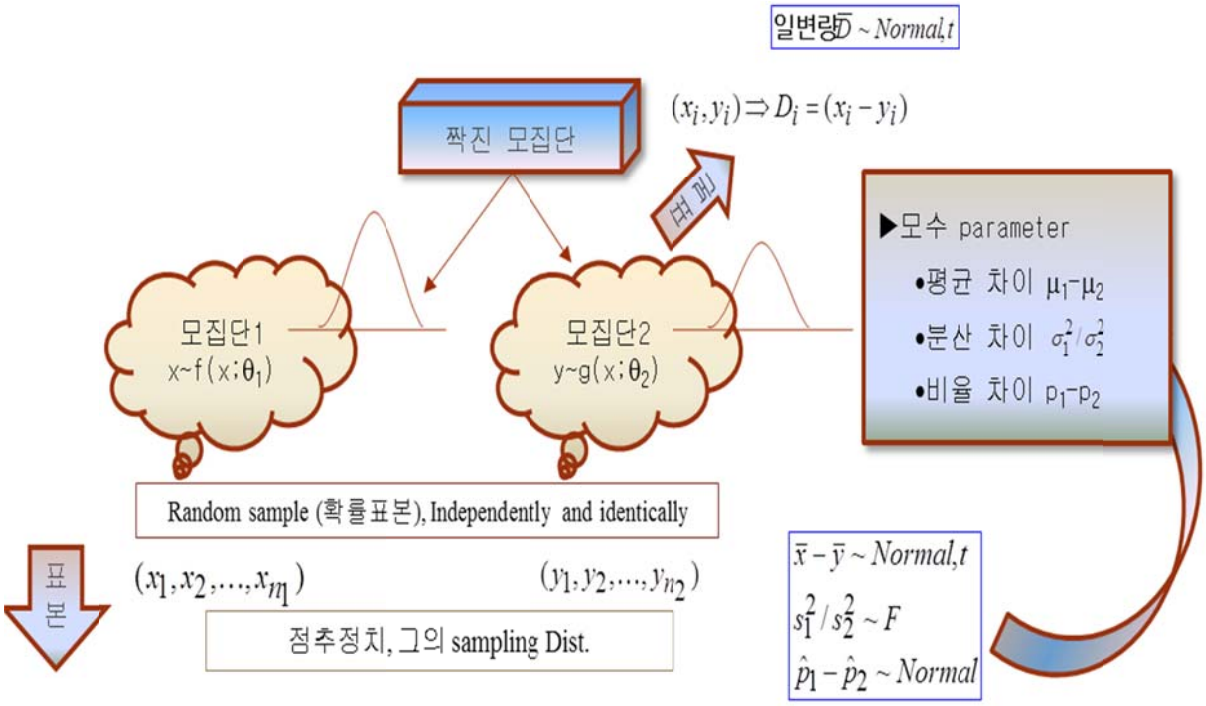
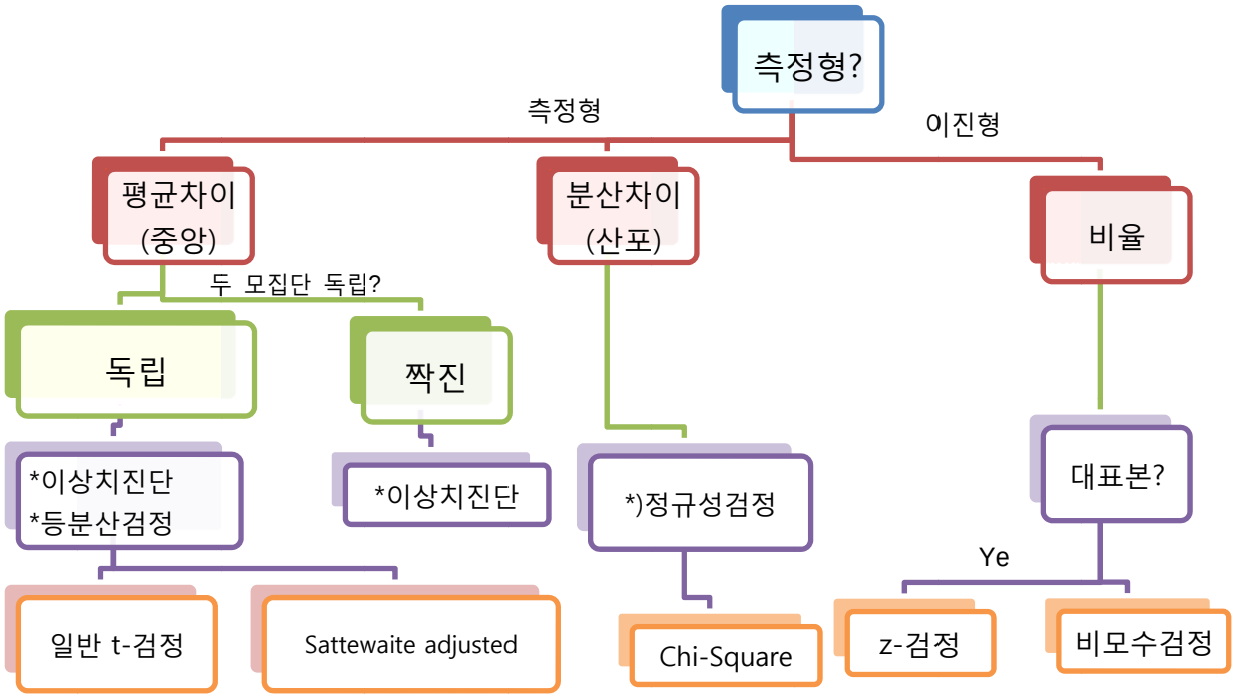
1. 두 모집단 차이 관련 추론 구성도



