CHAPTER 4.

SAS 함수

SAS는 변수에 대한 함수 값 계산이나 계산에 필요한 함수가 내장되어 있다. 수학적 계산을 위한 절대값, 제곱근과 승(power), 로그, 지수 함수, 통계 계산을 위한 평균, 분산, CV 등 많은 함수들이 있다. 이 함수를 사용하는 방법은 다음과 같다. 수식처럼 오른쪽 함수 결과가 왼쪽 변수에 저장된다. 함수는 변수의 각 관측치에 적용되므로 결과는 변수의 관측치 수만큼 계산된다. 즉 함수의 계산은 데이터에서 행으로 이루어진다.

•변수 이름 = 함수 이름(변수, 다른 함수, 숫자 등)

`Y=LOG(X);` 변수 X의 자연 로그(natural LOG) 값이 변수 Y에 저장

•변수 이름 = 함수 이름(변수1, 변수2, …)

`Y=SUM(X, Z, W);` 변수 X, Z, W의 합이 변수 Y에 저장

•변수 이름 = 함수 이름(of 연속된 변수 이름)

`Y=MEAN(OF X1-X10);` 변수 X1, X2, …, X10 10개 변수의 평균이 변수 Y에 저장

PROC MEAMS 사용하면 변수의 평균, 표준편차 등을 얻을 수 있는데 이는 MEAN 함수와 달리 설정된 변수에 대해 통계 값이 출력되게 된다. PROC 단계는 데이터의 열(변수)에 대한 계산 결과를 얻는다.

4.1 맛보기

다음 프로그램을 실행해 결과를 보자. 각 함수의 의미는 다음과 같다.

⑴EXP 함수는 () 안의 변수 관측치의 지수 값을 구하는 함수이다.

⑵RANNOR(seed)는 평균이 0이고 표준편차가 1인 정규분포함수를 따르는 관측치를 생성하는 함수이다. SEED(시드)는 값을 생성할 때 시작하는 위치를 나타내는 값으로 1~(2^{31}-1) 사이의 정수 값이나 0을 사용할 수 있다. 0을 사용하면 프로그램 실행되는 시각이 시드 값으로 설정된다. 변수 ENG는 평균이 90이고 표준편차가 10인 정규분포를 따르며, 변수 MATH는 평균이 80, 표준편차가 10인 정규분포를 따르는 분포에서 얻는(생성, generating) 관측치 변수이다.

⑶MEAN 함수는 ()에 지정된 변수들의 평균을 내는 함수이다.

확장 편집기 - 제목없음1 *

```
DATA FN;
     DO ID=1 TO 10;
          Y=EXP(ID);
          ENG=90+RANNOR(1)*10;
          MATH=80+RANNOR(1)*10;
          AVG=MEAN(ENG,MATH);
          OUTPUT;
     END;
RUN;

PROC PRINT DATA=FN;
RUN;
```

ID	Y	ENG	MATH	AVG
1	2.72	108.048	79.2008	93.6245
2	7.39	93.966	69.1668	81.5663
3	20.09	112.383	73.7577	93.0703
4	54.60	95.137	79.1339	87.1352
5	148.41	84.058	80.3189	82.1886
6	403.43	82.622	77.4986	80.0603
7	1096.63	96.850	71.9584	84.4042
8	2980.96	82.557	72.0450	77.3011
9	8103.08	93.407	76.9949	85.2010
10	22026.47	76.502	84.3270	80.4143

4.2 함수(1)

다음은 수학, 통계, 연산에 관련된 함수를 예제와 함께 정리한 것이다. 각 변수에 대한 기초 통계량 계산 혹은 변수들간의 관계를 보려면 적절한 PROC을 사용하며 함수, 수식에 의해 변수 생성(변수명=함수이름(argument))을 할 수 있다. **argument** 부분에는 변수, 숫자를 사용할 수 있다. V, A, B, ..., X 등은 변수 이름이다.

수학 계산 관련 함수

함수 형태	내용	예제
ARCOS(argument)	COS의 inverse 값을 계산 −1≤argument≤1	**V=arcos(a);** **V=arcos(0.3);**
COS(argument)	COS 값을 계산 argument은 실수나 Radian값	**V=cos(a);** **V=cos(3.14159/3);**
ARSIN(argument)	SIN의 inverse 값을 계산 argument은 실수나 Radian값	**V=arsin(a);** **V=arsin(0.3);**
SIN(argument)	SIN 값을 계산 −1≤argument≤1	**V=sin(a);** **V=sin(3*3.14159);**
TAN(argument)	TAN 값을 계산, −1≤argument≤1	**V=tan(a);** **V=tan(2);**
EXP(변수명**)**	지수함수로 지수 값을 계산한다.	**V=EXP(X);** **V=EXP(3.2);**
SQRT(변수명**)**	제곱근 값을 계산한다. () 안의 수는 0보다 커야 한다. 제곱은 **x**2**, 세제곱근 **x**3**	**V=SQRT(X);** **V=SQRT(3.2);**
LOG(변수명**)**	자연 로그(natural log) 값을 계산한다. $Log_e^X = Ln(X)$	**V=LOG(X);** **V=LOG(3.2);**
LOGN(변수명**)**	로그의 밑이 n인 로그 값을 계산한다. N=10 이면 상용로그 값	**V=LOG(X);** **V=LOG(3.2);** **V=LOG10(3.2);**

EXAMPLE: 수학 함수 사용 예제

FN.sas *

```
DATA FN;
     INPUT X Y Z;
     CARDS;
     -1.3 1   11  27
     4.3  2   12  .
     -7   3   13  26
     5    4   .   25
     2.7  .   15  24
     -2.7 10  16  23
     0    100 17  22
RUN;
```

FN.sas *

```
DATA FN1;
     SET FN;
     C=COS(Y);
     T=TAN(Y);
     EXP=EXP(Y);
     SQ=SQRT(Y);
     LN=LOG(Y);
     LOG=LOG10(Y);
     DROP X Z;
RUN;

PROC PRINT DATA=FN1;
RUN;
```

Obs	Y	C	T	EXP	SQ	LN	LOG
1	1	0.54030	1.55741	2.72	1.0000	0.00000	0.00000
2	2	-0.41615	-2.18504	7.39	1.4142	0.69315	0.30103
3	3	-0.98999	-0.14255	20.09	1.7321	1.09861	0.47712
4	4	-0.65364	1.15782	54.60	2.0000	1.38629	0.60206
5	.	.	.	.	.	.	.
6	10	-0.83907	0.64836	22026.47	3.1623	2.30259	1.00000
7	100	0.86232	-0.58721	2.6881E43	10.0000	4.60517	2.00000

정수 및 절대값 얻기 관련 함수

함수 형태	내용	예제
ABS(argument)	절대값을 계산	**V=abs(-2.4);** **V=abs(a);**
CEIL(argument)	argument 값 이상이면서 가장 작은 정수	**V=ceil(a);** **V=ceil(-2.4);**
INT(argument)	정수 값을 출력한다.	**V=int(a);**
FLOOR(arguments)	argument 값 이하이면서 가장 큰 정수	**V=floor(a);**

SIGN(argument)	값의 부호를 출력한다. 양수면 1, 음수면 -1, 0이면 0의 값이 저장된다.	**V=sign(a);** **V=sign(name-90);**
MOD(숫자1,숫자2)	숫자1을 숫자2로 나눈 나머지 계산.	**V=mod(14,3);**

EXAMPLE: 수학 함수 사용 예제

FN.sas *

```
DATA FN2;
     SET FN;
     V1=ABS(X);
     V2=CEIL(X);
     V3=INT(X);
     V4=FLOOR(X);
     V5=SIGN(X);
     DROP Y Z;
RUN;

PROC PRINT DATA=FN2;
RUN;
```

Obs	X	V1	V2	V3	V4	V5
1	-1.3	1.3	-1	-1	-2	-1
2	4.3	4.3	5	4	4	1
3	-7.0	7.0	-7	-7	-7	-1
4	5.0	5.0	5	5	5	1
5	2.7	2.7	3	2	2	1
6	-2.7	2.7	-2	-2	-3	-1
7	0.0	0.0	0	0	0	0

통계 계산 관련 함수

함수 형태	내용	예제
FACT(n)	Factorial 값을 구한다. ()안은 반드시 정수 값이어야 한다. $n!$	**V=FACT(a);** **V=FACT(6);**
COMB(n,r)	combination 값을 구한다. n, r은 정수이어야 한다. $nCr = \frac{n!}{r!(n-r)!}$	**V=COMB(5,2);** **V=COMB(a,b);**
PERM(n,r)	Permutation 값을 구한다. n, r은 정수이어야 한다. $n\Pr = \frac{n!}{(n-r)!}$	**V=PERM(5,2);** **V=PERM(a,b);**

MAX(arguments)	최대값을 구한다. **V=max(x1,x2,x3,x4);** 연속일 때 **V=max(of x1-x4);** 이 형식은 아래 함수에도 적용	**V=max(a,b,c);** **V=max(1,5,-1,7);**
MIN(arguments)	변수 관측치 중 최소값을 구한다.	**V=min(a,b,c);**
N(arguments)	변수 관측치 개수(결측치 제외) 계산	**V=n(a,b,c);**
SUM(arguments)	변수 관측치들의 합을 구한다.	**V=sum(x,y,z);**
MEAN(arguments)	변수 관측치들의 평균을 구한다.	**V=mean(x,y,z);**
RANGE(arguments)	변수 관측치들의 범위를 구한다.	**V=range(x,y,z);**
STD(arguments)	변수 관측치들의 표준편차를 구한다.	**V=std(x,y,z);**
STDERR(arguments)	변수 관측치들의 표준오차를 구한다.	**V=stderr(x,y,z);**
VAR(arguments)	변수 관측치들의 분산을 구한다.	**V=var(x,y,z);**

[참고] 변동계수와 표준오차

▪변동계수: $CV = \frac{s}{\bar{x}} \times 100(\%)$ 집단 간 분산을 비교하기 위하여 사용되는 통계량

▪표준오차: $s/\sqrt{n}$ 표준편차를 표본 개수의 제곱근으로 나눈 값으로 표본평균의 표준편차

FN.sas *

```
DATA FN3;
     SET FN;
     FAC=FACT(Z);
     N=N(X,Y,Z);
     MAX=MAX(X,Y,Z);
     MN=MEAN(X,Y,Z);
     SUM=SUM(X,Y,Z);
     STD=STD(X,Y,Z);
     STDERR=STDERR(X,Y,Z);
     CV=CV(X,Y,Z);
RUN;

PROC PRINT DATA=FN3;
RUN;
```

X	Y	Z	FAC	N	MAX	MN	SUM	STD	STDERR	CV
-1.3	1	11	39916800	3	11	3.5667	10.7	6.5394	3.7755	183.347
4.3	2	12	479001600	3	12	6.1000	18.3	5.2374	3.0238	85.858
-7.0	3	13	6227020800	3	13	3.0000	9.0	10.0000	5.7735	333.333
5.0	4	.	.	2	5	4.5000	9.0	0.7071	0.5000	15.713
2.7	.	15	1.30767E12	2	15	8.8500	17.7	8.6974	6.1500	98.276
-2.7	10	16	2.09228E13	3	16	7.7667	23.3	9.5479	5.5125	122.935
0.0	100	17	3.55687E14	3	100	39.0000	117.0	53.5070	30.8923	137.197

[참고] E13의 의미는 $\times 10^{13}$ 이다. 그러므로 3.55E14=355,000,000,000,000이다.

차분(difference) 관련 함수

함수 형태	내용	예제
LAG(변수명**)**	이전 관측치를 가져온다.	**V=LAG(X);**
LAGN(변수명**)**	n번째 이전 관측치를 가져온다.	**V=LAG2(X);**
V=DIF(변수명**)**	현재 관측치와 이전 관측치의 차이를 구한다. 그러므로 **DIF=X-LAG(X);**이다.	**V=DIF(X);**
V=DIFN(변수명**)**	현재 관측치와 이전 관측치의 차이를 구한다. 그러므로 **DIFN=X-LAGN(X);**이다.	**V=DIF2(X);**

FN.sas *

```
DATA FN4;
     SET FN;
     LAG1=LAG(X);
     LAG3=LAG3(X);
     DIF1=DIF1(X);
     DIF=X-LAG(X);
     DROP Y Z;
RUN;

PROC PRINT DATA=FN4;
RUN;
```

Obs	X	LAG1	LAG3	DIF1	DIF
1	-1.3	.	.	.	.
2	4.3	-1.3	.	5.6	5.6
3	-7.0	4.3	.	-11.3	-11.3
4	5.0	-7.0	-1.3	12.0	12.0
5	2.7	5.0	4.3	-2.3	-2.3
6	-2.7	2.7	-7.0	-5.4	-5.4
7	0.0	-2.7	5.0	2.7	2.7

관측치 변환 함수

함수 형태	내용	예제
LENGTH(변수명);	문자 변수 문자열(string) 길이 출력한다.	**V=length(z);** **V=length('se');**
LEFT(변수명);	문자열의 왼쪽 정렬한다.	**V=left(z);** **V=left(' se');**
RIGHT(변수명);	문자열의 오른쪽 정렬한다.	**V=right(z);** **V=right(' se ');**
SUBSTR(argument, 시작, 길이**);**	문자열 관측치의 일부를 얻는데 사용된다. 시작은 문자열 시작 위치, 길이는 문자 개수를 지정한다.	**V=substr(z,1,3);** **V=substr('hi ',1,2);**
TRIM(변수명);	문자열 뒤쪽 공백 없앤다.	**V=trim(z);** **V=trim(' hi ');**
UPCASE(변수명);	문자열을 대문자로 변환한다.	**V=upcase(z);** **V=upcase ('hi');**

FN.sas *

```
DATA FNO;
     INPUT X $ 1-12;
     L=LEFT(X);   R=RIGHT(X); SR=SUBSTR(X,3,5);
     TR=TRIM(X); U=UPCASE(X);
     LN1=LENGTH(X);   LN2=LENGTH(R);
     LN3=LENGTH(SR); LN4=LENGTH(TR);
     CARDS;
Think Global
Act Local
Wolfpack
   ncsu
RUN;
PROC PRINT DATA=FNO;
RUN;
```

문자 변수 지정하는 $ 뒤에 1-12의 의미는 첫 열부터 12번째 열까지 변수 X의 관측치로 읽어 들이라는 의미이다. 프로그램에서 3번째 행부터는 프로그램이 길어지는 것을 피하기

위하여 두 개 문장을 한 라인에 적었다. 한 라인에 적었더라도 세미콜론으로 문장의 끝을 알렸으므로 각각 다른 문장으로 실행된다.

X	L	R	SR	TR	U
Think Global Act Local Wolfpack ncsu	Think Global Act Local Wolfpack ncsu	Think Global Act Local Wolfpack ncsu	ink G t Loc lfpac su	Think Global Act Local Wolfpack ncsu	THINK GLOBAL ACT LOCAL WOLFPACK NCSU

LN1	LN2	LN3	LN4
12	12	5	12
9	12	5	9
8	12	5	8
4	12	2	4

EXAMPLE: 함수 사용하기

🖫IQ.TXT 데이터를 SAS 데이터 ONE으로 만들고, IQ1, IQ2, IQ3 중 가장 높은 값을 IQ_MAX, 가장 낮은 값을 IQ_MIN으로 만들어보자. 그리고 평균은 IQ_MEAN으로 하고, 함수를 사용하여 자료를 만들어보자. SAS 데이터 이름은 TWO로 하자.

확장 편집기 - 제목없음1 *

```
DATA ONE;
     INFILE 'C:/TEMP/IQ.TXT';
     INPUT ID IQ1 IQ2 IQ3 GENDER $ SIZE;
RUN;

DATA TWO;
     SET ONE;
     IQ_MAX=MAX(IQ1,IQ2,IQ3);
     IQ_MIN=MIN(IQ1,IQ2,IQ3);
     IQ_MEAN=MEAN(IQ1,IQ2,IQ3);
RUN;

PROC PRINT DATA=TWO;
RUN;
```

Obs	ID	IQ1	IQ2	IQ3	GENDER	SIZE	IQ_MAX	IQ_MIN	IQ_MEAN
1	1	133	132	124	Female	816932	133	124	129.667
2	2	140	150	124	Male	1001121	150	124	138.000
3	3	139	123	150	Male	1038437	150	123	137.333
4	4	133	129	128	Male	965353	133	128	130.000
5	5	137	132	134	Female	951545	137	132	134.333
6	6	99	90	110	Female	928799	110	90	99.667
7	7	138	136	131	Female	991305	138	131	135.000
8	8	92	90	98	Female	854258	98	90	93.333
9	9	89	93	84	Male	904959	93	84	88.667

EXAMPLE: 함수 사용하기(2)

SAS 데이터 TWO에서 남자(여자)는 IQ_MEAN가 100(110)보다 크면 1, 작으면 -1, 같으면 0이 되는 변수 W를 만들고, SAS 데이터 이름은 THREE로 하자.

