5.6 PROC MEANS

연속형 변수(continuous, metric, measurable)에 대한 기초통계량(elementary statistics, 평균, 중앙값, 백분위 값, 분산, 범위 등)을 계산하거나 모집단 평균에 대한 가설 검정을 하는데 사용되는 procedure이다.

5.6.1 중앙 위치

평균(mean): 평균은 관측치의 절대 크기의 중앙이므로 모든 관측치를 더한 값을 관측치 수(n)로 나눈 값이므로 산술 평균(arithmetic average)과 동일한 개념이다. n개의 관측치 $(x_1, x_2, \ldots, x_n)$ 의 평균은 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$ 이다. 예제 (1, 4, 6, 5, 6, 2)의 평균은 (1+4+6+5+6+2)/6=4이다.

중앙값(median): 자료의 크기 중심인 평균과는 달리 중앙값은 자료의 순서의 중심이다. 자료의 중앙값을 계산하기 위하여 자료의 순서 통계량(order statistics)을 먼저 구해야 한다. 순서 통계량이란 관측치를 크기 순으로 정렬한 후 제일 작은 값부터 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ 으로 표기하고 이를 순서 통계량(order statistics) 이라 한다. 만약 표본의 크기 n이 홀수이면 $M = x_{(\{n+1\}/2)}$, 짝수이면 $M = [x_{(n/2)} + x_{(\{n+2\}/2)}]/2$ 이 중앙값이 된다. 순서통계량(order statistics): 크기가 n인 표본 자료 관측치(observation) $(x_1, x_2, \ldots, x_n)$ 을 크기 순으로 정렬한 후 가장 작은 관측치를 $x_{(1)}$, 가장 큰 관측치를 $x_{(n)}$ 이라 표현하고 $x_{(1)}$, $x_{(2)}$, .., $x_{(n)}$ 을 순서 통계량이라 한다. 순서 통계량에 대해 다음이 성립한다.

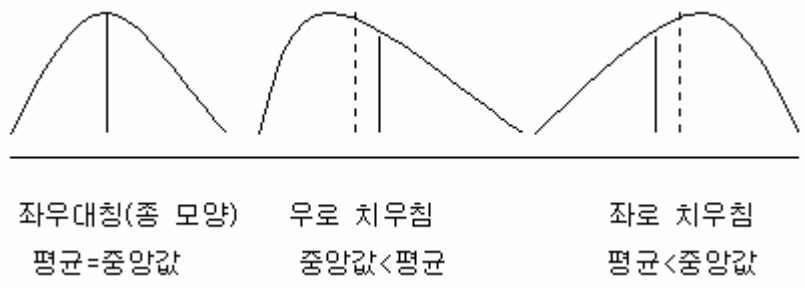
ⅰ) $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ ⅱ) 최소값 $x_{(1)}$ ⅲ) 최대값 $x_{(n)}$) 범위(range): $x_{(n)} - x_{(1)}$

예제 크기 6인 표본 관측치 (1, 4, 6, 5, 6, 2)의 순서 통계량은 (1, 2, 4, 5, 6, 6)이다. 최소값은 $x_{(1)} = 1$, 최대값은 $x_{(6)} = 6$ 이고 범위는 $x_{(6)} - x_{(1)} = 5$ 이다.

깊이(depth): 사분위 값을 구하려면 자료의 깊이 (depth) 개념을 이용하면 편리하다. (Tukey 제안) 관측치를 크기 순으로 정렬한 후 각 양쪽 끝에서 1부터 번호를 매겨 그 번호를 자료의 깊이라 정의한다. 즉 최대값, 최소값의 깊이는 각 1이다. Depth(중앙값=M)= $(n+1)/2$ 이고 사분위 깊이는 Depth(Q1)=Depth(Q3)=([Depth(M)]+1)/2이다. (기호)[x]=x를 넘지 않는 최대 정수.

예제 크기 6인 표본 관측치 (1, 4, 6, 5, 6, 2)의 중앙값 길이는 (6+1)/2=3.5이고, 사분위 길이는 ([3.5]+1)/2=(3+1)/2=2이다.

평균과 중앙값 비교: 자료의 측정치 중 다른 측정치에 비해 아주 크거나 아주 작은 측정치(극단치)가 존재하는 경우 순서의 중심인 중앙값과는 달리 크기의 중심인 평균은 극단치가 존재하는 쪽으로 치우치게 된다. 극단치 중 수집 자료에 포함하여 분석하기에는 부적절하게 크거나 작은 측정치를 이상치(outlier)라 한다. 다음은 확률 분포 함수 (히스토그램) 형태에 따른 중앙값과 평균의 관계이다.