2. 두 모집단 추론 과정



Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

3. 독립인 independent 두 모집단 평균 ($\mu_1 - \mu_2$) 차이 검정

Task I 연구문제 정의

(1) 비즈니스 문제를 통계적 문제로 기술

- 같은 도시지만 커피 소비가 많은 시애틀과 Atlanta 는 우유 소비량이 많아 우유 값이 높을 것이다.
- 모수 parameter $(\mu_1 - \mu_2) = ?$
- 두 도시(모집단)는 서로 독립

(2) 수집 데이터 (MILK.xls) 내용 정리

- Seattle 지역 21 개, Atlanta 지역 18 개 도매점 가격조사

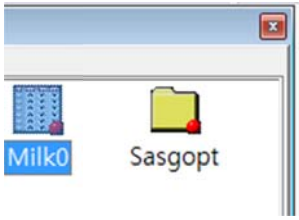
	A	B
1	Seattle	Atlanta
2	2.55	2.25
3	2.67	2.3
4	2.5	2.49

(3) 적절한 통계적 방법론 : 독립인 두 모집단 평균 차이 검정

SAS 엑셀 데이터 읽기, 각 열에는 오직 하나의 확률변수만 갖도록

```
/*외부 엑셀 데이터 읽어 SAS 데이터 만들기 */
PROC IMPORT DATAFILE="D:₩TMP₩MILK.XLS" OUT=MILK DBMS=EXCEL REPLACE;
        SHEET="DATA"; GETNAMES=YES;
RUN;

/*한 열에 한 변수(측정항목)만 있다면 필요없는 프로그램 (시작)*/
DATA MILK0;
        SET MILK;
        PRICE=SEATTLE;GROUP="SEATTLE";OUTPUT;
        PRICE=ATLANTA;GROUP="ATLANTA";OUTPUT;
        KEEP PRICE GROUP;
RUN;
```



	price	group
1	2.55	seattle
2	2.25	atlanta
3	2.67	seattle
4	2.3	atlanta
5	2.5	seattle
6	2.49	atlanta

Task II 데이터 검증

(1) 이상치

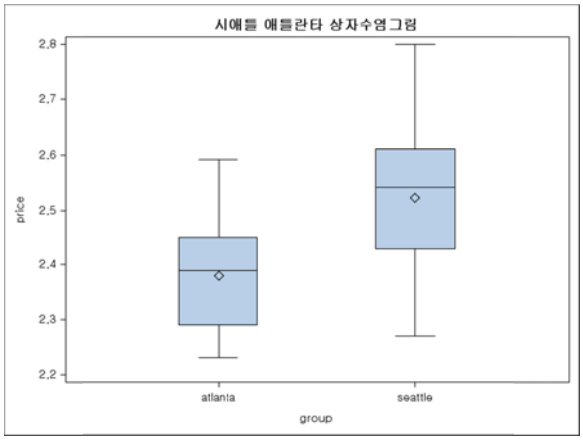
- 집단별 상자-수염 그림으로 진단 ($< Q_1 - 1.5 * IQR,\ > Q_3 + 1.5 * IQR$)
- 이상치 제거

(2) 등분산 검정

- 통계소프트웨어(Minitab 제외) 자동적으로 등분산 검정을 실시함
- 귀무가설 : $\sigma_1^2 = \sigma_2^2$
- 검정통계량 : $TS = \dfrac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)} \sim F(n_{분자} - 1, n_{분모} - 1)$