확장 편집기 - 제목없음1 *

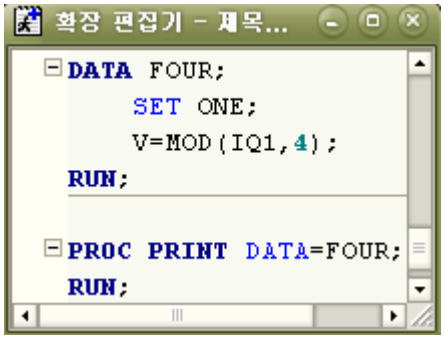
```
DATA THREE;
     SET TWO;
     W=-1;
     IF (GENDER='Male' AND IQ_MEAN>100) THEN W=1;
     IF (GENDER='Male' AND IQ_MEAN=100) THEN W=0;
     IF (GENDER='Female' AND IQ_MEAN>110) THEN W=1;
     IF (GENDER='Female' AND IQ_MEAN=110) THEN W=0;
RUN;

PROC PRINT DATA=THREE;
RUN;
```

Obs	ID	IQ1	IQ2	IQ3	GENDER	SIZE	IQ_MAX	IQ_MIN	IQ_MEAN	W
1	1	133	132	124	Female	816932	133	124	129.667	1
2	2	140	150	124	Male	1001121	150	124	138.000	1
3	3	139	123	150	Male	1038437	150	123	137.333	1
4	4	133	129	128	Male	965353	133	128	130.000	1
5	5	137	132	134	Female	951545	137	132	134.333	1
6	6	99	90	110	Female	928799	110	90	99.667	-1

EXAMPLE: 함수 사용하기(3)

SAS 데이터 ONE에서 IQ1이 4의 배수이면 0 그렇지 않으면 나머지가 되는 변수 V를 만들고, SAS 데이터 이름은 FOUR로 하자.



SAS 시스템 11

Obs	ID	IQ1	IQ2	IQ3	GENDER	SIZE	V
1	1	133	132	124	Female	816932	1
2	2	140	150	124	Male	1001121	0
3	3	139	123	150	Male	1038437	3
4	4	133	129	128	Male	965353	1
5	5	137	132	134	Female	951545	1
6	6	99	90	110	Female	928799	3
7	7	138	136	131	Female	991305	2
8	8	92	90	98	Female	854258	0
9	9	89	93	84	Male	904858	1

EXAMPLE: 함수 사용하기(4) 확률분포함수 그리기

평균이 0이고 분산이 2인 정규분포 확률분포함수(probability density function)를 그려보자.

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad -\infty<x<\infty$$

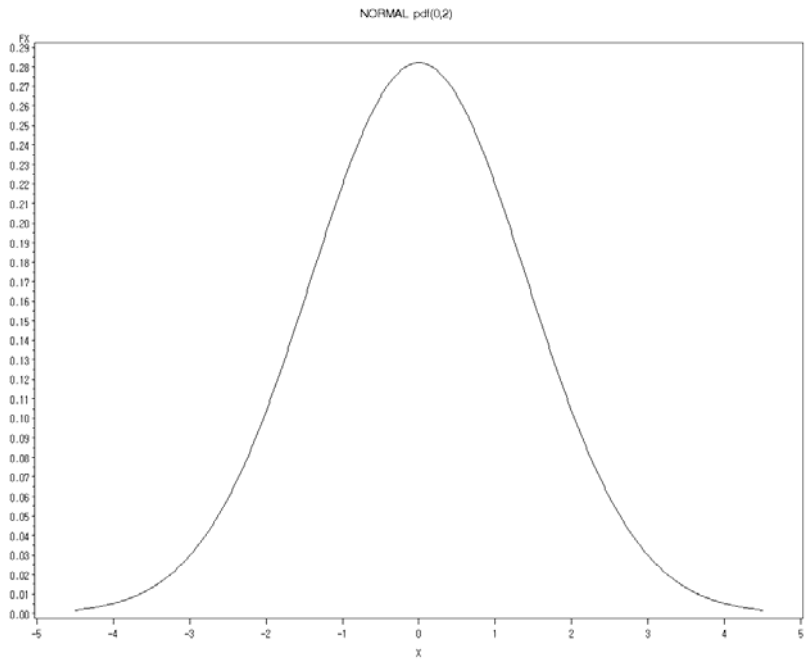
확률분포함수의 X-축은 확률변수(X)가 가질 수 있는 값이며 Y-축은 확률($f(x)$)이다. 이론적으로는 정규분포를 따르는 확률변수가 $(-\infty,\infty)$ 사이지만 실제 데이터는 $\mu\pm3\sigma$ 안에 대부분 다 포함된다. (경험적 규칙: empirical rule) 여기서 $\mu=2,\sigma=\sqrt{2}$ 이므로 (-4.24, 4.24)이다. 그래서 그래프를 위한 최소값 -4.5, 최대값 4.5을 사용하였다.

DO문은 X-축의 값을 설정하기 사용하였고 증가분 BY는 그래프를 smooth하게 그리기 위하여 0.01로 작게 잡았다. 그래서 X=-4.5, -4.49, -2.48, ..., 4.49, 4.5 이런 값들이 연속적으로

사용된다. 그러므로 관측치 개수는 901개이다. X-축 구간을 매우 작게 잡았기 때문에 PROC GPLOT에서 I(interpolation)옵션을 JOIN(점들을 연결)으로 사용해도 그래프가 smooth 하다. 만약 X-축의 구간을 0.5로 잡았으면 I=SPLINE을 사용하면 그래프가 곡선화 된다. 컴퓨터 속도의 발달로 관측치가 많아도 실행되므로 구간을 작게 설정해도 된다.

FN1.sas * PROC GPLOT 실행 중

```
DATA FN1;
    DO X=-4.5 TO 4.5 BY 0.01;
        FX=1/(SQRT(2*3.141592654*2))*EXP(-(X)**2/(2*2));
        OUTPUT;
    END;
RUN;
TITLE H=1.2 F=SWISS 'NORMAL pdf(0,2)';
PROC GPLOT DATA=FN1;
    SYMBOL I=JOIN V=NONE;
    PLOT FX*X;
RUN;
```



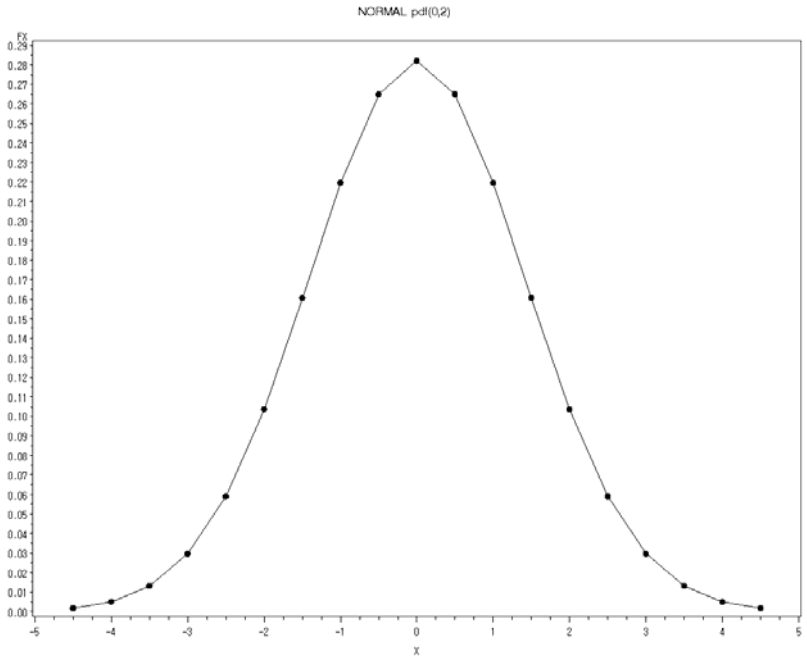
TITLE문은 그래프의 제목을 붙이기 위해 사용되었으며 H는 글자 크기(height)를 설정한 것이고 F(font) 옵션은 글꼴을 설정한 옵션이다. PROC GPLOT에서 SYMBOL 문장은 그래프의 점들을 정의하는 것으로 I는 점들을 연결하는 옵션이고 V(value)는 점들의 형태를 설정

(triangle, dot, square, ‘a’, ‘o’ 등을 사용할 수 있다)하는 옵션이다. 자세한 내용은 5장을 참고하기 바란다.

만약 위 프로그램에서 0.01대신 0.5를 사용하였다면 어떻게 될까? 점들이 너무 떨어져 있어 직선 연결하면 smooth하지 못하다. 이런 경우 I=SPLINE을 사용하면 이런 문제는 해결된다.

FN1.sas * PROC GPLOT 실행 중

```
DATA FN1;
    DO X=-4.5 TO 4.5 BY 0.5;
        FX=1/(SQRT(2*3.141592654*2))*EXP(-(X)**2/(2*2));
        OUTPUT;
    END;
RUN;
TITLE H=1.2 F=SWISS 'NORMAL pdf(0,2)';
PROC GPLOT DATA=FN1;
    SYMBOL I=JOIN V=DOT;
    PLOT FX*X;
RUN;
```



EXAMPLE: 함수 사용하기(4)

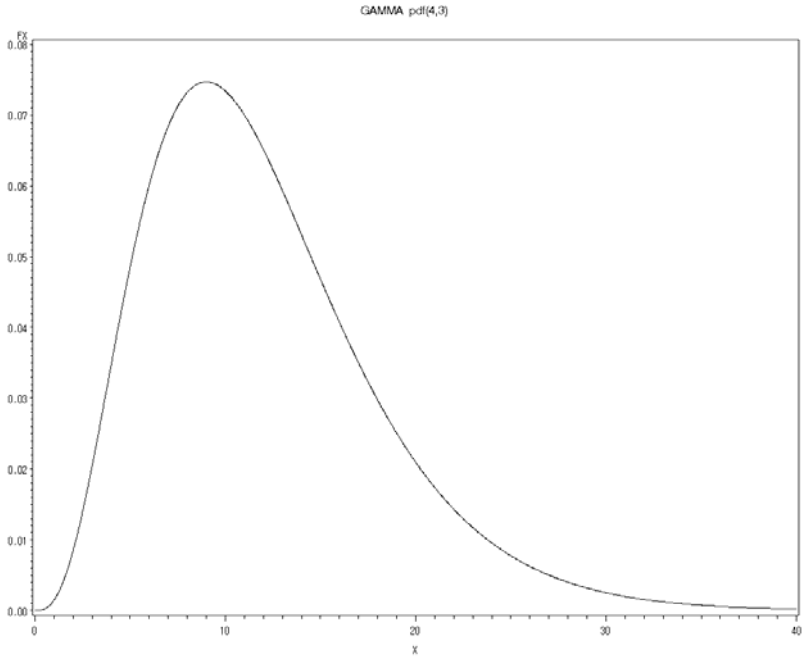
모수(parameter) $(\alpha, \beta) = (4,3)$ 인 감마분포의 확률밀도함수를 그려보자.

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

모수는 확률분포함수의 형태를 결정하는 것으로 모수 값을 알면 확률밀도함수를 그릴 수 있다. 정규 확률분포함수의 모수는 평균(μ)과 표준편차(σ)이다. 감마분포는 우측으로 치우친(skewed to the right, positive skewed) 형태이므로 평균 $\alpha\beta = 12$, 표준편차 $\sqrt{\alpha\beta^2} = 6$ 이나 최대값을 30대신 40을 사용하였다.

FN1.sas * PROC GPLOT 실행 중

```
DATA FN1;
     AL=4;BE=3;
     DO X=0 TO 40 BY 0.01;
          FX=1/(GAMMA(AL)*(BE**AL))*(X**(AL-1)*EXP(-X/BE));
          OUTPUT;
     END;
RUN;
TITLE H=1.2 F=SWISS 'GAMMA pdf(4,3)';
PROC GPLOT DATA=FN1;
     SYMBOL I=JOIN V=NONE;
     PLOT FX*X;
RUN;
```



EXAMPLE: 함수 사용하기(5) 이항분포 그리기

이산형 확률밀도함수인 이항분포를 함수를 이용하여 그려보자. 모수가 $(n=20,\ p=0.2)$ 인 이항분포 확률분포함수는 다음과 같다.

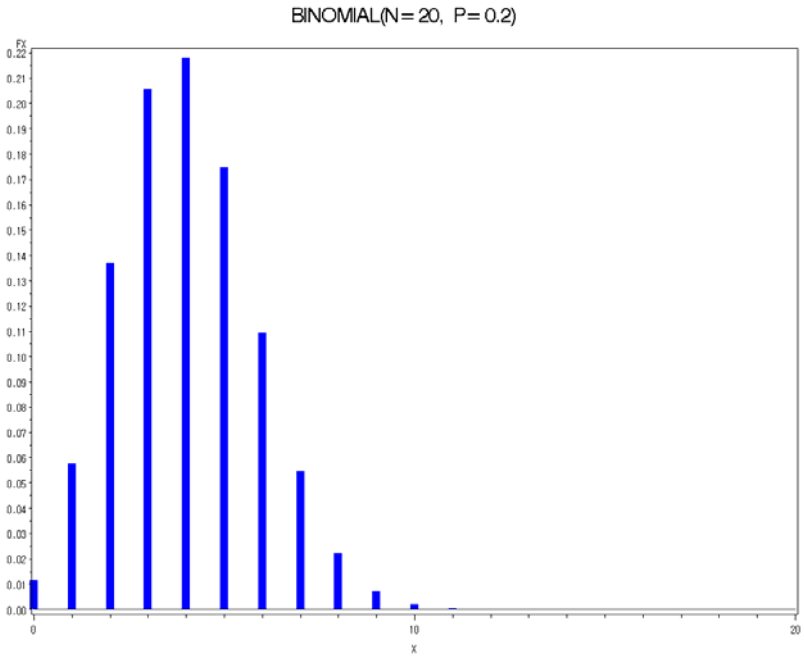
$$p(x)=\binom{n}{x}p^{x}q^{n-x}=\binom{20}{x}0.2^{x}0.8^{20-x},\ x=0,1,2,...,20$$

Combination 계산을 위하여 COMB 함수를 사용하였다. 모수 $(n=20,\ p=0.2)$ 인 이항분포를 따르는 이상형 확률변수가 가질 수 있는 값이 0, 1, …, 20이므로 이를 DO문에 초기값과 말기 값으로 설정하였다. 증가분은 디폴트인 1(BY 1)을 사용하였다. 이산형의 확률은 막대 높이로 표시되므로 I 옵션에는 NEEDLE 옵션을 사용하였다. CI는 막대 안의 색깔을 지정하는 옵션이고 W(weight)는 막대의 넓이를 설정하는 옵션이다.

FN1.sas * PROC GPLOT 실행 중

```
DATA FN3;
    DO X=0 TO 20;
        FX=COMB(20,X)*(0.2)**X*(0.8)**(20-X);
        OUTPUT;
    END;
RUN;

PROC GPLOT DATA=FN3;
    SYMBOL I=NEEDLE V=NONE CI=BLUE W=10;
    PLOT FX*X;
RUN;
```

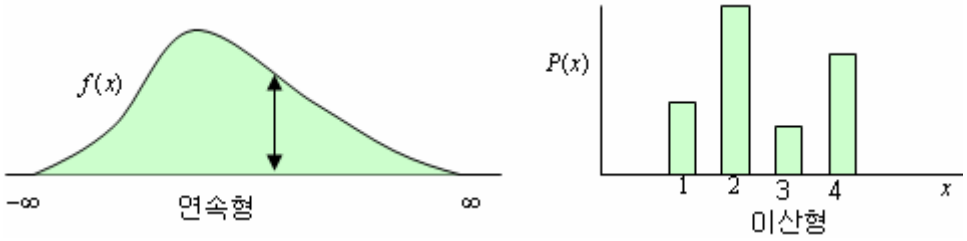


4.3 함수2

확률밀도함수의 확률 값을 얻거나 그래프를 그리기 위하여 4.1절의 방법을 사용할 수 있으나 함수 형태를 적어 주는 것이 번거롭다. 그래서 통계소프트웨어인 SAS는 통계학에서 가장 많이 사용하는 확률분포함수(PDF), 누적확률분포함수(CDF), 백분위(percentile), 그리고 임의의 분포를 따르는 확률변수를 생성(generating)하는 방법을 함수화 하여 내장하고 있다.