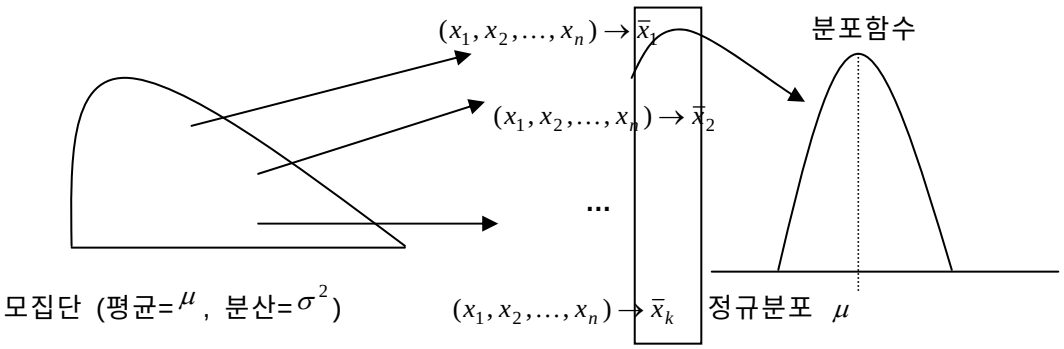


예제 크기 10인 자료 (1, 2, 3, 4, 5, 6, 7, 8, 9, 55)의 경우 평균과 중앙값을 구하면 평균은 $\bar{x} = 10$ 이고 중앙값은 $M = [x_{(5)} + x_{(6)}]/2 = (5+6)/2 = 5.5$ 이다. 중앙값 길이는 (10+1)/2=5.5 이고 사분위 길이는 ([5.5+1])/2=3이다. 그러므로 제 일사분위 Q_1은 $x_{(3)} = 4$ 이다.

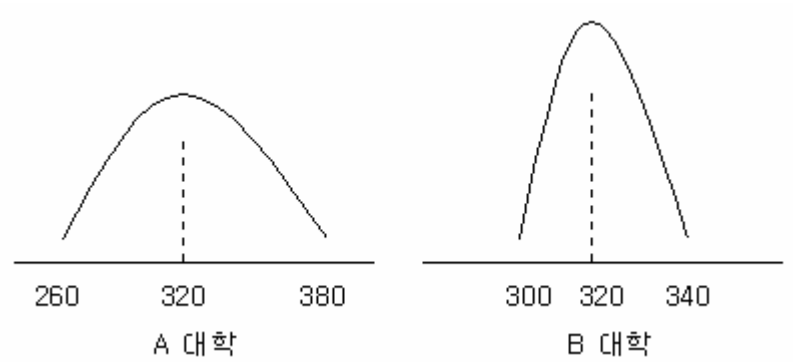
위의 자료에서 중앙 위치를 나타내는 숫자 요약으로는 중앙값인 5.5가 평균인 10보다 더 합리적이다. 이와 같이 극단치가 존재하는 경우 자료의 중앙 위치를 나타내는 통계량으로 중앙값이 평균보다 더 이상적인 값이다. 그리고 자료에 극단치가 존재하지 않으면 중앙값과 평균이 거의 일치하므로 자료의 중앙 위치에 대한 요약 값이라는 측면에서는 중앙값이 평균보다 더 합리적이다. 중앙값이 평균에 비해 더 합리적인 중앙 위치 값임에도 불구하고 평균을 사용하는 이유는 다음과 같다. 평균에 대한 ①수학적 전개가 용이하고 중앙값과는 달리 ②평균의 분포함수를 쉽게 구할 수 있기 때문이다. 중심극한정리(Central Limit Theorem)에 의하면 표본의 크기 n인 큰 경우(대표본) 모집단의 분포에 상관없이 표본 평균의 분포 함수는 정규분포를 따른다. 통계량의 분포함수를 구할 수 있는 것이 장점이 되는 이유는 통계량의 분포를 알아야 모수에 대한 신뢰구간을 구하거나 모수에 대한 가설 검정이 가능하다.