SAS | 이상치 진단

```
/*상자수염 그림 BY 집단 */
TITLE "시애틀 애틀란타 상자수염그림";
PROC SGPLOT DATA=MILK0;
        VBOX PRICE / CATEGORY=GROUP;
RUN;
```



이상치 없음

Task III 통계적 검정

(1) 통계적 가설

- 귀무가설 : $H_0 : \mu_1 = \mu_2$ 시애틀과 애틀란타 우유 가격은 동일하다
- 대립가설 : $H_a : \mu_1 < \mu_2$ 시애틀 우유 가격이 높다

(2) 등분산 검정 / 검정통계량 및 유의확률 계산

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

- 검정통계량 : $TS = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s(\bar{x}_1 - \bar{x}_2)} \sim t(???)$,
- 등분산 $s(\bar{x}_1 - \bar{x}_2) = SE = s_p = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}}$, $TS \sim t(n_1 + n_1 - 2)$
- 이분산 : $TS = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n2}}} \sim t(df * complicated)$

SAS 독립 집단 차이 검정

```
/*외부 엑셀 데이터 읽어 SAS 데이터 만들기 */
PROC TTEST DATA=MILK0;
        CLASS GROUP;
        VAR PRICE;
RUN;
```

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	20	17	1.62	0.3182

- 등분산 검정 : 유의확률이 0.32 이므로 귀무가설 채택 => 등분산

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	37	-3.73	0.0006
Satterthwaite	Unequal	36.751	-3.80	0.0005

- 등분산이므로 Pooled 행에 있는 검정 통계량 -3.73, 유의확률 0.0006/2=0.0003 (단측검정이므로) => 귀무가설 기각

Task IV 비즈니스 리포트

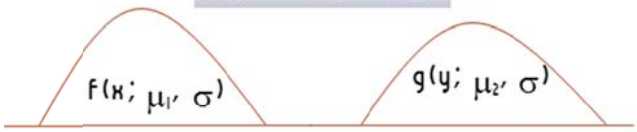
다음 요약표와 함께 95% $(\mu_1 - \mu_2)$ 신뢰구간 제시 : $(\bar{x}_1 - \bar{x}_2) \pm t(???)SE$

도시	n	평균 (M)	표준편차 (SD)	t-통계량	유의확률
Atlanta	18	2.38	0.101	3.73	0.0003
Seattle	21	2.52	0.129		

- 시애틀 지역의 우유 값이 애틀란타 지역보다 높다.
- 95% 시애틀과 애틀란타 우유 가격 평균 차이 신뢰구간 : (-$0.22, -$0.06)
- Diff (1-2) Pooled -0.1403 -0.2165 -0.0641

4. 짝진 paired 두 모집단 평균 $(\mu_1 - \mu_2)$ 차이 검정

$$\mu_d = (\mu_1 - \mu_2)$$



$$d_i = (x_i - y_i)$$

$$\bar{d} = \sum_{i=1}^{n} d / n_1 \qquad \bar{d} \sim univariate$$

$$s_d^2 = \sum_{i=1}^{n} (d_i - \bar{d}) / (n-1)$$

Task I 연구문제 정의

(1) 비즈니스 문제를 통계적 문제로 기술

- 광고는 고객의 구매욕구를 높일 것이다. => 모수 parameter $(\mu_2 - \mu_1) = ?$
- 광고 전후 동일 고객의 판단이므로 짝진 표본

(2) 수집 데이터 (AD.xls) 내용 정리

-

individual	after	before
1	6	5
2	6	4

(3) 적절한 통계적 방법론 : 짝진 두 모집단 평균 차이 검정

SAS 엑셀 데이터 읽기, 각 열에는 오직 하나의 확률변수만 갖도록

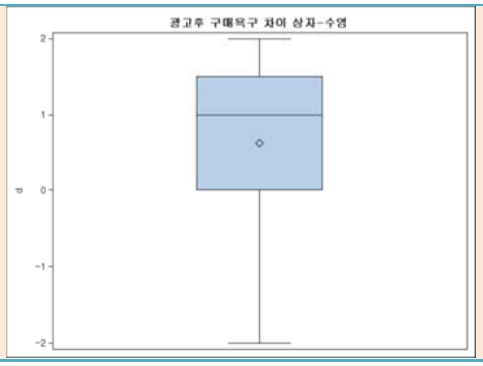
```
/*외부 엑셀 데이터 읽어 SAS 데이터 만들기 */
PROC IMPORT DATAFILE="D:₩TMP₩TVAD.XLS" OUT=AD DBMS=EXCEL REPLACE;
        SHEET="SHEET2"; GETNAMES=YES;
RUN;
```

