인 이유는 모수 하나를 추정하였기 때문이다.

|| EXAMPLE3 ||

다음은 한남대학교 정문을 통과하는 차량의 수가 Poisson 분포를 따르는지 알아보기 위하여 30초 마다 차량 통과 회수를 300회 조사하였다. 아래 자료를 이용하여 Poisson 분포를 따르는지 검정하시오. (유의수준=0.05)

통과 차량	0	1	2	3	4	5	6	7
관측 빈도	20	54	74	67	45	25	11	4

○ 귀무가설: 위의 자료는 Poisson 분포를 따른다.

○ 대립가설: Poisson 분포를 따르지 않는다.

각 셀의 기대 빈도를 구하기 위해서는 Poisson 분포의 모수(λ)를 알아야 한다.

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

표본 자료로부터 모수 p의 추정치($\hat{l}$)를 구하면

$$\hat{l} = (0\times20 + 1\times54 + \ldots + 7\times4)/300 = 2.67$$

그러므로 기대빈도는 다음 Poisson 확률 분포에 의해 구한다.

$$p(x) = \frac{e^{-2.67} 2.67^{x}}{x!}$$

통과 차량	0	1	2	3	4	5	6	7
관측 빈도	20	54	74	67	45	25	11	4
기대 확률	0.069	0.185	0.247	0.22	0.147	0.078	0.035	0.013
기대 빈도	20.7	55.5	74.1	66	44.1	23.4	10.5	3.9

○ 검정 통계량:

$$T = \frac{(20-20.7)^2}{20.7} + \frac{(54-55.5)^2}{55.5} + \ldots + \frac{(4-3.9)^2}{3.9} = -0.234$$

○ 결론: 표로부터 c^2(자유도=8-1-1=6, $a = 0.05$) =15.5이므로 귀무가설이 채택되고 이 자료는 Poisson 분포를 따른다고 할 수 있다.

II HOMEWORK#15 II (due JUNE 12)

다음은 1962-1967년 사이 대전-청주간 국도에서 사망 교통 사고 발생 건수이다. 매년 균등(uniformly distributed) 발생하였는지 검정하시오. (유의수준=0.05)

통과 차량	1962	1963	1964	1965	1966	1967
관측 빈도	7	12	6	10	12	6

II HOMEWORK#15 II (계속: due JUNE 12)

다음 자료가 Poisson 분포를 따르는지 검정하시오. 필요하면 셀을 합치시오. 셀을 합칠 때 중간에 있는 셀들을 합치면 안된다. 마지막의 셀들을 합해야 한다. 즉 ≥ 8셀을 만들어 $(8+9+10+11+12)$ 셀의 관측치를 합치면 된다. 기대 확률은 0-7까지 구하고 1에서 이 확률의 합을 빼면 된다. 그러나 $\hat{l}$을 구할 때는 전체 셀의 자료를 이용해야 한다. (유의수준=0.05)

발생 회수	0	1	2	3	4	5	6	7	8	9	10	11	12
관측 빈도	24	16	16	18	15	9	6	5	3	4	3	0	1

7.2. KOLMOGOROV-SMIRNOV one-sample test

c^2 (chi-square) 검정 방법은 연속형 변수(측정형 변수)에도 사용할 수 있다. 연속형 구간을 분석자가 임의로 8-10개의 구간으로 나누어 각 구간(셀)의 관측 빈도와 기대 빈도를 이용하여 c^2(chi-square) 검정 방법을 사용할 수 있다. (page 88 하단 그림 참조)

그러나 일반적으로 연속형 변수에 대해서는 K-S 검정을 주로 사용하는데 K-S 검정은 누적 분포 함수 (cumulative distribution) 개념을 이용한다. $F(x) = P(X \le x)$ 두 변수의 누적 확률 분포 함수가 동일하면 확률 밀도 함수도 동일하다는 정리(Theorem)가 이용된다.