4.3.1 확률밀도함수

확률밀도함수(PDF, probability density function)는 input(X-축)은 확률변수가 가지는 값이고, output($f(x)$, Y-축)은 확률인 함수이다. 그리고 확률이 갖추어야 할 조건 (1)확률 $f(x)$ 은 항상 0보다 크고 (2)확률변수 전 영역의 $f(x)$ 적분 값은 1이다. (모든 확률을 더하면 1이다)



이처럼 X-축은 확률변수이고 Y-축은 확률 $f(x)$ (이산형일 경우에는 $p(x)$)는 확률밀도함수이다. 화살표의 높이는 확률이다. 이산형인 경우 확률변수 한 값에 대해 확률이 존재하나 연속형인 경우는 0이다. 이산형 확률밀도함수인 경우 확률은 높이이고 연속형 확률밀도함수의 확률은 면적이다. 확률이므로 짙은 부분 면적의 합(전체 확률의 합)은 1이다. 확률변수 X 가 가지는 값의 범위는 (a,b) 이다.

SAS에서 확률밀도함수를 얻는 함수는 PDF이다.

PDF('분포이름', x, arguments)

X는 확률변수가 가질 수 있는 영역(구간) 안에 임의의 값이고 argument는 확률변수의 모수들이다. 모수는 확률변수함수 $f(x)$ 을 결정하는 값으로 정규분포확률밀도함수는 평균과 μ , 분산 σ^2 이다. 지수분포의 모수는 평균인 β 이다.

이산형 확률밀도함수

분포	확률분포함수	모수	SAS 함수
베르누이 분포 (Bernoulli)	$p(x) = p^x q^{1-x}$ $x = 0,1$ 평균: p 분산: pq	p	**PDF('BERNOULLI',x,p)**

이항분포 (Binomial)	$p(x)=\binom{n}{x}p^x q^{n-x}$ $x=0,1,2,...,n$ 평균: np 분산: npq	n, p	**PDF('BINOMIAL',x,p,n)**
기하분포 (Geometric)	$p(x)=q^{x-1}p$ $x=1,2,...$ 평균: q/p 분산: q/p^2	p	**PDF('GEOMETRIC',x,p)**
음이항 분포 (Negative binomial)	$p(x)=\binom{x-1}{r-1}p^{r-1}q^{x-r}p$ $x=r,r+1,...$ 평균: rq/p 분산: rq/p^2	r, p	**PDF('NEGB',x,p,r)**
포아송 분포 (Poisson)	$p(x)=\frac{e^{-\lambda}\lambda^x}{x!}$ $x=0,1,2...$ 평균: λ 분산: λ	λ	**PDF('POISSON',x,λ)**
초기하 분포 (Hyper-geometric)	$p(x)=\frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}}$ $M=1,2...,K=0,1,2,...,M$ $n=1,2,...,M$ 평균: $n(\frac{K}{M})$ 분산: 복잡	(M,K,n)	**PDF('HYPER',x,M,K,n)**

연속형 확률밀도함수

분포	확률분포함수	모수	SAS 함수
정규분포 (Normal)	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $-\infty<x<\infty$ 평균: μ, 분산: σ^2	μ,σ	**PDF('NORMAL',x,μ,σ)**

감마분포 (Gamma)	$f(x)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-x/\beta}$ $0<x<\infty$ 평균: $\alpha\beta$, 분산: $\alpha\beta^2$	α,β	**PDF('GAMMA',x,** α **,** β **)**
베타분포 (Beta)	$f(x)=\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}\,0<x<\infty$ 평균: $a/(a+b)$ 분산: $\frac{ab}{(a+b)^2(a+b+1)}$	a, b	**PDF('BETA',x,a,b)**
지수분포 (Exponential)	$\alpha=1$인 감마분포. $f(x)=\frac{1}{\beta}e^{-x/\beta}$ 평균: β, 분산: β^2	β	**PDF('EXPO',x,** β **)**
카이제곱 분포 (Chi- squared)	$\alpha=r/2$, $\beta=2$ 감마분포. $f(x)=\frac{1}{\Gamma(r/2)2^{r/2}}x^{r/2-1}e^{-x/2}$ 평균: r, 분산: $r/2$	자유도 r	**PDF('CHISQ',x,r)**
T-분포	$T=W/\sqrt{V/r}$, $0<x<\infty$ $f(x)$ 복잡, 자유도 r $W\sim Normal(0,1), V\sim\chi^2(r)$ 평균: 0, 분산: $r/(r-2)$	자유도 r	**PDF('T',x,**자유도**)**
F-분포	$F=\frac{U/r_1}{V/r_2}$, $0<x<\infty$ $f(x)$ 복잡, 자유도 (r_1,r_2) $W\sim V\sim\chi^2(r_1), V\sim\chi^2(r_2)$ 평균: $n/(n-2)$, 분산: 복잡	분자, 분모 자유도 (r_1,r_2)	**PDF('F',x,**분자 자유도**,**분모 자유도**)**
균일분포 (Uniform)	$f(x)=\frac{1}{b-a}, a\le x\le b$ 평균: $(a+b)/2$, 분산: $(b-a)^2/12$	영역 최소값, 최대값	**PDF('UNIFORM',x,a,b)**

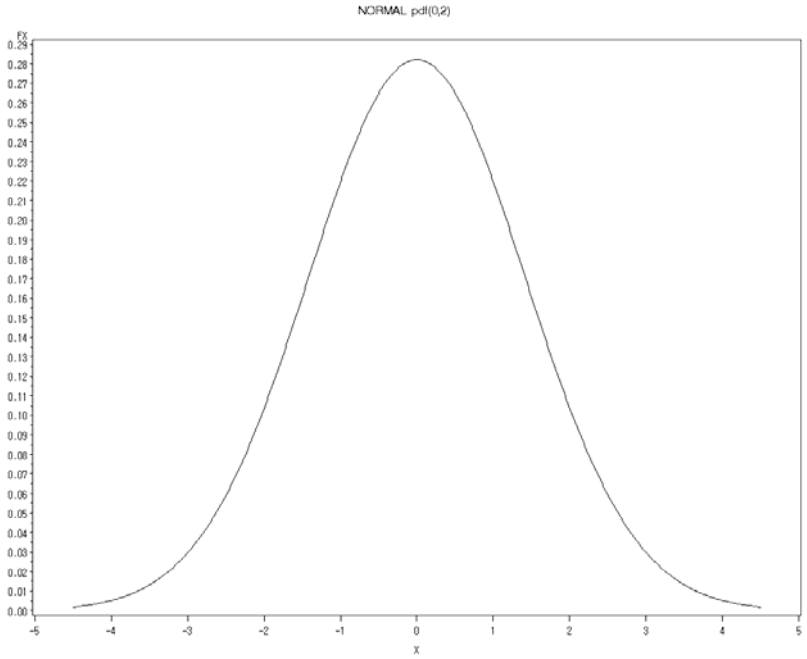
EXAMPLE: PDF 함수 사용하기, 정규분포(평균=0, 분산=2)

평균이 0이고 분산이 2인 정규분포 확률분포함수(probability density function)를 그려보자. 앞에서는 함수 식을 이용하였지만 이제 PDF 함수를 이용해 보자.

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad -\infty<x<\infty$$

FN1.sas * PROC GPLOT 실행 중

```
DATA FN1;
     DO X=-4.5 TO 4.5 BY 0.01;
          FX=PDF('NORMAL',X,0,SQRT(2));
          OUTPUT;
     END;
RUN;
TITLE H=1.2 F=SWISS 'NORMAL pdf(0,2)';
PROC GPLOT DATA=FN1;
     SYMBOL I=JOIN V=NONE;
     PLOT FX*X;
RUN;
```



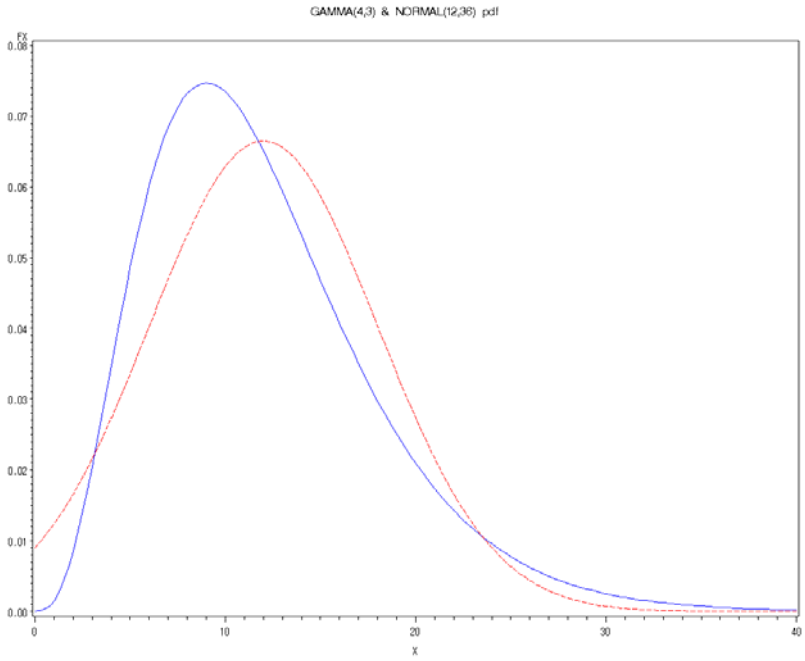
EXAMPLE: PDF 함수 사용하기(2)

모수(parameter) $(\alpha, \beta) = (4,3)$ 인 감마분포의 확률밀도함수와 평균이 12이고 표준편차가 6인 정규분포의 확률밀도함수 그래프를 하나의 그래프에 그려보자. 정규분포는 좌우 대칭인 반면 감마 분포는 우측으로 치우친 형태를 갖는다. 지금 그리려는 두 분포의 평균과 분산은 동일하다.

SYMBOL1, SYMBOL2는 그려지는 그래프에 대한 설정이다. C(color)는 선의 색을 L은 라인의 속성을 나타내는 것으로 L=1(디폴트)은 연속, 값이 커질수록 선이 많이 끊어진다. X-축의 값을 1단위로 얻었으므로 SPLINE을 사용하여 선들을 매끄럽게(smooth) 하였다. OVERLAY 옵션은 두 개 이상의 그래프를 하나의 산점도에 그리라는 옵션이다.

FN1.sas * PROC GPLOT 실행 중

```
DATA FN2;
    DO X=0 TO 40;
        FX=PDF('GAMMA',X,4,3);
        FX1=PDF('NORMAL',X,12,6);
        OUTPUT;
    END;
RUN;
TITLE H=1.2 F=SWISS 'GAMMA(4,3) & NORMAL(12,36) pdf';
PROC GPLOT DATA=FN2;
    SYMBOL1 I=SPLINE V=NONE C=BLUE;
    SYMBOL2 I=SPLINE V=NONE C=RED L=3;
    PLOT (FX FX1)*X/OVERLAY;
RUN;
```



EXAMPLE: PDF 함수 사용하기(3) 이항분포 그리기

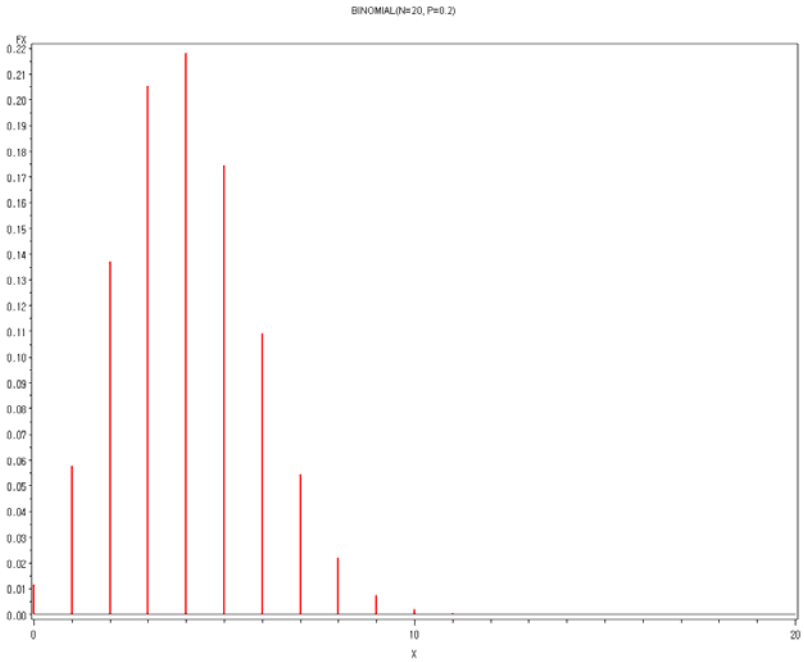
이항분포 $(n=20, p=0.2)$ 인 확률분포함수를 PDF 함수를 이용하여 그려보자.

$$P(x)=\binom{20}{x}0.2^{x}0.8^{20-x},\quad x=0,1,2,...,20$$

막대의 색깔은 빨간, 막대의 넓이는 2로 설정하였다. 막대 높이(확률)는 이전 것과 동일하다. 그래프에 관련된 모든 옵션을 reset하려면 GOPTIONS; 문장을 한 번 사용해 주면 된다. 제목을 RESET하는 TITLE; 문장과 사용방법은 동일하다.

FN1.sas * PROC GPLOT 실행 중

```
DATA FN4;
    DO X=0 TO 20;
        FX=PDF('BINOMIAL',X,0.2,20);
        OUTPUT;
    END;
RUN;
TITLE H=1.2 F=ARIAL 'BINOMIAL(N=20, P=0.2)';
PROC GPLOT DATA=FN4;
    SYMBOL I=NEEDLE V=NONE CI=RED W=2;
    PLOT FX*X;
RUN;
```

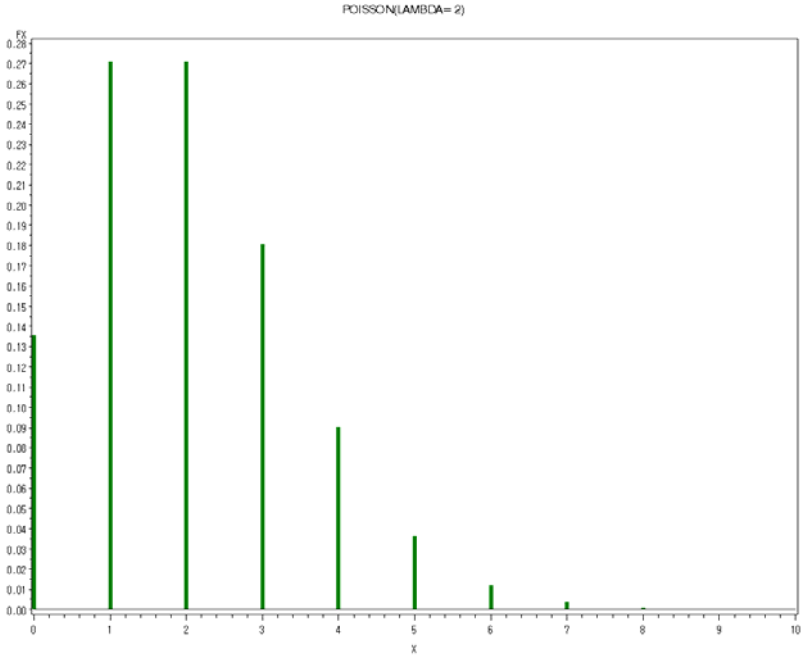


EXAMPLE: PDF 함수 사용하기(4) 포아송분포 그리기

$\lambda = 2$ 인 포아송 분포의 확률밀도함수를 그리시오. X-축 최대값을 얼마로 설정하는 것이 좋은가? 실제 이론적으로는 무한대(∞)의 값을 갖지만 평균이 2인 경우 10 이상이면 확률이 0이므로 10까지만 고려하였다.