⏭중심극한정리 (Central Limit Theorem): 표본의 크기 n이 큰 경우 (20~30 이상) 모집단의 분포 형태에 관계없이 표본 평균의 분포는 정규분포를 따른다.

$$\bar{x} \sim Normal(\mu, \frac{\sigma^2}{n})$$



6.6.2 흩어진 정도(산포도: spread)



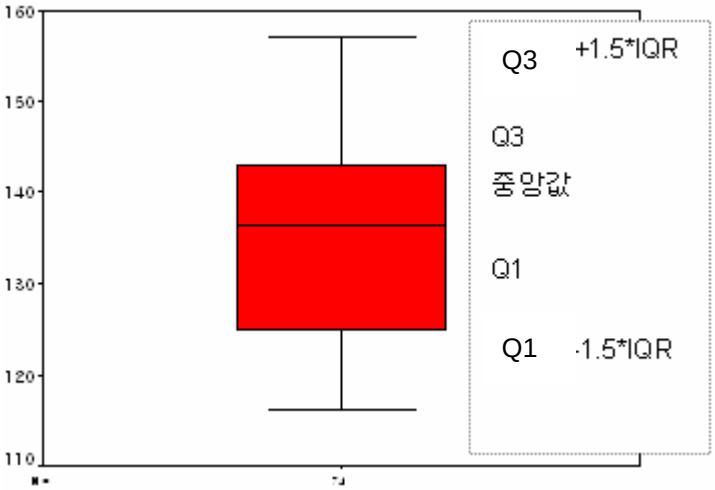
위의 그림은 두 대학의 수능 성적 자료 분포 함수이다. 자료의 중앙 위치 값(320점)만으로 보면 두 대학 신입생들의 성적은 동일하다고 말할 수 있을 것이다. 그러나 확률 분포 함수를 살펴보면 두 대학 신입생들의 수능 성적은 전혀 다른 정보를 가지고 있음을 쉽게 알 수 있다. A 대학의 신입생은 매우 우수한 학생이 있고 상대적으로 성적이 낮은 학생들도 있으나 B 대학 신입생들의 성적은 큰 차이를 보이지 않고 평균 점수에 집중되어 있다.

범위와 IQR: 측정 자료의 최대값과 최소값의 차이를 범위(range)라 한다. 범위는 계산이 편리하다는 장점이 있으나 범위 계산할 때는 두 측정치(최대값, 최소값)만 사용되므로 다른 측정치들의 정보가 전혀 고려되지 않고 이상치나 극단치가 존재하는 경우 범위가 커지는 단점이 있다. 크기가 10인 자료 (1, 2, 3, 4, 5, 6, 7, 8, 9, 55)와 같이 극단치가 존재하면 자료의 범위는 54

로 커지게 된다. 극단치가 존재하는 경우 산포도의 계산 값인 범위가 커지는 단점을 보완하기 위하여 삼 사분위 값과 일 사분위 값의 차이인 IQR 값을 산포도로 사용하기도 하지만 이것 역시 다른 측정치의 정보는 무시되는 단점을 가지고 있다.

p%-percentile(백분위 값): 데이터 관측치 중 p%가 그 값보다 작고 (1-p)%가 그 값보다 클 때 그 값을 p% 백분위 값이라 한다. 일사분위(First Quartile, Low Quartile) Q1은 관측치 중 25%가 그 값보다 작고 75%가 그 값보다 클 때 그 값을 일사분위라 정의한다. 이사분위(Second Quartile, Median) Q2은 관측치 중 50%가 그 값보다 작고 자료의 50%가 그 값보다 클 때 그 값을 이사분위라 정의하고 이를 특히 중앙값이라 한다. 삼사분위(Third Quartile, Upper Quartile) Q3는 관측치 중 75%가 그 값보다 작고 자료의 25%가 그 값보다 클 때 그 값을 삼사분위라 정의한다. 그리고 (Q3-Q1)을 자료의 IQR(Inter-Quartile Range)라 한다.

예제 크기 6인 표본 관측치 (1, 4, 6, 5, 6, 2)의 순서 통계량은 (1, 2, 4, 5, 6, 6)이다. 중앙값의 깊이는 $(6+1)/2 = 3.5$ 이고 일사분위와 삼사분위 깊이는 $([3.5]+1)/2 = 2$ 이다. 그러므로 일사분위 $Q_1 = x_{(2)} = 2$, 이산분위 중앙값 $Q_2 = (4+5) = 4.5$, 삼사분위 $Q_3 = x_{(5)} = 6$ 이다. 그러므로 $IQR = 6-2 = 4$ 이다.