가정

표본 관측치는 $(x_1, x_2, \ldots, x_n)$ 서로 독립이다. (확률 표본: random sample)

가설

귀무가설: $F(x) = F_0(x)$. 모집단의 누적 확률 분포 함수는 F_0 이다.

대립가설: $F(x) \ne F_0(x)$. 모집단의 누적 확률 분포 함수는 F_0 를 따르지 않는다.

검정통계량 :

크기 n의 표본 관측치로부터 자료 누적 함수 $S(x)$ =(x이하인 관측치의 개수)/(n)를 계산하고 다음 통계량을 이용하여 가설을 검정한다.

$$D = \sup_x | S(x) - F_0(x) |$$

결론

K-M 표를 이용하여 가설을 검정하면 된다.

모수 추정

모수를 추정해야 하는 경우 K-M 방법은 사용하는 경우 Massey는 모수가 추정된다면 1종 오류가 적어질 것이므로 가설 검정이 보수적으로 될 것이라고 지적하였다.

모수 추정 후 가설을 검정하면 된다.

II EXAMPLE II

다음 자료가 정규분포(평균=85, 표준편차=15)를 따름을 보이시오. (유의수준=0.05)

58	78	84	90	97	70	90	86	82
59	90	70	74	83	90	76	88	84
68	93	70	94	70	110	67	68	75
80	68	82	104	92	112	84	98	80

- **귀무가설**: **위의 자료는** $Normal(m = 85, s = 15)$ **분포를 따른다.**
- **대립가설**: $Normal(m = 85, s = 15)$ **를 따르지 않는다.**

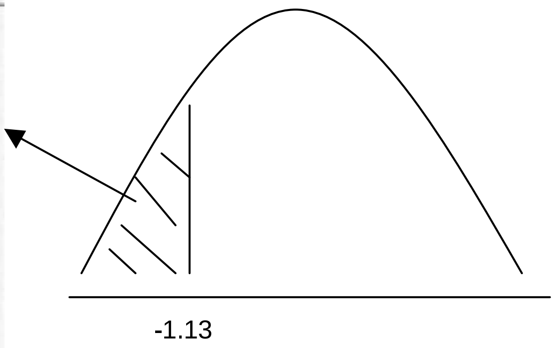
우선 자료를 크기 순으로 정렬한다.

관측 누적 분포함수: $S(x) = (\# of \le x)/n$

표본의 크기가 36이므로 각 x값에 대해 누적 빈도를 구하는 것과 같다. 물론 같은 관측치가 있으면 둘을 합쳐 누적 빈도를 구한다.

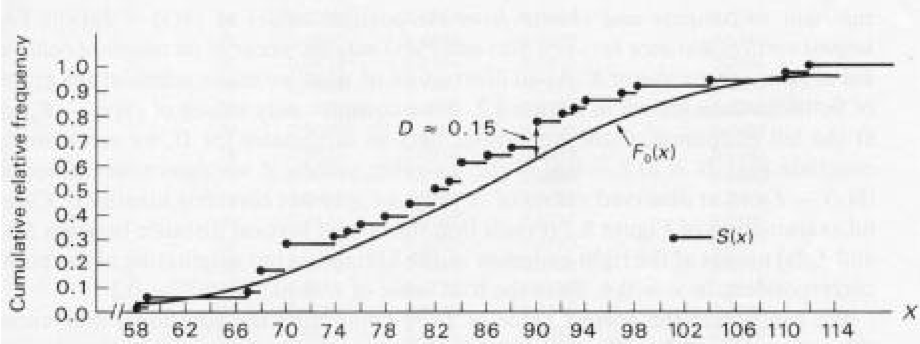
이론 분포 함수는 귀무가설에 설정한 분포 함수를 이용하여 계산한다.