FN1.sas * PROC GPLOT 실행 중

```
DATA FN4;
    DO X=0 TO 10;
        FX=PDF('POISSON',X,2);
        OUTPUT;
    END;
RUN;
GOPTIONS;
TITLE H=1.2 'POISSON(LAMBDA=2)';
PROC GPLOT DATA=FN4;
    SYMBOL I=NEEDLE V=NONE CI=GREEN W=5;
    PLOT FX*X;
RUN;
```

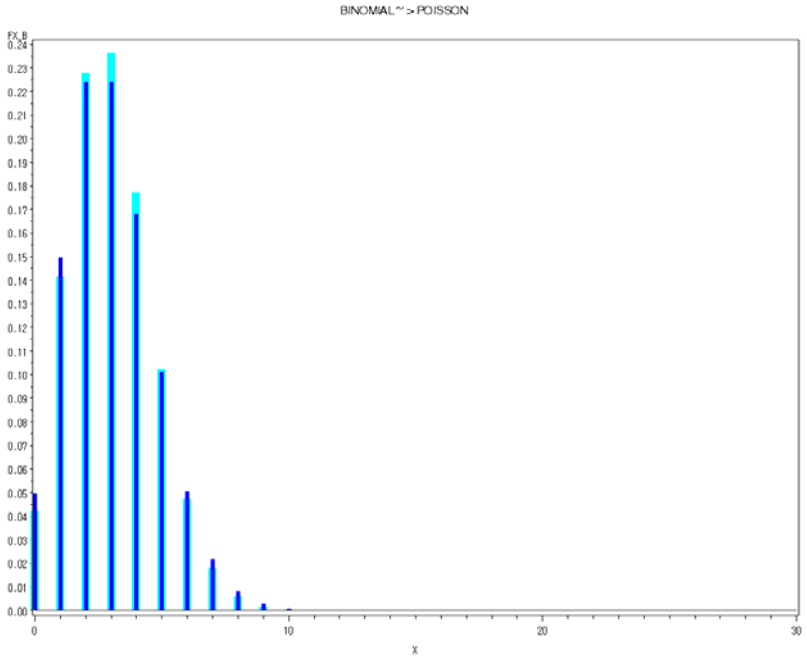


EXAMPLE: PDF 함수 사용하기(5) 이항분포의 포아송분포 근사

이항분포의 n이 크고 p가 매우 작은 경우 이항분포 (n, p) 는 모수 $\lambda = np$ 인 포아송 분포에 근사 한다. 이를 PDF 함수를 이용하여 살펴보기로 하자. 모수 $(n = 30, p = 0.1)$ 인 이항분포(cyan 색, 굵은 막대)와 모수 $\lambda = 3$ 인 포아송 분포(blue색, 얇은 막대)를 한 그래프에 그렸다. GOPTIONS; 문장은 모든 그래프 관련 설정을 RESET하는 문장이다.

FN1.sas * PROC GPLOT 실행 중

```
DATA FN7;
     DO X=0 TO 30;
          FX_B=PDF('BINOMIAL',X,0.1,30);
          FX_P=PDF('POISSON',X,3);
          OUTPUT;
     END;
RUN;
GOPTIONS;
TITLE H=1.2 'BINOMIAL~>POISSON';
PROC GPLOT DATA=FN7;
     SYMBOL1 I=NEEDLE V=NONE CI=CYAN W=10;
     SYMBOL2 I=NEEDLE V=NONE CI=BLUE W=5;
     PLOT (FX_B FX_P)*X/OVERLAY;
RUN;
```



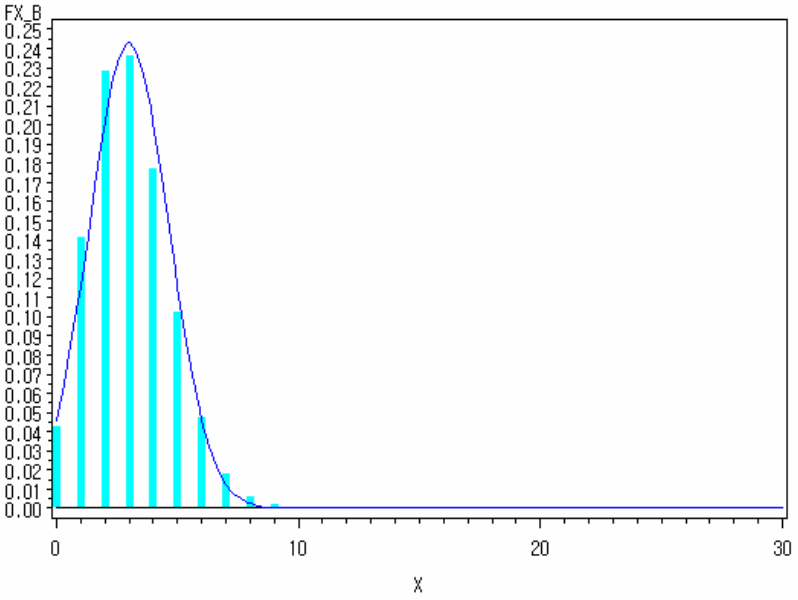
EXAMPLE: PDF 함수 사용하기(6) 이항분포의 정규분포 근사

n이 큰 경우 모수 (n, p) 인 이항분포는 평균 np, 분산이 npq 인 정규분포에 근사 한다. 모수 $\lambda = np$ 인 포아송 분포에 근사 한다. 이를 PDF 함수를 이용하여 살펴보기로 하자. 모수 $(n = 30, p = 0.1)$ 인 이항 분포와 평균 3, 분산 2.7인 정규분포를 함께 그려보자.

FN1.sas * PROC GPLOT 실행 중

```
DATA FN8;
    DO X=0 TO 30;
        FX_B=PDF('BINOMIAL',X,0.1,30);
        FX_N=PDF('NORMAL',X,3,SQRT(2.7));
        OUTPUT;
    END;
RUN;
GOPTIONS;
TITLE H=1.2 'BINOMIAL~>NORMAL';
PROC GPLOT DATA=FN8;
    SYMBOL1 I=NEEDLE V=NONE CI=CYAN W=5;
    SYMBOL2 I=SPLINE V=NONE CI=BLUE W=1;
    PLOT (FX_B FX_N)*X/OVERLAY;
RUN;
```

BINOMIAL~>NORMAL



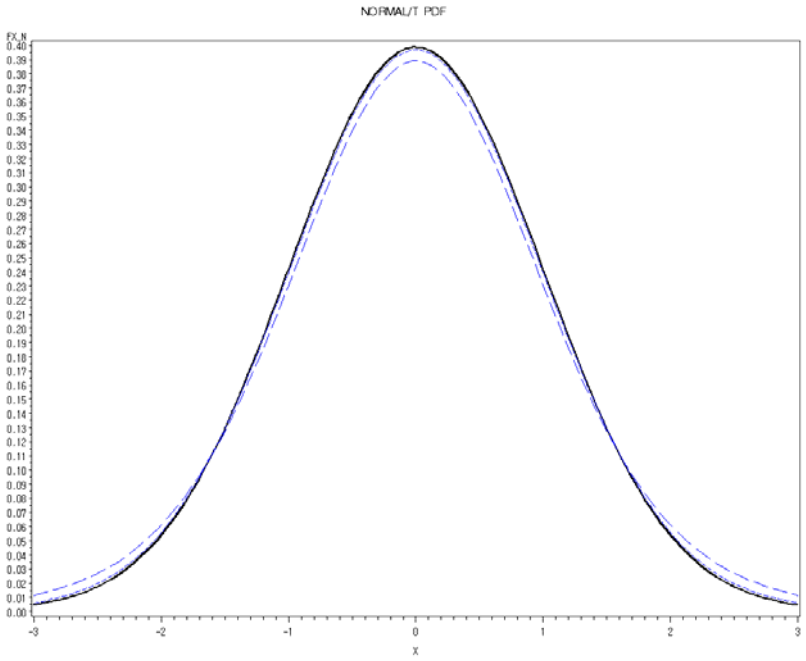
EXAMPLE: PDF 함수 사용하기(7) 정규분포와 T-분포

표준정규분포(평균=0, 분산=1), 자유도 10인 t분포(평균=0, 분산=10/8=1.25), 자유도 50인 t-분포(평균=0, 분산=1.042) 하나의 그래프에 그려보자. t-분포는 표준정규분포처럼 좌우 대

칭이나 분산(자유도가 n인 경우 $n/(n-2)$)이 약간 크므로 표준정규분포에 비해 좌우 꼬리 부분이 두껍다. T-분포의 자유도가 커질수록 표준정규분포에 근사 한다.

PDF.sas * PROC GPLOT 실행 중

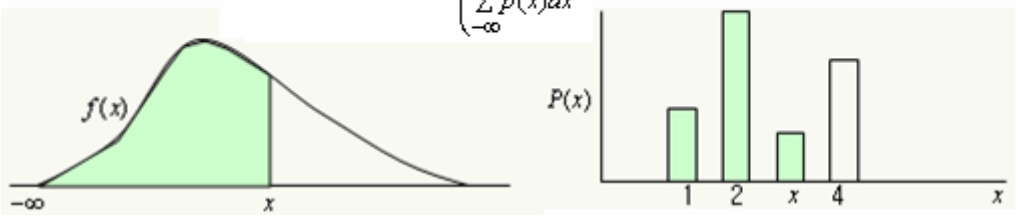
```
DATA PDF1;
     DO X=-3 TO 3 BY 0.5;
          FX_N=PDF('NORMAL',X,0,1);
          FX_T1=PDF('T',X,10);
          FX_T2=PDF('T',X,50);
          OUTPUT;
     END;
RUN;
TITLE H=1.2 F=SWISS 'NORMAL/T PDF';
PROC GPLOT DATA=PDF1;
     SYMBOL1 I=SPLINE V=NONE C=BLACK L=1 W=2;
     SYMBOL2 I=SPLINE V=NONE C=BLUE L=2;
     SYMBOL3 I=SPLINE V=NONE C=BLUE L=3;
     PLOT (FX_N FX_T1 FX_T2)*X/OVERLAY;
RUN;
```



4.3.2 누적확률분포함수

누적분포함수(CDF, Cumulative Density Function, 분포함수라 한다)는 확률분포함수의 가장 왼쪽 $-\infty$ 으로부터 임의의 점 x 까지 적분한 값을 의미한다. 초록색 부분의 면적이 $F(x)$ 이다. $F(-\infty)=0$, $F(\infty)=1$ 이다. 누적분포함수 값을 얻으려면 확률분포함수를 적분하여야 한다. SAS는 이를 CDF 함수로 내장하고 있다. CDF 함수는 사용방법, 옵션 등이 PDF 함수와 동일하다.

$$F(x) = P(X \le x) = \begin{cases} \int_{-\infty}^{x} f(x)dx \\ \sum_{-\infty}^{x} p(x)dx \end{cases}$$

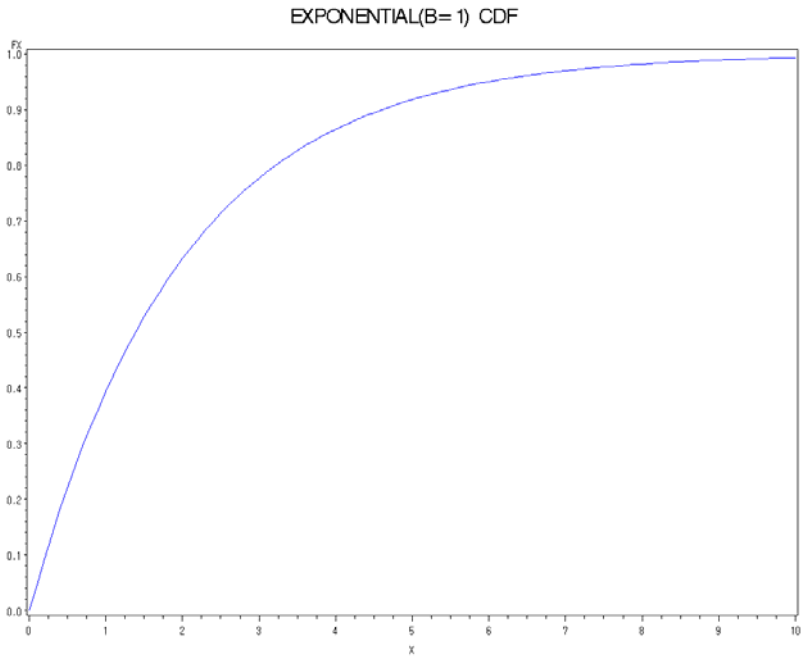


EXAMPLE: CDF 함수 사용하기, 지수분포의 누적분포함수

평균이 β 인 지수분포의 누적분포함수는 $F(x)=\int_{-\infty}^{x}\frac{1}{\beta}e^{-t/\beta}dt=1-e^{-x/\beta}$ 이다. 평균이 2인 지수분포의 누적분포함수를 그려보자.

CDF.sas *

```
DATA CDF1;
     DO X=0 TO 10 BY 0.5;
          FX=1-EXP(-X/2);
          OUTPUT;
     END;
RUN;
TITLE 'EXPONENTIAL(B=1) CDF';
PROC GPLOT DATA=CDF1;
     SYMBOL I=SPLINE V=NONE C=BLUE;
     PLOT FX*X;
RUN;
```



다음과 같이 CDF 함수를 이용하여 동일한 결과를 얻는다.

CDF.sas * PROC GPLOT 실행 중

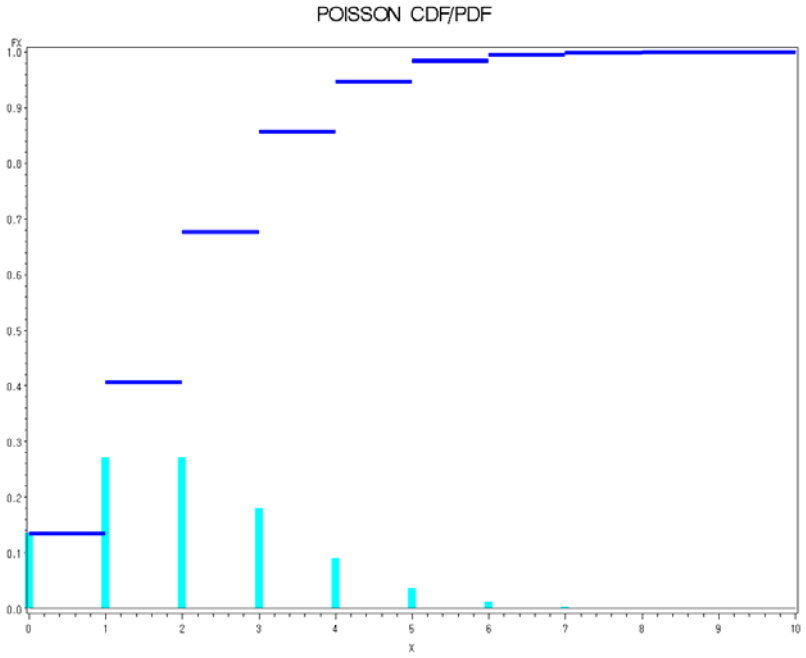
```
DATA CDF1;
    DO X=0 TO 10 BY 0.5;
        FX=CDF('EXPONENTIAL',X,2);
        OUTPUT;
    END;
RUN;
```

EXAMPLE: CDF 함수 사용하기(2) 포아송 분포

모수 $\lambda = 2$ 인 포아송 분포의 확률분포함수와 누적분포함수 하나의 그래프에 그리시오. CDF 그릴 때는 Interpolate 옵션을 STEP으로 하면 된다. X 값이 증가할 때마다 막대의 높이만큼 증가한다. 즉 X=1의 누적 값은 0, 1의 막대 높이 합이다.

CDF.sas * PROC GPLOT 실행 중

```
DATA CDF1;
    DO X=0 TO 10;
        FX=PDF('POISSON',X,2);
        FX1=CDF('POISSON',X,2);
        OUTPUT;
    END;
RUN;
TITLE 'POISSON CDF/PDF';
PROC GPLOT DATA=CDF1;
    SYMBOL1 I=NEEDLE V=NONE CI=CYAN W=10;
    SYMBOL2 I=STEP V=NONE C=BLUE W=5;
    PLOT (FX FX1)*X/OVERLAY;
RUN;
```



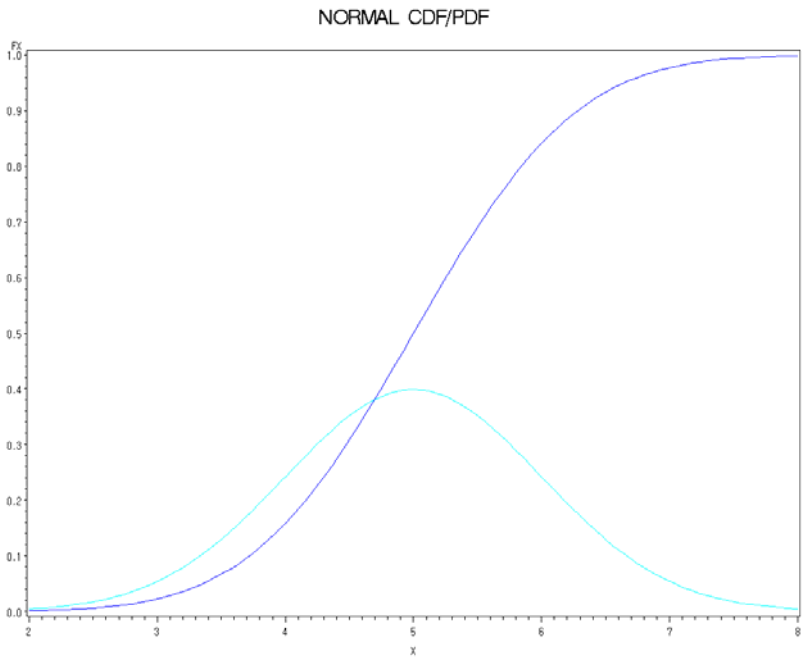
EXAMPLE: CDF 함수 사용하기(3) 정규 분포

평균이 5이고 표준편차가 1인 정규분포 확률분포함수와 누적확률분포함수를 하나의 그래프에 그리시오. 한동안 CDF 함수가 PDF 함수보다 작다? 맞나? 그렇다. 누적확률분포함수

는 거기까지의 확률밀도함수 면적임을 상기하기 바란다.

CDF.sas * PROC GPLOT 실행 중

```
DATA CDF1;
    DO X=2 TO 8 BY 0.5;
        FX=PDF('NORMAL',X,5,1);
        FX1=CDF('NORMAL',X,5,1);
        OUTPUT;
    END;
RUN;
TITLE 'NORMAL CDF/PDF';
PROC GPLOT DATA=CDF1;
    SYMBOL1 I=SPLINE V=NONE C=CYAN W=1;
    SYMBOL2 I=SPLINE V=NONE C=BLUE W=1;
    PLOT (FX FX1)*X/OVERLAY;
RUN;
```



누적확률밀도함수에 대한 SAS 함수로 CDF 대신 다음을 사용하기도 한다. 그러나 CDF 를 사용하는 것이 정형화된 형태라 편리하다.

분포	누적 분포 함수
포아송 분포	POISSON

베타 분포	PROBBETA
이항 분포	PROBBNML
카이제곱 분포	PROBCHI
F-분포	PROBF
감마 분포	PROBGAM
초기하 분포	PROBHYPR
음이항 분포	PROBNEGB
정규분포	PROBNORM
T-분포	PROBT

4.3.3 백분위

데이터를 크기 순으로 배열 했을 때 데이터의 p %가 어떤 임의의 값보다 작고 $(100-p)$% 가 그 값보다 큰 경우 그 임의의 값을 $p-th$ 백분위(percentile) 값이라 한다. 50% 백분위 값을 중앙값(median), 25% 백분위 값을 일사분위(first quartile), 75% 백분위 값을 삼사분위(third quartile)이라 한다. 중앙값을 이사분위(second quartile)이라고도 한다.

데이터의 백분위 값을 계산하기 위하여 데이터의 순서 통계량(order statistics)을 구해야 한다. 관측치를 크기 순으로 정렬한 후 제일 작은 값부터 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ 으로 표기하고 이를 순서 통계량(order statistics) 이라 한다.

•순서통계량: $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$

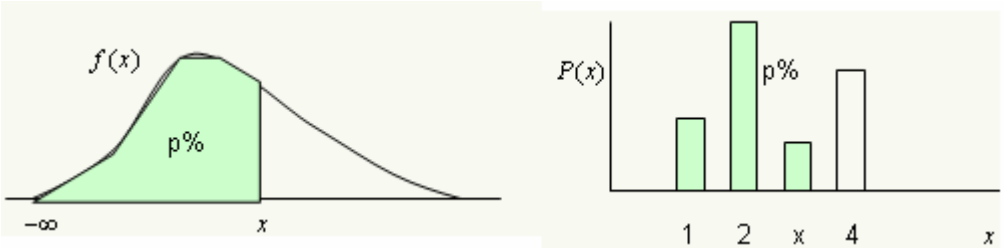
•최소값(min): $x_{(1)}$, 최대값(max): $x_{(n)}$

•범위(range): $x_{(n)} - x_{(1)}$

백분위 값이나 사분위 값을 구하려면 자료의 깊이(depth) 개념을 이용하면 편리하다. (Tukey가 제안) 관측치를 크기 순으로 정렬한 후 각 양쪽 끝에서 1부터 번호를 매겨 그 번호를 자료의 깊이라 정의한다. 즉 최대값, 최소값의 깊이는 각 1이다. 중앙값의 깊이는 $(n+1)/2$ 이고 사분위 깊이는 q=(중앙값의 깊이의 정수 함수+1)/2이다. 즉 중앙값은 $x_{((n+1)/2)}$ 이고 일사분위는 $x_{(q)}$ 이고 삼사분위는 $x_{(n-q+1)}$ 이다.

표본의 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. 그러므로 중앙값은 $x_{(3.5)=}$ (4+5)/2=4.5이고 일사분위는 $x_{(2)} = 2$ 이고 삼사분위는 $x_{(6-2+1)} = 6$ 이다.

확률분포함수에서 백분위 값의 개념을 알아보자. 초록색 부분의 확률이 p인 x값을 찾으면 이것이 p% 백분위 값이다. 표준 정규분포인 경우 95% 백분위 값을 찾으면 1.645이다. 이처럼 백분위는 누적분포함수의 역함수 형태가 된다.



일반적으로 이산형 확률밀도함수의 경우 정확한 p% 백분위 값을 구할 수 없다. 예를 들어 모수가 $(n=20, p=0.1)$인 경우 95% 백분위 값을 구해보자. 95% 백분위 값은 3과 4 사이의 값이다? 이처럼 이산형 확률분포함수의 경우 백분위 값을 구할 수 없으므로 SAS도 함수를 제공하지 않고 있다.

PCN.sas *

```
DATA PCN1;
    DO X=0 TO 20;
        FX=CDF('BINOMIAL',X,0.1,20);
        OUTPUT;
    END;
RUN;
TITLE 'BINOMIAL CDF';
PROC PRINT DATA=PCN1;
RUN;
```

Obs	X	FX
1	0	0.12158
2	1	0.39175
3	2	0.67693
4	3	0.86705
5	4	0.95683
6	5	0.98875
7	6	0.99761
8	7	0.99958
9	8	0.99994
10	9	0.99999
11	10	1.00000
12	11	1.00000
13	12	1.00000
14	13	1.00000
15	14	1.00000

연속형 확률분포함수의 경우 확률변수 이름 첫 글자와 INV(inverse, 이 이름을 사용한 이유는 백분위 값과 누적분포함수는 역함수 관계이므로)를 결합하여 함수로 내장하고 있다. 단 표준정규분포의 경우에는 PROBIT 함수로 되어 있다.

분포함수	SAS 함수	사용 예제

표준정규분포
(Standard Normal)
	PROBIT(확률**)**
	평균 μ 이고 표준편차 σ 인 정규분포의 백분위 값을 어떻게 구하나?
	$z=\frac{x-\mu}{\sigma}$ 이므로
	$\mu+$**PROBIT(**확률**)***σ
		x=probit(0.95);
		x1=10+probit(0.95)*2;
		x x1
		1.64485 13.2897

감마분포
(Gamma)
	GAMINV(확률, α **)**
	만약 **GAMINV(**확률, α , β **)**인 경우에는
	β***GAMINV(**확률, α **)**
	$\because X \sim Gamma(\alpha,1) \Rightarrow \frac{X}{\beta} \sim Gamma(\alpha,\beta)$
		data one;
		x=gaminv(0.95,2);
		x1=2*gaminv(0.95,2);
		run;
		x x1
		4.74386 9.48773

지수분포
(Exponential)
	β***GAMINV(**확률, 1**)**
	$\because X \sim Gamma(\alpha=1,\beta)=Exppnential(\beta)$
		data one;
		x=gaminv(0.95,1);
		x1=2*gaminv(0.95,1);
		run;
		x x1
		2.99573 5.99146

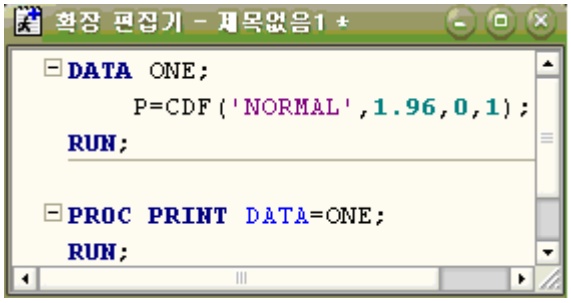
카이제곱
분포
(Chi-squared)
	CINV(확률, 자유도**)**	data one;
		x=cinv(0.95,1);
		x1=cinv(0.95,2);
		run;
		x x1
		3.84146 5.99146

T-분포	**TINV(**확률, 자유도**)**	data one;
		x=tinv(0.975,20);
		x1=tinv(0.025,20);
		run;
		x x1
		2.08596 -2.08596

F-분포	**FINV(**확률, 분자자유도, 분모자유도**)**	data one;
		x=finv(0.975,5,7);
		x1=finv(0.025,5,7);
		run;
		x x1
		5.28524 0.14592

EXAMPLE: 백분위 함수 사용하기, 표준정규분포

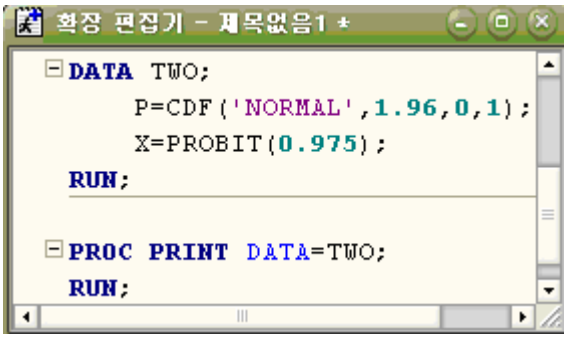
표준 정규분포의 경우 예를 들어보자. $x=1.96$ 까지 누적분포함수 값(확률)을 구하려면 다음 프로그램을 실행하면 된다.



Obs	P
1	0.97500

$$\int_{-\infty}^{1.96} \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}} dx = ?$$

누적확률이 0.975(즉 97.5% 백분위)인 확률변수 값을 구하려면 다음과 같이 하면 된다.



Obs	P	X
1	0.97500	1.95996

$$\int_{-\infty}^{?} \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}} dx = 0.975$$

백분위 함수의 대표적인 사용 예제는 통계적 가설 검정을 위한 검정 통계량으로부터 유의확률을 계산할 때이다. 전체 데이터를 가진 경우 데이터 분석은 PROC 단계를 이용하여 관련 통계량과 유의확률을 모두 얻을 수 있으나 통계학 원론에서 자주 등장하는 수작업 문제는 원 데이터를 주는 것이 아니라 필요한 통계량만 주고 가설 검정하라는 경우 함수가 사용된다.

EXAMPLE: PROC와 함수 이용 구별하기

대학생들의 평균 IQ 120이라고 한다. 우리 대학 학생들의 IQ가 전체 대학생 평균 IQ와 같은지 알아보기 위하여 10명을 무작위 추출하여 IQ를 측정하였다. 학생들의 IQ 분포는 정규분포를 따른다고 가정하자.

표본 데이터: 120, 125, 130, 125, 135, 110, 120, 125, 130, 140

①귀무가설: $\mu = 120$, 대립가설 $\mu \neq 120$, μ는 모집단 평균

②검정통계량=2.25, 유의확률(p-값)은 0.051이므로 귀무가설을 기각하지 못한다. 우리 대학 학생들의 IQ는 120과 같다고 할 수 있다. (자세한 내용은 4.4절 참고)

PROC TTEST(PROC 단계)를 이용하면 검정 통계량과 유의확률을 얻는다.

확장 편집기 - 제목없음1 *

```
DATA IQ;
    INPUT IQ @@;
    CARDS;
    120 125 130 125 135
    110 120 125 130 140
RUN;
PROC TTEST DATA=IQ HO=120;
RUN;
```

Variable	Lower CL Mean	Mean	Upper CL Mean
IQ	119.97	126	132.03

T-Tests

Variable	DF	t Value	Pr > |t|
IQ	9	2.25	0.0510

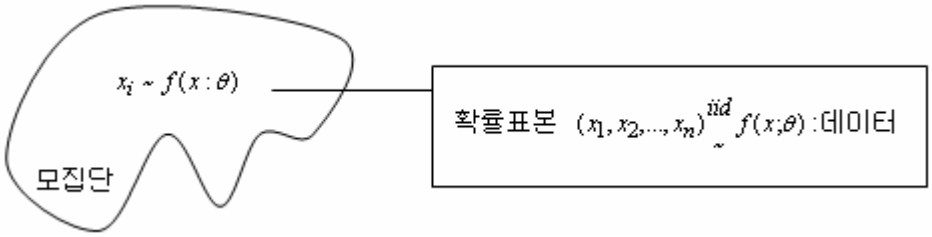
원 데이터가 주어진 것이 아니라 표본평균 126, 표본 표준편차 8.433만 주어졌다면 가설검정을 위하여 PROC 단계를 사용할 수 없다. 수작업에 의해 검정 통계량을 계산하고 확률분포 표에 의해 주어진 유의수준에 대응하는 임계치를 찾아 비교하여 귀무가설 채택 여부를 결정한다. 수작업 계산 대신 함수를 이용하여 통계량과 유의확률을 구할 수 있다. 이는 다음 절의 예제를 참고하기 바란다.

4.4 통계적 가설 검정

4.4.1 기본 개념

모수와 통계량

모집단은 알고자 하는 대상이 되는 집단 의미한다. 모집단을 통계학에서 표현할 때는 확률분포함수 $f(x;\theta)$ 으로 정의한다. θ 는 모수이다. 그러므로 모집단에서 우리가 알지 못하는 것은 확률 분포 함수 f 와 모집단 특성의 요약 값인 모수 θ 이다. 그러나 실제 모집단의 확률밀도함수 $f(x)$ 을 추정하기는 불가능할 뿐 아리나 실제 관심의 대상은 모수이다. 즉, 우리의 관심은 모수(parameter) 값에 대한 정보이다. 이것들을 표본 데이터를 이용하여 추정하게 된다. 표본 데이터로부터 계산된 값들을 통계량(statistic)이라 한다.



모집단(population)	표본(sample) $(x_1, x_2, ..., x_n)$
θ (모수) 중 관심이 있는 값은 모평균 μ, 모분산 σ^2, 모비율 p 이다.	모수에 대한 좋은 추정치($\hat{\theta}$)로 표본평균 $\bar{x}$, 표본분산 s^2, 표본비율 $\hat{p}$

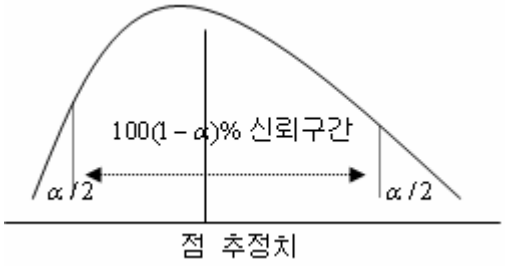
표본으로부터 계산된 통계량이 모수 추정에 사용되면 추정치(estimate)라 하고 가설 검정에 사용되면 검정 통계량(test statistic)이라 한다.

추정(estimation)

표본 데이터 $(x_1, x_2, ..., x_n)$ 으로부터 계산된 통계량을 이용하여 모수의 값에 대한 정보를 얻는 것을 추정치를 얻는다고 한다. 모수 추정에는 모수를 하나의 값으로 추정하는 점 추정

과 구간으로 추정하는 구간 추정이 있다. 모수를 어떤 통계량으로 추정하는 것이 좋은가? 통계학에서는 좋은 추정치를 Best Linear Unbiased Estimator(혹은 최소 분산 추정치 Minimum Variance Unbiased Estimator)라 한다.

모수 중 가장 많은 관심의 대상은 모집단 평균(μ), 모집단 분산(σ^2), 모집단 비율(p)이다. 이에 대한 좋은 점 추정치(MVUE)는 표본 평균 $\bar{x} = \frac{\sum x_i}{n}$, 표본 분산 $s^2 = \frac{\sum (x_i - \bar{X})^2}{n-1}$, 그리고 표본 비율 $\hat{p} = \#of성공/n$ 이다. 점 추정치와 점 추정치의 확률분포함수를 이용하여 모수에 대한 구간 추정치(interval estimation)를 얻을 수 있다. 다음 그림은 점 추정치와 구간 추정치를 나타낸 것이다. 확률분포함수는 점 추정치(통계량)의 확률분포함수이다. 모평균에 대한 MVUE는 표본 평균이다. 대표본인 경우 표본 평균의 정규분포에 근사하므로 이를 이용하여 신뢰구간을 구한다.



가설검정

통계적 가설 검정은 (1)서로 배반인 두 개의 가설(귀무가설, 대립가설)을 설정하고 (2)표본 데이터로부터 적절한 검정 통계량 값을 계산하고 (3)이를 이용하여 두 가설 중 하나를 선택하는 순서로 진행된다.

예를 들어 모집단의 평균에 관심이 있다고 하자. 모수는 $\theta = \mu$ 이다. 모집단 평균에 대한 MVUE는 표본평균 $\bar{x}$ 이다. 그러므로 $\hat{\theta} = \bar{x}$ 는 모수 μ 의 점 추정치(point estimate)이다. 그리고 우리는 $\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$ 사실을 알고 있으므로 모평균에 대한 $100(1-\alpha)\%$ 신뢰구간을 구하면 $\bar{x} \pm t(n-1;1-\alpha/2)\frac{s}{\sqrt{n}}$ 이다. $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ 는 통계적 가설 $H_0 : \mu = \mu_0$ 에 대한 검정을 위해 사용되는 검정통계량이다.