표준편차 및 분산: 자료의 산포도 측정치로 가장 많이 사용되는 분산(variance)은 각 측정치(x_i)들이 평균($\bar{x}$)으로부터 떨어진 정도(차이)를 제곱한 값들을 합한 후 자료의 수로 나눈 값이고 표준 편차(standard deviation)는 분산의 양의 제곱근 값으로 정의된다. 모집단 전체 자료의 분산을 모집단 분산(σ^2), 표준 편차를 모집단 표준 편차(σ: sigma)라 하고 표본 자료의 경우는 표본 분산(s^2), 표본 표준 편차(s)라 한다.

표준편차나 분포는 순서 통계량에 의해 자료의 흩어진 정도를 나타내는 범위나 IQR과는 달리 측정치들이 평균으로부터 떨어진 정도를 숫자로 나타낸다. 측정치와 평균의 차이를 제곱함으로써 멀리 떨어질수록 자료의 흩어진 정도에 더 많은 영향을 미치게 된다. 즉 표준 편차나 분산은 자료들이 평균으로부터 평균적으로 얼마나 떨어져 있는지를 나타내는 수치이다.

⏭ 표본 분산, 표준 편차 계산식: $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2]$, $s = \sqrt{s^2}$

예제 크기가 6인 자료 (1, 4, 6, 5, 6, 2)의 분산과 표준 편차를 구해보자.

$$s^2 = \frac{1}{6-1}\sum_{i=1}^{6}(x_i - 4)^2 = \frac{(1-4)^2 + (4-4)^2 + \cdots + (2-4)^2}{5} = 4.4, \quad s = \sqrt{4.4} = 2.098$$

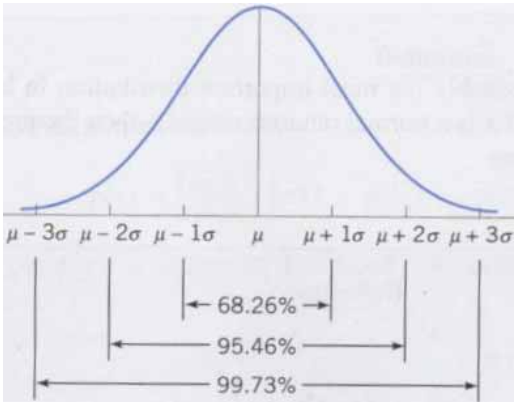
평균과 표준편차: 측정형 자료에 대한 기초 통계량을 정리할 때 평균과 표준 편차를 사용하는 것이 좋다. 분산은 측정치들을 제곱한 값들을 합한 것이므로 측정치들을 단순히 합한 평균과는 단위가 다르지만 제곱근을 구한 표준 편차는 평균과 단위가 일치하기 때문이다.

참고 Empirical Rule(경험적 규칙): 자료의 분포가 좌우 대칭이면 다음이 성립한다.

(1)전체 관측치 중 적어도 68.3%는 평균 $\pm$ 표준편차 범위에 있다

(2)전체 관측치 중 적어도 95.5%는 평균 $\pm 2 *$표준편차 범위에 있다

(3)전체 관측치 중 적어도 99.7%(대부분)는 평균 $\pm 3 *$표준편차 범위에 있다



변동계수(CV: Coefficient of Variation): 측정 단위에 따라 표준 편차의 값의 크기가 달라지므로 단위가 다른 두 집단을 비교하는 경우 두 표준 편차의 단위를 같게 할 필요가 있다. 이를 위하여 표준편차를 평균으로 나눈 값에 100을 곱한 값을 변동 계수(CV: Coefficient of

Variation)라 하고 상대 변동(분산) 개념으로 정의하고 있다.

⏭ 변동계수: 표본 자료의 평균이 $\bar{x}$, 표준 편차가 s인 경우 $CV = \frac{s}{\bar{x}} \times 100(\%)$

예제 학생 A, B의 공부 습관을 조사하였더니 A학생은 평균 3시간, 표준 편차는 0.5, B 학생은 6시간 표준 편차 0.8인 결과를 얻었다. 어느 학생이 더 꾸준히 공부하는 습관을 가지고 있을까? 계산 결과 B 학생이 더 꾸준히 공부하는 습관을 가지고 있었다.

A 학생 공부시간에 대한 변동 계수 $= 0.5/3 \times 100(\%) = 16.7\ (\%)$
B 학생 공부시간에 대한 변동 계수 $= 0.8/6 \times 100(\%) = 13.3\ (\%)$

표준편차 와 표준오차(standard error): 표준편차 (standard deviation)는 자료의 표준 편차이고 표준오차는 통계량(여기서는 표본평균)의 표준편차이다. 그러므로 표본평균의 표준오차는 $\sigma/\sqrt{n}$ (모집단 표준 편차를 알 경우) 혹은 $s/\sqrt{n}$ (모집단 표준 편차 모를 경우) 이다.

6.6.3 예제 보기

POTTERY.txt 데이터에서 도자기의 칼슘 함유량(CA)과 마그네슘 함유량(MG)의 기초통계량을 구해 보자.

①원하는 기초통계량의 KEY-word를 적으면 된다. 만약 key-word를 지정하지 않으면 mean, std, min, max 4개 기초통계량이 출력된다. 이외에도 첨도($\frac{E(X-\mu)^3}{\sigma^3}$)는 SKEWNESS(정규분포=0), 왜도는 KURTOSIS $\frac{E(X-\mu)^4}{\sigma^4}$ (정규분포=3)를 이용하면 된다. ②원하는 변수를 지정한다.

```
PROC MEANS DATA=POTTERY MEAN STD VAR CV ①
           P5 P10 Q1 MEDIAN Q3 P90 P95 QRANGE RANGE;
  ②VAR MG CA;
RUN;
```

변수	평균값	표준편차	분산	변동계수	제5 백분위수
Mg	3.1415385	2.1797260	4.7512055	69.3840311	0.5600000
Ca	0.1465385	0.1012301	0.0102475	69.0809261	0.0100000

제10 백분위수	제1 사분위수	중간값	제3 사분위수
0.6000000	0.6700000	3.8250000	4.5200000
0.0100000	0.0600000	0.1550000	0.2200000

제90 백분위수	제95 백분위수	사분위 범위	범위
5.6900000	5.9100000	3.8500000	6.7000000
0.2900000	0.3000000	0.1600000	0.3000000

만약 분류형 변수에 따라 보기 원하면 CLASS 문이나 BY문을 이용할 수 있다. BY문은 그 변수에 의해 정렬되어 있어야 하니 CLASS 문을 이용하는 것이 편리하다. 출력 형태만 다를 뿐 출력 결과 내용은 동일하다.

```
PROC MEANS DATA=POTTERY MEAN STD VAR CV
           P5 P10 Q1 MEDIAN Q3 P90 P95 QRANGE RANGE;
    VAR MG CA;
    CLASS SITE;
RUN;
```

Site	관측치수	변수	평균값	표준편차	
A	5	Mg	0.6060000	0.0634823	0.(
		Ca	0.0520000	0.0342053	0.(
C	2	Mg	3.8550000	0.1202082	0.(
		Ca	0.2950000	0.0070711	0.00(
I	5	Mg	0.6740000	0.0320936	0.(
		Ca	0.0260000	0.0260768	0.00(
L	14	Mg	4.8264286	1.0882209	1.
		Ca	0.2021429	0.0582011	0.(

분석 결과를 SAS 데이터에 저장하려면 다음과 같이 OUTPUT 문을 사용하면 된다. 다음은 변수 MG에 대한 기초 통계량의 일부를 저장하는 프로그램이다. ①저장하고자 하는 SAS 데이터 이름을 지정한다 ②key-woard=변수명의 형태로 지정한다. 변수명에 기초통계량이 저장된다. ③저장된 기초통계량을 출력해 보자. ④이 부분은 저장되는 통계량이 아니라 출력 결과에 출력되는 기초통계량을 지정한다.

```
PROC MEANS DATA=POTTERY MEAN STD VAR CV ④
           P5 P10 Q1 MEDIAN Q3 P90 P95 QRANGE RANGE;
    VAR MG;
    OUTPUT OUT=OUT1 MEAN=MG_M STD=MG_STD MEDIAN=MG_MED;
RUN;          ①                  ②

PROC PRINT DATA=OUT1; ③
RUN;
```

만약 두 개의 변수 이상이 한꺼번에 사용되는 경우 OUTPUT 문의 변수명은 다음과 같이 지정한다.

```
PROC MEANS DATA=POTTERY;  이 부분에 아무 통계량도 지정하지 않았으므로 MEAN,STD,MIN,MAX만 출력
    VAR MG CA;
    OUTPUT OUT=OUT1 MEAN=MG_M CA_M  STD=MG_STD CA_STD;
RUN;

PROC PRINT DATA=OUT1;
RUN;
```

Obs	_TYPE_	_FREQ_	MG_M	CA_M	MG_STD	CA_STD
1	0	26	3.14154	0.14654	2.17973	0.10123

그럼 CLASS 문이 사용된다면 SAS 데이터에는 어떻게 저장되어 있을까? _TYPE_ 변수에 의해 구분될 수 있는데 0이면 전체 기초 통계량을 1이면 CLASS에 지정된 변수의 수준에 따라 출력된다.

```
PROC MEANS DATA=POTTERY;
    VAR MG CA;
    CLASS SITE;
    OUTPUT OUT=OUT1 MEAN=MG_M CA_M  STD=MG_STD CA_STD;
RUN;

PROC PRINT DATA=OUT1;
RUN;
```

Site	_TYPE_	_FREQ_	MG_M	CA_M	MG_STD	CA_STD
	0	26	3.14154	0.14654	2.17973	0.10123
A	1	5	0.60600	0.05200	0.06348	0.03421
C	1	2	3.85500	0.29500	0.12021	0.00707
I	1	5	0.67400	0.02600	0.03209	0.02608
L	1	14	4.82643	0.20214	1.08822	0.05820

신뢰구간 구하기

①유의수준을 지정한다. 유의수준은 5%이므로 신뢰수준은 95%가 된다. ②STD는 표준편차(s)이고 STDERR는 표준오차로 표본평균의 편차이다. $s/\sqrt{n}$ ③신뢰구간을 구하는 옵션이다.

```
PROC MEANS DATA=POTTERY ALPHA=0.05 MEAN STD STDERR CLM;
    VAR MG CA;
RUN;
```

(① ALPHA=0.05, ② STD, ③ CLM)

변수	평균값	표준편차	표준오차	평균에 대한95% 신뢰하한	평균에 대한95% 신뢰상한
Mg	3.1415385	2.1797260	0.4274794	2.2611281	4.0219489
Ca	0.1465385	0.1012301	0.0198529	0.1056507	0.1874262

단일 집단 모평균 추론 $H_0 : \mu = \mu_0$

▨검정통계량: $T = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1) \; or \; N(0,1)$

▶소표본인 모집단 정규분포 가정 혹은 비모수 방법(Sign 검정, Wilcoxon Ranks Sum)

▨신뢰구간: $\bar{x} \pm t(n-1;1-\alpha/2)\dfrac{s}{\sqrt{n}}$

연습문제 **A:\JUNE08\MEAN1.doc**

FITNESS.txt에서 산소량(OXYGEN)의 대해 평균과 표준편차를 구하시오.
평균과 표준편차를 저장한 SAS 데이터 OUT1을 만들고 여기에 저장된 값을 이용하여 검정통계량(T), p-값, 95% 신뢰구간(계산 결과가 맞는지 CLM 옵션을 이용하여 확인)을 계산하시오.
귀무가설: 산소량의 평균은 50이다. 대립가설: 50이 아니다.
프로그램, 결과, 그리고 해석(결론)에 대해 워드 문서를 “MEAN1”으로 저장한다.

T, PROBT 옵션 사용하기

PROC MEANS에 검정통계량 T과 P-값(유의확률)을 출력하는 옵션 T, PROBT가 있으나 이는 귀무가설의 모집단 평균이 0인 경우 검정하는 통계량과 p-값이 출력된다.

```
PROC MEANS DATA=FITNESS T PROBT;
    VAR OXYGEN;
RUN;
```

① ②

분석 변수 : Oxygen

t Value	Pr > |t|
49.51	<.0001

그러므로 굳이 PROC MEANS를 사용하여 모집단 평균 검정을 하려면 우선 데이터에서 귀무가설의 모집단 평균 값으로 빼준 후 검정하면 된다.

```
DATA FITNESS2;
    SET FITNESS;
    OXYGENO=OXYGEN-50;
RUN;

PROC MEANS DATA=FITNESS2 T PROBT;
    VAR OXYGENO;
RUN;
```

분석 변수 : OXYGENO

t Value	Pr > |t|
-2.74	0.0102

이런 작업을 한 번에 할 수 있는 PROCEDURE가 있는데 그것이 TTEST이다. ①은 유의수준을 지정한다. 이는 신뢰수준을 정해주기 위함이다. 95% 신뢰구간 ②귀무가설의 모집단 평균을 지정한다.

```
PROC TTEST DATA=FITNESS ALPHA=0.05 HO=50;
    VAR OXYGEN;
RUN;
```

① ②

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
Oxygen	31	45.422	47.376	49.33	4.2571	5.3272	7.1208	0.9568

t-검정통계량은 위와 동일하다.

T-Tests

Variable	DF	t Value	Pr > |t|
Oxygen	30	-2.74	0.0102