x	$z = (x - 85)/15$	S(x)	$F_0(x)$
58	−1.80	1/36=.0278	0.0359
59	−1.73	2/36=.0556	0.0418
67	−1.20		0.1151
68	−1.13		0.1292
70	−1.00		0.1587
74	−0.73		0.2327
75	−0.67		0.2514
76	−0.60		0.2743
78	−0.47		0.3192
80	−0.33		0.3707
82	−0.20		0.4207
83	−0.13		0.4483
84	−0.07		0.4721
86	0.07		0.5279
88	0.20		0.5793
90	0.33		0.6293
92	0.47		0.6808
93	0.53		0.7019
94	0.60		0.7257
97	0.80		0.7881
98	0.87		0.8078
104	1.27		0.8980
110	1.67		0.9525
112	1.80		0.9641



x	$S(x)$	$F_0(x)$	$\lvert S(x) - F_0(x) \rvert$
58	0.0278	0.0359	0.0081
59	0.0556	0.0418	0.0138
67	0.0833	0.1151	0.0318
68	0.1667	0.1292	0.0375
70	0.2778	0.1587	0.1191
74	0.3056	0.2327	0.0729
75	0.3333	0.2514	0.0819
76	0.3611	0.2743	0.0868
78	0.3889	0.3192	0.0697
80	0.4444	0.3707	0.0737
82	0.5000	0.4207	0.0793
83	0.5278	0.4483	0.0795
84	0.6111	0.4721	0.1390
86	0.6389	0.5279	0.1110
88	0.6667	0.5793	0.0874
90	0.7778	0.6293	D = 0.1485
92	0.8056	0.6808	0.1248
93	0.8333	0.7019	0.1314
94	0.8611	0.7257	0.1354
97	0.8889	0.7881	0.1008
98	0.9167	0.8078	0.1089
104	0.9444	0.8980	0.0464
110	0.9722	0.9525	0.0197
112	1.0000	0.9641	0.0359

FIGURE 8.1 $S(x)$ and $F_0(x)$ for Example 8.5



○ **검정 통계량**: $D = \sup_x | S(x) - F_0(x) | = 0.1485$

○ **결론**: K-M**표로부터** n=36**이고 양측 검정인 경우 기각역 값은** 0.221**이다. 그러므로 귀무가설이 채택되고 이 자료는 정규분포**$(m = 85, s = 15)$ **를 따른다고 할 수 있다**.

II HOMEWORK#15 II (**계속**: due JUNE 12)

다음 자료가 정규분포를 따르는지 검정하시오. (**유의수준**=0.05)

1112	1435	1789	1489	1805	1738	1932	750	2513	3201
1205	935	2085	988	2450					

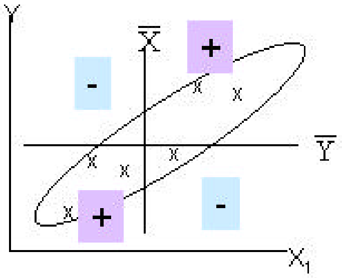
8. 비모수 상관 계수

분류형 두 변수간의 관계(association)에 대해서는 5장 독립성 검정에서 살펴보았다. c^2-검정 방법은 분류형 변수 중 명목형(nominal) 변수에 주로 사용된다.

이 장에서는 측정형 두 변수의 순위 상관 계수와 순서형 분류형 변수간의 상관 계수에 대해 살펴보기로 하자.

8.1. 모수적 방법

측정형 변수들의 상관 관계를 정도를 나타내는 상관 계수 계산식은 같다. 다음 계산식으로부터 얻어지는 상관 계수는 특히 Pearson 상관 계수라 한다.



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}$$

$$\hat{\beta}_1 = \frac{s_y}{s_x} r$$

$$s_x = \sqrt{s_{xx}} = \sqrt{\sum (x_i - \bar{x})^2}$$

가정　두 변수에 대한 표본 관측치는 (x_i, y_i) 이고 모두 측정형 변수이다.

가설

귀무가설: 두 변수 (x, y) 는 선형 상관 관계가 없다. 상관 계수가 0이다. $r = 0$

대립가설: 선형 상관 관계가 존재한다. $r \neq 0$

검정통계량 :

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} \sim t(df = n-2)$$, r =상관계수, n=표본의 크기

만약 귀무가설이 $H_0 : r = r_0 \neq 0$ 이라면 다음 검정 통계량을 사용해야 한다.

$$T = 0.5 \ln \frac{1+r}{1-r} \sim Normal(0.5 \ln \frac{1+r}{1-r}, 1/(n-3)) \rightarrow Z = \frac{T - 0.5 \ln \frac{1+r_0}{1-r_0}}{\sqrt{1/(n-3)}} \sim Normal(0,1)$$

결론

t (자유도=n-2) 분포표, 혹은 표준정규분포표로부터 기각역을 구하고 표본으로부터 계산된 검정 통계량 값과 비교하여 검정통계량이 기각역에 속하면 귀무가설을 기각한다.

II EXAMPLE II

키와 몸무게의 상관 관계를 보기 위하여 19명의 학생을 무작위 추출하여 옆의 자료를 얻었다. 표본 상관계수 0.878이었다. (유의수준=0.05)

○ 귀무가설: 키와 몸무게 상관 관계 없다. $r = 0$

○ 대립가설: 키와 몸무게 상관 관계 있다. $r \neq 0$

○ 검정통계량:

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.878}{\sqrt{(1-0.878^2)/(19-2)}} = 4.12$$

○ 결론: $t(df = 17, a/2 = 0.025) = 2.11$ 이므로 귀무가설이 기각되고 몸무게와 키는 양의 상관 관계를 갖는다고 결론 지을 수 있다.

학생	키	몸무게	
1	65	98	(x1, y1)
2	69	113	(x2, y2)
3	57	84	
4	63	103	
5	64	103	
6	57	83	
7	60	85	
8	63	113	
9	63	84	
10	59	100	
11	51	51	
12	64	90	
13	56	77	
14	67	112	
15	72	150	
16	65	128	
17	67	133	
18	58	85	
19	67	112	(x19, y19)

8.2. Spearman rank correlation coefficient (순위 상관 계수)

측정형 변수나 순서형 분류형 변수들의 상관 관계를 정도를 자료의 순위 값에 의하여 계산하는 방법으로 주로 순서형 분류형 변수들의 상관 관계를 계산한다.

가정

두 변수에 대한 표본 관측치는 (x_i, y_i) 이고 각 변수는 크기 순으로 정렬이 가능하다.

가설

귀무가설: 두 변수 (x, y) 는 서로 독립이다.

대립가설: 독립이 아니다. ←→ 선형 상관 관계가 존재한다.

검정통계량:

$$r_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad \text{where} \quad \sum d_i^2 = \sum_{i=1}^{n}[R(x_i) - R(y_i)]^2$$

$R(x_i)$ 는 x변수의 i번째 관측치의 순위이다. $R(y_i)$ 는 y변수의 i번째 관측치의 순위이다.

결론

○ 만약 두 변수 (x, y)가 완전한 양의 상관관계가 있다면 모든 d_i 는 0이고 $r_S = 1$이다.

○ 만약 두 변수 (x,y)가 완전한 음의 상관 관계가 있다면 d_i 는 다음과 같으므로 $r_S = -1$

$$[R(x)=1, R(y)=n],\ [R(x)=2, R(y)=n-1],\ \ldots,\ [R(x)=n, R(y)=1]$$

Spearman 표에서 기각역 값을 구하고 계산된 검정통계량과 비교한다. (본 강의에서는 이 표를 따로 싣지 않겠다. 대신 SAS 이용 방법을 예제에서 설명하기로 한다)

Large Sample approximation

$$z = r_S \sqrt{n-1} \sim Normal(0,1)$$

8.3. Kendall's Tau

측정형 변수나 순서형 분류형 변수들의 상관 관계를 정도를 자료의 순위 값에 의하여 계산하는 방법으로 주로 순서형 분류형 변수들의 상관 관계를 계산한다.

가정

두 변수에 대한 표본 관측치는 (x_i, y_i) 이다.

가설

귀무가설: 두 변수 (x, y) 는 서로 독립이다.

대립가설: 독립이 아니다. ←→ 선형 상관 관계가 존재한다.

검정통계량:

$$\tau = \frac{\sum_{i<j} \operatorname{sgn}(x_i - x_j)\operatorname{sgn}(y_i - y_j)}{\sqrt{(T_0 - T_1)(T_0 - T_2)}} \text{, where } \begin{array}{l} T_0 = n(n-1)/2 \\ T_1 = \sum t_i (t_i - 1)/2 \\ T_2 = \sum u_i (u_i - 1)/2 \end{array}$$

$n =$ 쌍의 관측치 수

$t_i =$ 주어진 순위에서 동일한(tied) X 관측치 수

$u_i =$ 주어진 순위에서 동일한(tied) Y 관측치 수

$$\operatorname{sgn}(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z = 0 \\ -1 & \text{if } z < 0 \end{cases}$$

결론

Kendall 표에 의해 가설을 검정하면 된다. (본 강의에서는 이 표를 따로 싣지 않겠다. 대신 SAS 이용 방법을 예제에서 설명하기로 한다)

Large Sample approximation

$$z = \frac{3\hat{\tau}\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \sim Normal(0,1)$$

8.4. Kendall's Tau와 Spearman 순위 상관 계수

○ 수작업의 경우는 Spearman이 더 유리하다.

○ n이 아주 크지 않은 경우 Kendall's Tau가 Spearman 보다 더 빨리 정규분포에 근사한다.

○ 둘은 상관 계수 값의 차이가 있으나 p-값은 거의 동일하다.

○ 일반적으로 Kendall's Tau가 모집단 상관 계수의 추정치이므로 더 선호된다.

8.5. 예제

다음은 5주 독서 클리닉에 참가한 학생들의 독서 능력 향상 정도와 IQ와의 상관계수를 구하고 해석하시오. (유의수준=0.05)

독서향상도	IQ
0.6	86
0.2	107
1.6	102
0.5	104
0.9	104
0.5	89
0.8	109
0.8	109
0.8	101
0.4	96
1.8	113
0.1	85
0.9	100
0.2	94
1.6	104
1.6	104
0.0	98
1.6	115
0.2	109
0.3	94

Editor - Untitled1 *

```
data one;
    input x y @@;
    cards;
    .6 86 .2 107 1.6 102 .5 104 .9 104 .5 89 .8 109
    .8 109 .8 101 .4 96 1.8 113 .1 85 .9 100 .2 94
    1.6 104 1.6 104 0 98 1.6 115 .2 109 .3 94
run;

proc corr data=one kendall spearman;
    var x y;
run;
```

○ KENDALL, SPEARMAN 옵션을 사용하지 않으면 SAS는 자동으로 PEARSON 상관 계수가 출력된다.

```
Spearman Correlation Coefficients, N = 20
        Prob > |r| under H0: Rho=0

              x              y

x       1.00000        0.49752
                        0.0256

y       0.49752        1.00000
         0.0256

Kendall Tau b Correlation Coefficients, N = 20
        Prob > |r| under H0: Rho=0

              x              y

x       1.00000        0.35957
                        0.0348

y       0.35957        1.00000
         0.0348
```

○ Spearman 순위 상관계수는 0.4975이고 p-값은 0.0256으로 0.05보다 작으므로 귀무가설이 기각된다. 향상 정도와 IQ와는 양의 상관 관계가 있다.

○ Kendall 상관계수는 0.3596이고 p-값은 0.0348으로 0.05보다 작으므로 귀무가설이 기각된다. 향상 정도와 IQ와는 양의 상관 관계가 있다.