통계적 가설(statistical hypothesis)

통계적 가설 검정이란 ①서로 배반인 두 개의 통계적 가설(귀무가설, 대립가설)을 설정하고 ②적절한 검정 통계량 값을 계산하고 ③이를 이용하여 두 가설 중 하나를 선택한다. 연구하고자 하는 내용을 표본 데이터로부터 계산한 통계량을 이용하여 사실 여부를 판단할 수 있도록 설정한 내용을 통계적 가설이라 한다.

가설 종류

(1)귀무가설(Null Hypothesis)

원래 그대로의 상태(state quo)를 의미하며 표본에 의해 거짓임이 판명되지 않으면 기각되지 않는다. 귀무가설은 상태 그대로를 의미하여 “null=nothing”이라 이름 붙였으며 “~같다”, “영향을 미치지 않는다”, “차이가 없다” 식으로 정의된다. 귀무가설은 모수에 대한 하나의 값을 설정한다.

(2)대립가설(Alternative Hypothesis)

귀무가설과 대립되는 가설로 얻고자 희망하는 모수의 조건이나 변수들간의 관계에 대한 문장 (statement)으로 연구 가설(research hypothesis)이라 한다. “보다 크다”, “같지 않다”, “영향을 미친다” 등으로 설정된다. 연구자가 원하는 내용이 대립가설에 있으므로 이를 연구가설(research hypothesis)이라 한다. 대립가설은 모수에 대한 영역으로 설정된다.

가설 종류(2)

대립 가설은 형태에 따라 단측 가설(one-sided)과 양측(two-sided) 가설로 나눈다. 단측 가설은 모수에 대한 한 쪽 영역만 설정한 것이고 양측 검정에서는 귀무가설에 설정한 모수 이외 영역이 설정된다. 양측검정은 “~와 같지 않다”, 단측 검정은 “<” 혹은 “>”으로 표현된다.

검정 오류

귀무가설이 사실인데 귀무가설을 기각할(대립가설 채택) 확률을 1종 오류(type I error)라

하고 귀무가설이 거짓인데도(대립가설이 사실) 귀무가설을 기각하지 않을(대립가설 채택) 확률을 2종 오류(type II error)라 한다.

실제 상황 / 검정 결과	귀무가설 진실	대립가설 진실
귀무가설 기각	1종 오류 α	옳은 판단
귀무가설 채택	옳은 판단	2종 오류 β

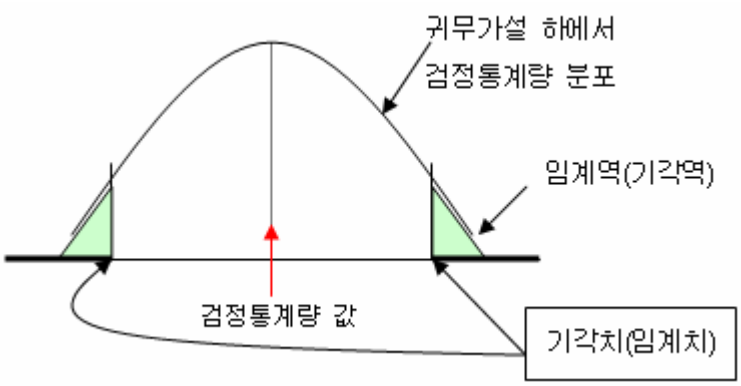
유의 수준(significant level)

가설 검정 방법 중 두 오류를 모두 줄일 수 있는 방법은 존재하지 않으므로 두 오류 중 하나를 고정하고 다른 오류를 줄일 수 있는 방법을 찾는다. 어느 오류를 고정할 것인가? 우리의 관심은 대립 가설에 있으므로 귀무가설을 기각할 확률을 고정하고(사용될 통계적 방법은 이 정도는 희생은 감수한다.) 대립가설이 사실인 경우 대립가설을 채택할 검정력 $(1-\beta)$을 최대화 할 수 있는 통계적 가설 검정 방법을 찾는다. 가설 검정을 위하여 설정한 1종 오류를 유의 확률이라 하며 일반적으로 10%, 5%(가장 일반적), 1%를 사용한다.

검정 통계량(Test Statistics), 기각역, 기각치(Reject region and Value)

귀무가설의 사실 여부를 판단하기 위하여 사용되는 통계량을 검정 통계량이라 한다. 검정 통계량은 표본 데이터로부터 계산되면 가설 검정하려면 검정 통계량의 분포를 알아야 한다. 검정 통계량의 분포는 귀무가설이 진실이라 가정하에 구하게 된다. 이래 그림에서 귀무가설 진실 하에서 구한 분포이므로 양쪽 극단(초록 부분)도 일어날 수 있으나 이런 극단이 발생하면 귀무가설을 기각한다. 그러므로 검정 오류이다. 이것을 1종 오류라 하고 미리 설정된 1종 오류를 유의수준이라 한다.

설정된 유의수준 하에서 귀무가설을 기각하게 되는 검정 통계량 값들의 영역(초록 부분)을 기각역(짙은 직선 영역)이라 하고 영역의 시작점을 기각치 혹은 임계치(critical region)라 한다. 대립가설 양측 가설이면 임계 영역이 양쪽 영역이고 양쪽 초록 부분의 합이 유의수준이다.

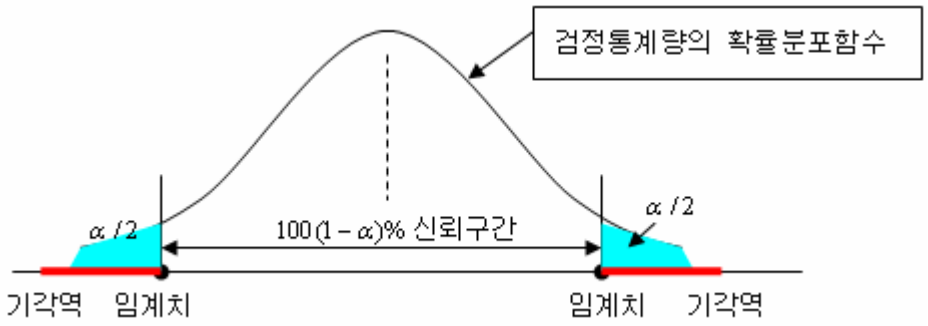


신뢰 구간 (Confidence interval)

임의의 모수에 대해 두 통계량(하한(L), 상한(U))이 존재하며 다음과 같이 쓸 수 있다면 (L,U) 을 모수(θ)에 대한 $100(1-\alpha)\%$ 신뢰구간이라 한다.

$$P(L(x_1,x_2,...x_n) < \theta < U(x_1,x_2,...x_n)) = 1-\alpha$$

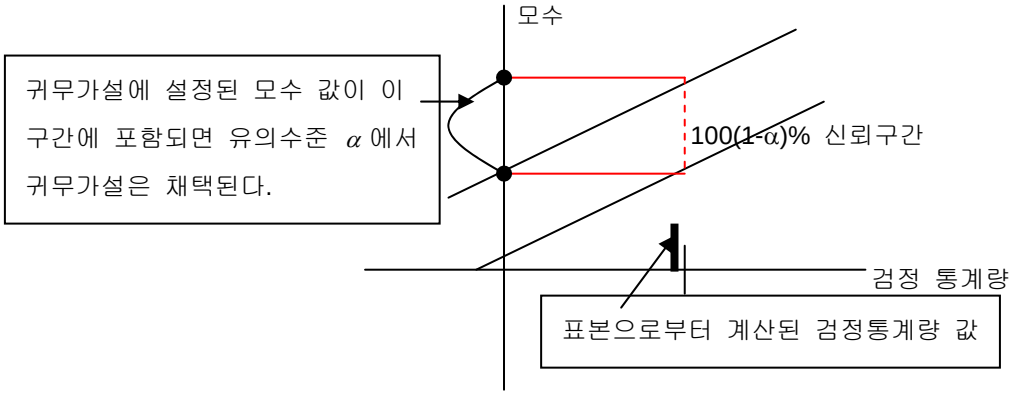
$100(1-\alpha)\%$ 을 신뢰 수준(confidence level), $L(x_1,x_2,...x_n)$ 을 하한(lower bound), $U(x_1,x_2,...x_n)$ 을 상한(upper bound)라 한다. 95% 신뢰구간의 실제 의미는 모수(모집단 인장 강도 평균)가 신뢰구간에 포함될 가능성(확률)이 95%를 의미하는 것이 아니다. 모집단으로부터 표본의 크기 20인 표본을 뽑아 신뢰구간을 구하고, 또 표본의 크기 20의 표본을 뽑아 신뢰구간을 구하고, 이런 과정을 100번 반복하면 그 중 95개의 신뢰구간은 모수를 포함하고 있다는 것이다.



100(1-α)% 신뢰 구간과 유의수준 α인 가설 검정(양측 검정)

일대일 대응 관계가 존재한다. 95% 신뢰구간에 포함되지 않은(포함되는) 모수 값이 유의

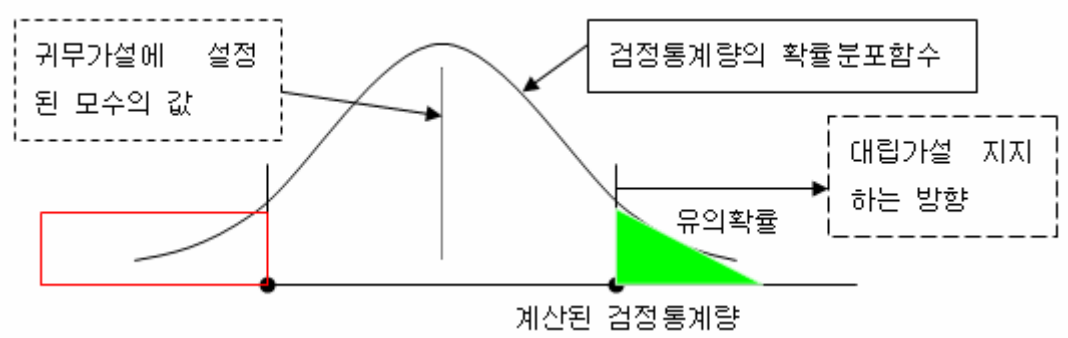
수준 5%하에서 검정되는 귀무가설에 설정되면 그 귀무가설은 기각(채택)된다. 이 관계를 그림으로 나타내면 다음과 같다.



유의확률(significant probability), p-값 (p-value)

귀무가설을 기각할 최소의 유의 확률을 의미한다. 즉 귀무가설을 기각하려면 유의 수준을 p-값으로 설정하면 된다. p-값은 확률로 검정 통계량 값이 계산되었을 때 이 값보다 클 확률(극한 상황 발생)을 의미한다. 그러므로 p-값이 0.05보다 크다면 귀무가설을 기각할 수 없고 작다면 귀무가설을 기각한다. P-값이 0.06이라면 유의 수준을 0.06으로 해야 귀무가설을 기각할 수 있다는 것이므로 유의 수준이 0.05(5%)이면 귀무가설이 기각되지 않는다.

다음 그림은 유의확률을 그림으로 나타낸 것이다. 그림에서 초록 부분은 대립가설이 “>귀무가설에 설정된 모수 값”인 단측 가설의 경우 유의확률이다. 대립가설이 “<귀무가설에서 설정된 모수 값” 형태이면 왼쪽 꼬리 부분이 유의확률이다. 대립가설이 양측 가설일 경우에는 한 쪽에서 얻어진 유의확률(초록색 부분)을 2배 하면 된다.



유의확률이 설정된 유의수준 하에서 기각역을 제시하는 것보다 다 많은 정보를 우리에게 주므로 통계소프트웨어는 검정통계량과 유의확률(대립가설이 양측가설인 경우)을 출력한다.

4.4.2 공식 정리

단일 집단 모평균 추론 $H_0 : \mu = \mu_0$

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

소표본인 경우에는 모집단 정규분포 가정이 필요하다. 만약 모집단이 정규분포를 따른다는 확신이 없는 소표본인 경우 비모수 방법(Sign검정, Wilcoxon Ranks Sum)을 실시한다.

②신뢰구간: $\bar{x} \pm t(n-1;1-\alpha/2)\dfrac{s}{\sqrt{n}}$ (표본의 크기가 커지면 T-분포는 정규분포에 근사하므로 수작업 시에는 표본의 크기가 20~30 이상인 대표본의 경우에는 정규분포 표 이용한다.)

단일 집단 모비율 추론 $H_0 : p = p_0$

①검정통계량: $T = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}} \sim N(0,1)$ (대표본)

소표본인 경우에는 유의확률 개념을 이용하여 가설 검정한다.

②신뢰구간: $\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ (대표본)

두 집단 모평균 추론 $H_0 : \mu_1 = \mu_2$

①검정통계량: $T = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2) = 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)[\text{소표본}] \; or \; N(0,1)[\text{대표본}]$,

통합 분산(pooled variance) $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$

소표본인 경우에는 두 모집단 정규분포 가정이 필요하다. 만약 두 모집단이 정규분포를 따른다는 확신이 없는 소표본 경우 비모수 방법(Median검정, Mann-Whitney Test)을 사용하여 가설 검정한다.

②신뢰구간: $(\bar{x}_1 - \bar{x}_2) \pm t(n_1 + n_2 - 2; 1-\alpha/2)s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

두 집단 모비율 추론 $H_0 : p_1 = p_2$

①검정통계량: $T = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2) = 0}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \sim N(0,1)$ (대표본)

②신뢰구간: $(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$ (대표본)

4.4.3 사용 예제

원 데이터기 주어진 경우에는 PROC 단계를 이용하여 통계량을 구하거나 가설 검정을 할 수 있다. 그러나 주요 통계량의 값만 주어진 경우에는 함수를 이용하여 검정통계량과 유의확률을 계산해야 가설 검정이 가능하다.

EXAMPLE: 모비율 검정 (대표본)

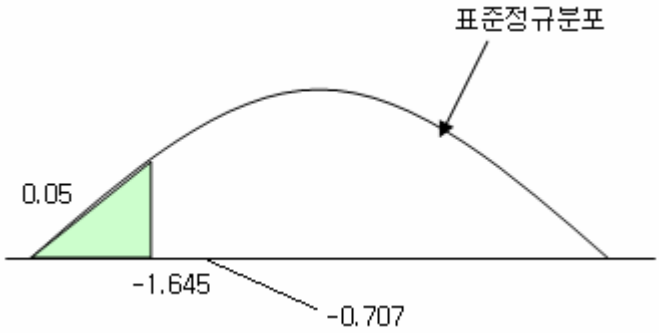
성인의 흡연 비율은 0.2라고 한다. 우리 대학 학생들의 흡연 성인 흡연 비율 0.2보다 낮은지 알아보기 위하여 50명을 임의 추출하여 8명이 흡연하고 있음을 알았다. 유의수준 0.05으로 가설 검정하고 95% 신뢰구간을 구하시오.

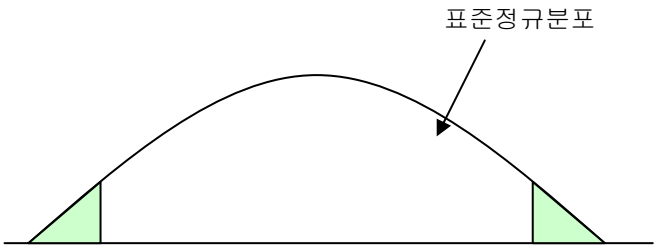
⑴귀무가설: 한남대학생 흡연 비율은 0.2이다. $p = 0.2$

대립가설: 비율은 0.2미만이다. $p < 0.2$

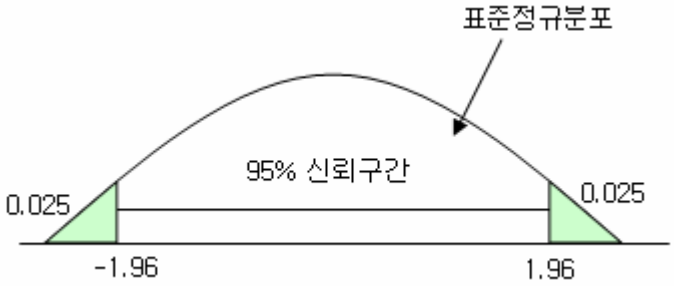
⑵검정통계량: $T = \dfrac{0.16 - 0.2}{\sqrt{\dfrac{0.2*(1-0.2)}{50}}} = -0.707 \sim Normal(0,1)$

대립가설이 단측가설(왼쪽)이므로 유의수준 5%를 왼쪽 부분에만 설정한다. 검정통계량이 표준정규분포를 따르므로 기각치는 -1.645이다. 표본으로부터 계산된 검정통계량 -0.707이 -1.645보다 작지 않기 때문에 귀무가설은 기각되지 않는다. 흡연 비율은 0.2라 할 수 있다.





⑶신뢰구간 $\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.16 \pm 1.96\sqrt{\frac{0.2*0.8}{50}} \Rightarrow (0.058, 0.26)$ 신뢰구간은 대립가설이 양측인 경우 구하는 것이므로 양쪽에 2.5%씩 할당하게 된다.



위의 과정을 SAS에서 함수를 이용하여 구하면 다음과 같다. ABS(T)의 의미는 검정통계량을 +(확률의 우측 부분 고려)로 하기 위함이다. 이는 CV(기각치)의 오른쪽 극한을 구했기 때문이다. CV는 단측검정의 기각치(임계치)를 계산하기 위한 것이다. 검정 통계량의 분포가 정규분포를 따르므로 누적 확률이 0.05가 되는 백분위 값을 구하면 그것이 왼쪽 부분 기각역이 된다. 양측이면 우측에 0.025, 왼쪽에 0.025 배정하면 된다.

검정통계량 T와 기각역 RC가 비교하여 T가 RC보다 작으면(귀무가설의 모수 값에서 멀어짐) 귀무가설을 기각하고 크면 귀무가설을 채택한다. 대립가설이 단측이고 귀무가설보다 작은 쪽만 고려하므로 유의확률은 검정통계량 값보다 작을 확률을 계산하면 된다. 출력 결과 유의확률은 0.23이므로 유의수준 0.05보다 크므로 귀무가설은 채택된다. 신뢰구간은 양측 신뢰구간을 구하는 것이므로 PROBIT 함수에는 0.975를 사용한다. 하한구간(변수 LOW) 구할 때 만약 0.975대신 0.025를 사용하기 원하면 앞에 “–“대신 “+”을 사용해야 한다. 정규

분포가 좌우 대칭임을 유의하기 바란다.

만약 상한 신뢰구간을 구하려면 UP=PHAT+PROBIT(0.95)*SQRT(0.2*0.8/50); 을 사용하면 된다.

확장 편집기 - 제목없음1 *

```
DATA ONE;
     PHAT=8/50;
     T=(PHAT-0.2)/SQRT(0.2*0.8/50); /*검정통계량*/
     RC=PROBIT(0.05); /*기각치, 왼쪽 영역 */
     IF (T<RC) THEN CON='귀무가설 기각';
     ELSE CON='귀무가설 채택'; /*가설 채택 여부 */
     P=CDF('NORMAL',T,0,1); /*유의확률 계산*/
     UP=PHAT+PROBIT(0.975)*SQRT(0.2*0.8/50);
     LOW=PHAT-PROBIT(0.975)*SQRT(0.2*0.8/50); /*95% 신뢰구간*/
RUN;
PROC PRINT DATA=ONE;
RUN;
```

PHAT	T	RC	CON	P	UP	LOW
0.16	-0.70711	-1.64485	귀무가설 채택	0.23975	0.27087	0.049128

EXAMPLE: 백분위 함수 사용하기, 모평균 검정(소표본)

대학생 평균 IQ가 110이라 한다. ○○한남대학교 학생들의 평균 IQ가 110인지 알아보기 15명을 조사하였더니 평균 115, 분산 10이었다. 가설 검정하고(유의수준=0.05) 95% 신뢰구간을 구하시오. IQ 데이터는 정규분포를 따른다고 가정하자.

⑴귀무가설: 우리 대학생 IQ 평균은 110이다. $\mu = 110$

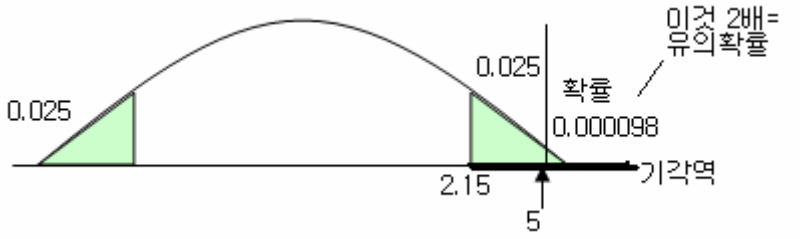
대립가설: $\mu \neq 110$

⑵검정통계량: $T = \frac{115-110}{\sqrt{10/15}} = 5 \sim t(n-1=14)$

검정통계량 T=5이고 기각치는 2.15이다. 검정통계량이 기각역에 속하므로 귀무가설이 기

각된다. 양측 검정이지만 일단 기각치 RC는 오른쪽 영역에서 구했다. 이는 T의 부호가 양이므로(표본평균이 귀무가설에 설정한 모집단 평균보다 크므로) 일단 오른쪽만 고려하면 된다. 그러나 단측 검정과는 달리 기각역을 계산할 때는 유의수준의 1/2인 0.025을 사용해야 한다.

양측 검정을 위한 유의확률을 계산할 때는 한 쪽 부분을 계산한 후 2배를 하면 된다. 1-CDF('T',T,N-1)은 검정통계량 보다 큰 부분의 확률을 계산한 것이므로 이것을 2배 하면 양측 검정을 위한 유의확률이 된다.



95% 신뢰구간에는 귀무가설에 설정된 값 110이 포함되어 있지 않으므로 유의수준 5%에서 귀무가설을 기각할 수 있다. 가설 검정 결과와 일치한다. 이는 앞에서 설명하였듯이 신뢰구간과 가설검정 간에는 일대일 대응 관계가 있다.

UNI.sas *

```
DATA TWO;
     M=115; S=SQRT(15); N=15; HO=110;
     T=(M-110)/(S/SQRT(N)); /*검정통계량*/
     RC=TINV(0.975,N-1); /*기각치, 오른쪽 영역*/
     IF (T>RC) THEN CON='귀무가설 기각';
     ELSE CON='귀무가설 채택'; /*가설 채택 여부 */
     P=2*(1-CDF('T',T,N-1)); /*유의확률 계산*/
     UP=M+TINV(0.975,N-1)*S/SQRT(N);
     LOW=M-TINV(0.975,N-1)*S/SQRT(N); /*95% 신뢰구간*/
RUN;
PROC PRINT DATA=TWO;
RUN;
```

HO	T	RC	CON	P	UP	LOW
110	5	2.14479	귀무가설 기각	.000194515	117.145	112.855

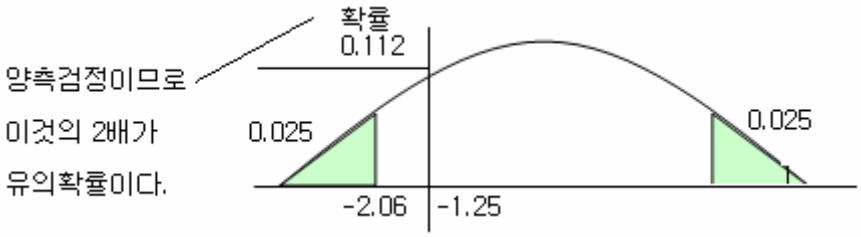
EXAMPLE: 백분위 함수 사용하기, 모평균 검정(대표본)

대학생 평균 키는 170이라 한다. ○○대학교 학생들의 평균 키가 170인지 알아보기 25명을 조사하였더니 평균 165, 분산 400이었다. 유의수준 5%에서 가설 검정하시오.

⑴귀무가설: 우리 대학생 IQ 평균은 170이다. $\mu = 170$

대립가설: $\mu \neq 170$

⑵검정통계량: $T = \frac{165-170}{20/\sqrt{25}} = -1.25 \sim t(24)$혹은$Normal(0,1)$



검정통계량의 부호가 음이므로 왼쪽 영역만 고려하여 기각역(유의수준의 1/2 설정)을 구하다. 검정통계량 -1.25, 기각치 -2.064이므로 귀무가설이 채택된다. 검정통계량 -1.25보다 작은 영역이 유의확률 계산에 사용된다. 양측 검정이므로 0.112의 2배가 유의확률이다. 유의확률이 0.224로 유의수준보가 크므로 귀무가설이 채택된다.

UNI.sas *

```
DATA THREE;
     M=165; S=SQRT(400); N=25; HO=170;
     T=(M-HO)/(S/SQRT(N)); /*검정통계량*/
     RC=TINV(0.025,N-1); /*기각치, 왼쪽 영역*/
     IF (T>RC) THEN CON='귀무가설 기각';
     ELSE CON='귀무가설 채택'; /*가설 채택 여부 */
     P=2*CDF('T',T,N-1); /*유의확률 계산*/
     UP=M+TINV(0.975,N-1)*S/SQRT(N);
     LOW=M-TINV(0.975,N-1)*S/SQRT(N); /*95% 신뢰구간*/
RUN;
PROC PRINT DATA=THREE;
RUN;
```

T	RC	CON	P	UP	LOW
-1.25	-2.06390	귀무가설 기각	0.22335	173.256	156.744

대표본의 경우 수작업 시에는 t-분포를 사용하는 것이 아니라 정규분포(중심극한정리 이용)를 이용하게 된다. 위와는 달리 정규분포를 이용할 경우 결과를 얻어보자. T-분포가 표준정규분포에 비해 꼬리가 두터우므로 t-분포 이용할 경우 기각역 값이 중심으로부터 정규분포에 비해 멀고 유의확률도 약간 크다. 통계소프트웨어는 모두 t-분포 이용하여 유의확률을 계산한다. 4.3.3절 PROC TTEST의 경우에도 t-통계량과 유의확률이 출력됨을 볼 수 있다.

UNI.sas *

```
DATA THREEO;
     M=165; S=SQRT(400); N=25; HO=170;
     T=(M-HO)/(S/SQRT(N)); /*검정통계량*/
     RC=PROBIT(0.025); /*기각치, 왼쪽 영역*/
     IF (T>RC) THEN CON='귀무가설 기각';
     ELSE CON='귀무가설 채택'; /*가설 채택 여부 */
     P=2*CDF('NORMAL',T,0,1); /*유의확률 계산*/
     UP=M+PROBIT(0.975)*S/SQRT(N);
     LOW=M-PROBIT(0.975)*S/SQRT(N); /*95% 신뢰구간*/
RUN;
PROC PRINT DATA=THREEO;
RUN;
```

T	RC	CON	P	UP	LOW
-1.25	-1.95996	귀무가설 기각	0.21130	172.840	157.160

EXAMPLE: 모분산 검정

품질 공정에서 분산이 0.2인 제품이 있다고 한다. 새로운 생산 공정이 제품의 분산을 낮추는지 알아보고자 한다. 제품 30개를 임의 추출하여 표본 분산을 계산하였더니 0.17이었다. 새로운 공정이 분산을 낮추었다고 할 수 있나? 유의수준 0.05에서 가설 검정하고 95% 상한 신뢰구간을 구하시오.

⑴귀무가설: 모집단 분산은 0.2이다. $\sigma^2 = 0.2$

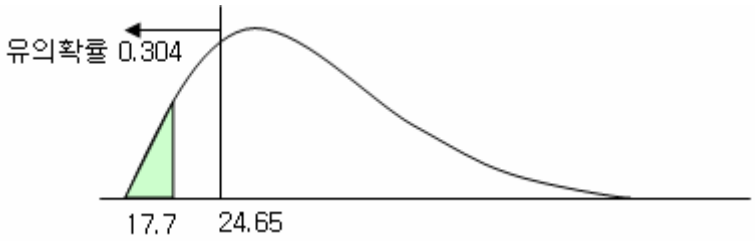
대립가설: $\sigma^2 < 0.2$

⑵검정통계량: $T = \frac{(n-1)s^2}{\sigma_0^2} = \frac{29*0.17}{0.2} = 24.65 \sim \chi^2(n-1)$

신뢰구간: $\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}$

대립검정통계량이 1보다 작으므로 카이제곱 분포에서 왼쪽 부분이 집중하면 된다. 그러므로 기각역도 왼쪽 부분만 보면 된다. 단측 검정이므로 기각역을 구할 때 0.05 사용하면 되고 유의확률도 검정통계량보다 작은 영역의 확률을 사용하면 된다.

검정통계량 24.65가 임계치보다 크므로 귀무가설이 채택된다. 유의확률 면에서도 귀무가설은 채택된다. 상한 신뢰구간을 구할 때는 양측 신뢰구간의 우측 분모에 $\chi^2_{\alpha=0.05}(df = 29)$을 사용하면 된다.



UNI.sas *

```
DATA TWO;
     S2=0.17; N=30; HO=0.2;
     T=(N-1)*S2/HO; /*검정통계량*/
     RC=CINV(0.05,N-1); /*기각치, 왼쪽 영역*/
     IF (T<RC) THEN CON='귀무가설 기각';
     ELSE CON='귀무가설 채택'; /*가설 채택 여부 */
     P=CDF('CHISQ',T,N-1); /*유의확률 계산*/
     UP=(N-1)*S2/CINV(0.05,N-1); /*95% 신뢰구간*/
RUN;
PROC PRINT DATA=TWO;
RUN;
```

S2	N	HO	T	RC	CON	P	UP
0.17	30	0.2	24.65	17.7084	귀무가설 채택	0.30372	0.27840

EXAMPLE: 두 모집단 평균 차이 검정

나무 해충 발생을 억제하는 치료제가 개발되었다. 효과가 있는지 알아보기 위하여 나무 14개를 임의 추출하여 7개는 치료제를 투여하고 나머지 개는 아무 처리도 하지 않았다. 일정 기간이 지난 후 나무의 해충 수를 조사하여 다음을 얻었다. 치료제가 효과가 있는지 유의수준 5%에서 가설 검정하시오.

치료제 투여 그룹: n=7, 표본 평균=28.57, 표본 분산=198.62

치료제 투여 않은 그룹: n=7, 표본 평균=40, 표본 분산=215.33

⑴귀무가설: 두 집단의 평균 해충 수는 같다. $\mu_1 = \mu_2$

대립가설: μ_1(치료제투여) $< \mu_2$

⑵검정통계량: $T = \dfrac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2) = 0}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{28.57 - 40}{14.39\sqrt{1/7 + 1/7}} = -1.486 \sim t(n_1 + n_2 - 2)$,

통합 분산(pooled variance) $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{6*198+6*215}{7+7-2}} = 14.39$

UNI.sas *

```
DATA TWO;
     M1=28.57;S11=198.62;N1=7;
     M2=40;S22=215.33;N2=7;
     SP=SQRT(((N1-1)*S11+(N2-1)*S22)/(N1+N2-2)); /*통합분산*/
     T=(M1-M2)/(SP*SQRT(1/N1+1/N2)); /*검정통계량*/
     RC=TINV(0.05,N1+N2-2); /*기각치, 왼쪽 영역*/
     IF (T<RC) THEN CON='귀무가설 기각';
     ELSE CON='귀무가설 채택'; /*가설 채택 여부 */
     P=CDF('T',T,N1+N2-2); /*유의확률 계산*/
RUN;
PROC PRINT DATA=TWO;
RUN;
```

SP	T	RC	CON	P
14.3866	-1.48635	-1.78229	귀무가설 채택	0.081490

대표본인 경우(표본의 크기가 20~30) 검정통계량의 분포는 표준정규분포를 따르므로 위의 프로그램에서 TINV대신 PROBIT, “T” 대신 “NORMAL”을 사용하면 된다. 위의 예제에서 치료제 투여 그룹의 표본의 크기 20, 투여하지 않은 그룹 표본의 크기는 30이었다면?

UNI.sas *

```
DATA TWOO;
     M1=28.57;S11=198.62;N1=20;
     M2=40;S22=215.33;N2=30;
     SP=SQRT(((N1-1)*S11+(N2-1)*S22)/(N1+N2-2)); /*통합분산*/
     T=(M1-M2)/(SP*SQRT(1/N1+1/N2)); /*검정통계량*/
     RC=PROBIT(0.05); /*기각치, 왼쪽 영역*/
     IF (T<RC) THEN CON='귀무가설 기각';
     ELSE CON='귀무가설 채택'; /*가설 채택 여부 */
     P=CDF('NORMAL',0,1); /*유의확률 계산*/
RUN;
```

SP	T	RC	CON	P
14.4470	-2.74069	-1.64485	귀무가설 기각	0.15866

EXAMPLE: 두 모집단 분산 차이 검정

두 생산 공정의 분산의 차이가 있는지 알아보기 위하여 생산공정1에서 크기 10인 표본을 추출하여 계산하였더니 0.51이었고 생산공정2의 표본분산 0.058(표본 크기=12)이었다. 생산공정의 분산 차이가 있는지 유의수준 5%에서 검정하시오.

⑴귀무가설: 두 생산공정의 분산은 같다. $\sigma_1^2 = \sigma_2^2$

대립가설: $\sigma_1^2 \neq \sigma_2^2$

⑵검정통계량: $T = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)} = \frac{0.105}{0.058} = 1.81 \sim F(n-1=9(\text{분자}), n-1=11(\text{분모}))$

검정통계량 계산할 때 항상 큰 표본 분산을 분자에 사용하므로 검정통계량 값은 항상 1보다 크므로 분포의 오른쪽 영역만 고려하면 된다. 그러나 여전히 양측 검정이므로 유의수준의 1/2을 사용한다. 즉 0.95가 아니라 0.975 사용해야 한다.

UNI.sas *

```
DATA FIVE;
     S11=0.105;N1=10;
     S22=0.058;N2=12;
     T=MAX(S11,S22)/MIN(S11,S22); /*검정통계량*/
     RC=FINV(0.975,N1-1,N2-1); /*기각치, 오른쪽 영역*/
     IF (T>RC) THEN CON='귀무가설 기각';
     ELSE CON='귀무가설 채택'; /*가설 채택 여부 */
     P=CDF('F',T,N1-1,N2-1); /*유의확률 계산*/
RUN;
PROC PRINT DATA=FIVE;
RUN;
```

S11	N1	S22	N2	T	RC	CON	P
0.105	10	0.058	12	1.81034	3.58790	귀무가설 채택	0.82498

4.5 확률변수 생성

임의의 분포를 따르는 확률변수를 만드는(생성, generate) 방법을 살펴보자. 생성된 데이터를 이용하여 사회 현상, 자연 현상을 컴퓨터에서 실현하여 결과를 미리 예측해 보는 것을 시뮬레이션(simulation)이라 한다.

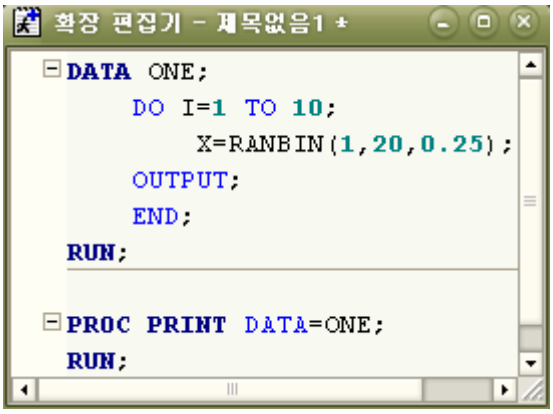
SAS에서 임의 분포를 따르는 확률변수 데이터를 생성하는 함수는 RAN* 이다. 난수를 생성할 때는 난수표(random number table)의 어디서 시작하느냐에 대한 seed 값을 지정하게 된다. seed 값은 0이나 $(2^{31}-1)$보다 적은 양의 정수 값을 사용하면 된다. seed를 0을 사용한 경우에는 컴퓨터가 실행된 시각이 seed로 들어간다. SEED 번호에 1 이상의 정수를 사용하면 프로그램 실행할 때마다 생성되는 데이터는 매번 동일하다.

4.5.1 이산형 변수

분포 함수	확률분포함수	SAS 함수
이항분포 (Binomial)	$p(x)=\binom{n}{x}p^x q^{n-x}$ $x=0,1,2,...,n$ 평균: np 분산: npq	**X=RANBIN(seed,n,p);** n 이 1 이면 Bernoulli 분포이다.
포아송분포 (Poisson)	$p(x)=\frac{e^{-\lambda}\lambda^x}{x!}$ $x=0,1,2...$ 평균: λ 분산: λ	**X=RANPOI(seed, λ);**
표 확률 분포 (Tabled Probability)	$p(x=i)=p_i, i=1,2,...,n$ $\sum_i p_i=1$	**X=RANTBL(seed,p1,p2,...,pn);** 1 이 나올 확률은 p_1, 2 나올 확률은 p_2,..., 정수 n 이 나올 확률이 p_n. $X=RANTBL(1,1/2,1/2)$ 는 1 이 나올 확률이 0.5, 2 가 0.5(즉 동전 던지는 실험)인 확률변수이다.

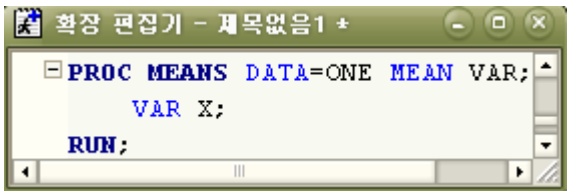
EXAMPLE: 이항 분포

20개의 사지선다형 문제를 찍을 때 맞는 개수 데이터를 10개 뽑아보자(생성). seed는 1로 하자. 사지 선다형 문제를 찍을 때 맞을 확률은 1/4, 틀린 확률은 3/4이고 결과는 맞거나(1) 혹은 틀리거나(0)이다. 그러므로 모수 ($n=20, p=1/4$) 이항분포에서 데이터를 10개 뽑으면 된다.



Obs	I	X
1	1	3
2	2	9
3	3	4
4	4	4
5	5	8
6	6	9
7	7	5
8	8	5
9	9	2
10	10	2

생성한 데이터의 평균과 분산은 얼마일까? 이론적으로는 $np=20*0.25=5, npq=3.75$ 이나 생성 결과는 평균 5.1, 분산은 7.21이다. 평균은 비슷하나 분산은 상당히 높다.



분석 변수 : X

평균값	분산
5.1000000	7.2111111

시드를 3으로 사용하면(X=RANBIN(3,20,1/4);) 다음과 같이 다른 결과를 얻는다. 이처럼 시드에 따라 결과가 다를 수 있다. 그러나 시뮬레이션 할 때는 생성되는 데이터의 개수가 상당히 많으므로 이런 문제는 해결된다.

분석 변수 : X

평균값	분산
5.3000000	3.5666667

EXAMPLE: 포아송 분포 생성

연못에서 작업을 시도하는 회수는 조사하였더니(오전 9시부터 오후 5시) 시간당 평균 4회이고 포아송 분포를 따른다고 하자. 매 30분 조사하였을 때 작업 회수 데이터 10개를 생성하시오. seed는 3으로 하시오. 시간당 평균 4이므로 포아송 분포의 성질에 의해 30분에는 평균 2이다.

확장 편집기 - 제목없...

```
DATA TWO;
    DO I=1 TO 10;
        X=RANPOI(3,2);
    OUTPUT;
    END;
RUN;

PROC PRINT DATA=TWO;
RUN;
```

SAS 시스템

Obs	I	X
1	1	2
2	2	4
3	3	1
4	4	3
5	5	3
6	6	4
7	7	2
8	8	2
9	9	1
10	10	1

EXAMPLE: 주사위 눈금

주사위 던지는 게임을 한다. 열 번 던질 때 나오는 수를 생성해보자. 눈금은 1부터 6까지의 정수이고 각 눈금이 나올 가능성이 동일하므로 표 확률분포를 생성하는 함수를 이용하면 된다. 시드는 3을 사용하였다. 주사위 각 면이 나타날 확률은 1/6로 동일하다. 그러나 10번 던지면 각 면의 상대 빈도(relative frequency) 값은 1/6이 되지 않는다.

```
DATA THREE;
    DO I=1 TO 10;
        X=RANTBL(3,1/6,1/6,1/6,1/6,1/6,1/6);
    OUTPUT;
    END;
RUN;

PROC PRINT DATA=THREE;
RUN;
```

SAS 시스템

Obs	I	X
1	1	4
2	2	6
3	3	2
4	4	5
5	5	5
6	6	6
7	7	4
8	8	4
9	9	1
10	10	2

EXAMPLE: 주사위 눈금 상대 빈도

주사위를 1000번 던져 각 눈금의 상대빈도를 구해보자. 데이터의 상대빈도를 구하는 PROC 단계는 FREQ이다. NOCUM 옵션은 누적 빈도 값을 출력하지 말라는 옵션이다. 1000번 던졌을 때는 각 눈금의 상대 빈도는 1/6=16.7(%)와 다소 차이는 있다. 그러나 더 많이 10,000번쯤 던지면? 상대 빈도가 1/16에 근사 한다.

```
DATA FOUR;
    DO I=1 TO 1000;
        X=RANTBL(3,1/6,1/6,1/6,1/6,1/6,1/6);
    OUTPUT;
    END;
RUN;

PROC FREQ DATA=FOUR;
    TABLE X;
RUN;
```

FREQ 프로시저

X	도수	백분율	누적 도수	누적 백분율
1	173	17.30	173	17.30
2	169	16.90	342	34.20
3	166	16.60	508	50.80
4	141	14.10	649	64.90
5	158	15.80	807	80.70
6	193	19.30	1000	100.00

GEN.sas *

```
DATA GEN1;
     DO I=1 TO 10000;
          X=RANTBL(3,1/6,1/6,1/6,1/6,1/6,1/6);
          OUTPUT;
     END;
RUN;

PROC FREQ DATA=GEN1;
     TABLE X/NOCUM;
RUN;
```

X	빈도	백분율
1	1694	16.94
2	1639	16.39
3	1660	16.60
4	1677	16.77
5	1650	16.50
6	1680	16.80

EXAMPLE: 기대 수익 계산

주사위 눈금에 1,000원을 곱해 상금을 준다. 한번 게임을 할 때 게임 참가비로 3,000원이다. 이 게임의 기대값(expected value)을 계산하시오. 이 게임을 할 때 기대되는 상금은 3500원($=(1+2+3+4+5+6)*1000/6$)이고 참가비가 3000원 이므로 기대 수익은 500원이다. 정말 기대 수익이 500원일까? 게임을 20번 할 때 매 게임 얼마를 따는지(잃는지) 알아보자. 그리고 PROC MEANS를 사용하여 게임의 기대값을 구해보자.

확장 편집기 - 제목없음1 *

```
DATA FIVE2;
     GAME=3000;
     DO I=1 TO 20;
          X=RANTBL(3,1/6,1/6,1/6,1/6,1/6,1/6);
          PRIZE=X*1000;
          TOTAL=PRIZE-GAME;
     OUTPUT;
     END;
RUN;

PROC PRINT DATA=FIVE2;
RUN;

PROC MEANS DATA=FIVE2 MEAN;
     VAR TOTAL;
RUN;
```

Obs	GAME	I	X	PRIZE	TOTAL
1	3000	1	4	4000	1000
2	3000	2	6	6000	3000
3	3000	3	2	2000	-1000
4	3000	4	5	5000	2000
5	3000	5	5	5000	2000
6	3000	6	6	6000	3000
7	3000	7	4	4000	1000
8	3000	8	4	4000	1000
9	3000	9	1	1000	-2000
10	3000	10	2	2000	-1000
11	3000	11	3	3000	0
12	3000	12	4	4000	1000
13	3000	13	4	4000	1000
14	3000	14	2	2000	-1000
15	3000	15	6	6000	3000
16	3000	16	6	6000	3000
17	3000	17	5	5000	2000
18	3000	18	1	1000	-2000
19	3000	19	6	6000	3000
20	3000	20	5	5000	2000

The MEANS Procedure

분석 변수 : TOTAL

평균값
1050.00

20번 게임 했을 때 기대 수익이 1050원이다. 그러나 1000번 게임 했다면? 아래와 같이 521원으로 이론적 기대 수익에 근사 한다.

분석 변수 : TOTAL

평균값
521.0000000

4.5.2 연속형 변수

분포 함수	확률분포함수	SAS 함수
표준 정규분포 (Standard Normal)	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $-\infty<x<\infty$ 평균: μ, 분산: σ^2	**X=RANNOR(seed);** 평균이 μ, 분산이 σ^2 인 정규분포를 따르는 확률변수 생성 $X=\mu+RANNOR(seed)*\sigma$
감마분포 (Gamma)	$f(x)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-x/\beta}$ $0<x<\infty$ 평균: $\alpha\beta$, 분산: $\alpha\beta^2$	**X=RANGAM(seed,** α **);** $\beta=1$ $X\sim RANGAM(\alpha,\beta)$ 인 경우는 **X=** β ***RANGAM(seed,** α **);** $X\sim\chi^2(df=2\alpha)$ 인 경우는 **X=** 2 ***RANGAM(seed,** α **);**
베타 분포 (Beta)	$f(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ $0<x<\infty$ 평균: $\alpha/(\alpha+\beta)$, 분산: $\alpha\beta^2$	**X1=RANGAM(seed,** α **);** **X2=RANGAM(seed,** β **);** **Y=X1/(X1+x2);**
지수분포 (Exponential)	$\alpha=1$인 감마분포. $f(x)=\frac{1}{\beta}e^{-x/\beta}$ 평균: β, 분산: β^2	**X=RANGAM(seed,1);** $\beta=1$ $X\sim Exponential(\beta)$ 인 경우는 **X=RANGAM(seed,1)*** β **;**
카이제곱 분포 (Chi-squared)	$\alpha=r/2$, $\beta=2$ 감마분포. $f(x)=\frac{1}{\Gamma(r/2)2^{r/2}}x^{r/2-1}e^{-x/2}$ 평균: r, 분산: $r/2$	**X=** 2 ***RANGAM(seed,** $r/2$ **);**
T-분포	$T=W/\sqrt{V/r}$, $0<x<\infty$ $f(x)$ 복잡, 자유도 r $W\sim Normal(0,1), V\sim\chi^2(r)$ 평균: 0, 분산: $r/(r-2)$	**Y1=RANNOR(seed);** **Y2=** 2 ***RANGAM(seed,** $r/2$ **);** **X=Y1/SQRT(Y2/R);**
F-분포	$F=\frac{U/r_1}{V/r_2}$, $0<x<\infty$ $f(x)$ 복잡, 자유도 (r_1,r_2) $W\sim\chi^2(r_1), V\sim\chi^2(r_2)$	**Y1=** 2 ***RANGAM(seed1,** $r_1/2$ **);** **Y2=** 2 ***RANGAM(seed2,** $r_2/2$ **);** **X=(Y1/R1)/(Y2/R2);**

	평균: $n/(n-2)$, 분산: 복잡	

EXAMPLE: 정규분포 생성

평균이 80, 분산이 7인 정규분포를 따르는 확률변수 데이터 10개를 생성해 보자.

확장 편집기 - 제목없음1 *

```
DATA ONE;
     DO I=1 TO 10;
          X=80+RANNOR(3)*SQRT(7);
     OUTPUT;
     END;
RUN;

PROC PRINT DATA=ONE;
RUN;
```

SAS 시스템

Obs	I	X
1	1	82.4271
2	2	80.8372
3	3	81.6216
4	4	77.8933
5	5	81.6035
6	6	77.0225
7	7	80.8537
8	8	81.0799
9	9	81.5191
10	10	79.5690

EXAMPLE: T-분포 생성

자유도 20 인 t-분포를 따르는 확률변수 데이터 10개를 생성하시오. 변수명은 T 로 하자.

```
확장 편집기 - 제목없음1 *
DATA TWO;
    DO I=1 TO 10;
        X1=RANNOR(3);
        X2=2*RANGAM(3,20/2);
        T=X1/SQRT(X2/20);
    OUTPUT;
    END;
RUN;

PROC PRINT DATA=TWO;
RUN;
```

SAS 시스템

X1	X2	T
0.91736	14.5405	1.07588
0.61290	22.5737	0.57690
0.60608	19.2413	0.61792
0.32266	28.2264	0.27161
-0.16289	10.8371	-0.22128
0.57160	18.7015	0.59111
0.05951	25.6622	0.05254
-1.82260	27.2420	-1.56166
0.56330	13.9714	0.67397
-2.55059	19.9923	-2.55108

EXAMPLE: F-분포 생성

자유도 (5,7) F-분포를 따르는 확률변수 데이터 10개를 생성하시오. 변수명은 F 로 하자.

```
확장 편집기 - 제목없음1 *
DATA THREE;
    DO I=1 TO 10;
        X1=2*RANGAM(3,5/2);
        X2=2*RANGAM(3,7/2);
        F=(X1/5)/(X2/7);
    OUTPUT;
    END;
RUN;

PROC PRINT DATA=THREE;
RUN;
```

SAS 시스템

I	X1	X2	F
1	5.5826	3.96939	1.96898
2	6.5095	3.44106	2.64841
3	4.5958	7.72659	0.83273
4	10.5943	2.35255	6.30463
5	7.2167	6.21184	1.62646
6	8.6085	3.82201	3.15328
7	2.2879	1.88341	1.70065
8	4.9958	9.92015	0.70504
9	2.0892	7.39972	0.39527
10	5.2917	4.12016	1.79807

EXAMPLE: 백분위 함수 사용하기, 모비율 검정

평균이 2인($\beta=2$) 지수분포를 생성하고 히스토그램을 그려 보자. MIDPOINT 옵션은 막

대 눈금은 중간 크기이다. CFILL는 막대 안의 색깔 지정 옵션이다. EXPONENTIAL은 이론적 지수분포의 확률밀도 함수를 그리라는 옵션이다. 모수 β 는 데이터 100개로부터 추정된 평균으로 추정된다.

확장 편집기 - 제목없음1 *

```
DATA ONE;
     DO I=1 TO 100;
          X=RANEXP(3)*2;
     OUTPUT;
     END;
RUN;

PROC UNIVARIATE DATA=ONE;
     VAR X;
     HISTOGRAM X/MIDPOINTS=0 TO 7 BY 0.2
                 CFILL=BLUE EXPONENTIAL;
RUN;
```