Task II 데이터 검증

(이상치) 짝진 개체의 값의 차이가 다른 개체에 비해 이상할 정도로 크거나 작은지 진단 (실험이나 측정오차로 기인할 수 있으므로)

SAS | 엑셀 데이터 읽기, 각 열에는 오직 하나의 확률변수만 갖도록

```
/*차이 이상치 진단 */
data ad0;
        set ad;  d=after-before;
run;
/*상자수염 그림 */
TITLE "광고후 구매욕구 차이 상자-수염";
PROC SGPLOT DATA=ad0;
        VBOX d;
RUN;
```



Task III 통계적 검정

(1) 통계적 가설

- 귀무가설 : $H_0 : \mu_1 = \mu_2$ 광고 전후의 고객 구매욕구는 동일하다
- 대립가설 : $H_a : \mu_1 < \mu_2$ 광고의 효과가 있다.

(3) 등분산 검정 / 검정통계량 및 유의확률 계산

- $TS = \dfrac{\bar{d} - \mu_{\bar{d}}}{SE(\bar{d}) = s(d)/\sqrt{n}} \sim t(n-1)$ 1 모집단 평균 추론

SAS | 짝진 집단 차이 검정

```
/*외부 엑셀 데이터 읽어 SAS 데이터 만들기 */
TITLE "짝진 T-검정";
PROC TTEST DATA=AD0;
        PAIRED BEFORE*AFTER;
RUN;
/*일변량 분석 */
PROC TTEST DATA=AD0;
        VAR D;
RUN;
```


Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

Difference: before − after

N	Mean	Std Dev	Std Err	Minimum	Maximum
8	-0.6250	1.3025	0.4605	-2.0000	2.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-0.6250	-1.7139	0.4639	1.3025	0.8612	2.6509

DF	t Value	Pr > |t|
7	-1.36	0.2168

N	Mean	Std Dev	Std Err	Minimum	Maximum
8	0.6250	1.3025	0.4605	-2.0000	2.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0.6250	-0.4639	1.7139	1.3025	0.8612	2.6509

DF	t Value	Pr > |t|
7	1.36	0.2168

Task IV 비즈니스 리포트

SAS 표 만들기

```
/*기초통계량 표 만들기 */
PROC TABULATE DATA=AD0;
        VAR BEFORE AFTER;
TABLE (BEFORE AFTER), (MEAN STD);
RUN;
```

	Mean	Std
before	5.38	1.60
after	6.00	1.85

광고	평균 (M)	표준편차 (SD)	t-통계량	유의확률
전	5.38	1.6	1.36	0.2168
후	6.00	1.85		

- 귀무가설을 기각할 수 없음. 광고 효과 없음
- 광고 효과 ⇔ 광고 전후 구매욕구 차이 95% 신뢰구간 : (-0.46, 1.72)

5. 두 모집단 비율 ($p_1 - p_2$) 차이 검정

Task I 연구문제 정의

(1) 비즈니스 문제를 통계적 문제로 기술

Mark McGwire 와 Sammy Sosa 홈런 경쟁이 있었던 1998 년 여성 야구팬이 늘었는지 알고 싶다.

(2) 수집 데이터 내용 정리

- (CNN/USA) 1998 년에 1082 여성들을 대상으로 야구팬인가를 묻는 질문에 682 명이 그렇다고 했다. 1995 년에는 1008 명 중 413 명이 야구 팬이라고 했다.

(3) 적절한 통계적 방법론 :두 모집단 비율 차이 검정

Task II 데이터 검증

대표본 이론 : $\min(n_1 p_1, n_1 q_1) > 5$, $\min(n_2 p_2, n_2 q_2) > 5$

Task III 통계적 검정

(1) 통계적 가설

- 귀무가설 : $H_0 : p_1 = p_2$ 1995 년 1998 년 여성 야구팬 응답 비율 동일
- 대립가설 : $H_a : p_1 < p_2$ 야구팬이라 응답한 비율 증가 ⇔ 홈런 경쟁효과 있음

(2) 검정통계량

$$T = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}} \sim z, \hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

-

	A	B	C	D
1		n	x	phat
2	1995	1008	413	0.409722
3	1998	1082	682	0.630314
4			pool phat=	0.523923
5			TS=	10.08993
6			p-value=	0

Task IV 비즈니스 리포트

귀무가설을 기각되므로 홈런 경쟁으로 인하여 여성 야구팬 증가

비율 차이 신뢰구간 : $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{(\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2})}$

z(a/2)=	1.959964
차이	0.220592
상한	0.262413
하한	0.178771

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr