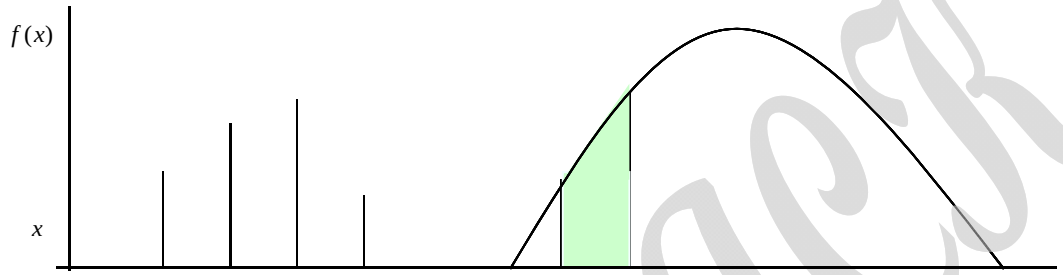
# 확률 모형(probability model)이란

확률 모형이란 확률 변수의 확률 분포 함수(probability density function)를 의미한다. 확률 분포 함수는 확률 변수($x$)가 가질 수 있는 각 값을 정의역(domain) 확률($f(x)$)을 치역으로(range) 한 함수이다. 아래 그림은 이산형 확률 밀도 함수와 연속형 확률 분포(밀도) 함수의 예이다. 이산형 경우에는 막대의 높이가 (히스토그램에서는 바의 높이) 연속형인 경우에는 면적이 확률이다. 연속형 확률 분포 함수에서는 $x$의 한 값에서 확률은 0 이다.




## 모집단 분포 가정

다음은 모집단의 분포 형태 $f(x)$을 가정하는 예제이다.

(1) 모집단 전체를 조사한 경우 전체 자료로부터 히스토그램을(이것이 확률 분포 함수이다. 물론 정확한 함수 식은 알 수 없지만) 그리거나 상대 빈도 개념으로 관심 구간의 확률을 구할 수 있다. (예) 한남대학교 학생 중 용돈이 30,000(원)~35,000(원)인 학생의 비율(확률)은? 전체 12,000 명 중 용돈이 이 구간에 속하는 학생 수가 확률이 된다.

(2) 자료를 시뮬레이션(simulation: 모집단 가정) 할 때 사용한다.

(3) 회귀분석이나 분산 분석에서 오차항에 대한 정규 분포 가정이 있다. 이 가정이 무너지면 회귀 계수(t-검정), 모형의 유의성(F-검정) 검정이 불가능하다.

(4) 소표본(표본의 크기 n<20~30)일 경우 모평균에 대한 가설 검정 시 모집단은 정규 분포임을 가정한다.

## 모집단 분포 가정이 불가능한 경우

모집단의 분포를 가정하지 않으면 어떻게 확률 분포를 알 수 있나? 표본 자료의 분포(즉 표본 분포)를 이용할 수 밖에 없다.

(1) (Goodness-of-fits:적합성 검정) : 빈도표 활용 검정통계량 방법 $\chi^2$-검정

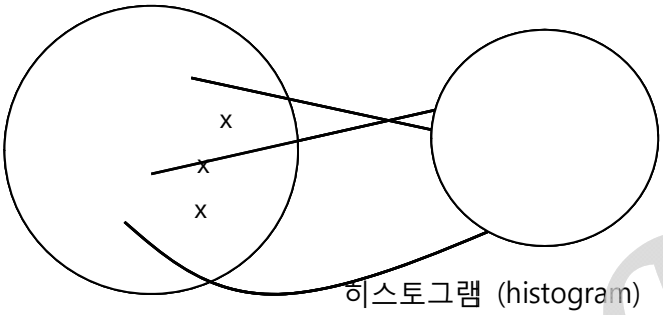
(2) 그래프 이용 방법: P-P plot 방법, 시각적 방법, rule of thumb



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

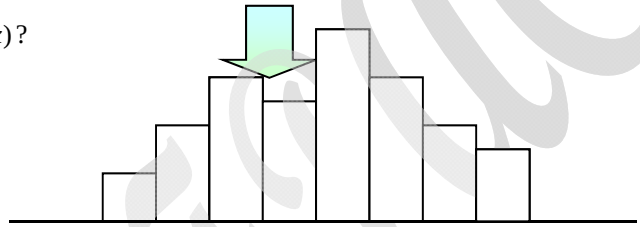
## 방법 1: Frequency table (빈도표) 개념 이용하기 ~ $\chi^2$ 분포 적합성 검정

표본 자료로부터 모집단 확률 밀도 함수를 어떻게 구할 수 있는가? 히스토그램으로부터 확률분포함수를 구한다? 불가능하다. 접근 방법은 표본 자료로부터 빈도표를(이를 관측 빈도) 만들고 (histogram 과 동일) 모집단이 따를 것 같은 분포로부터(예:정규분포) 빈도표(이를 기대 빈도)를 만들어 비교하면 빈도의 차이가 거의 없으면 모집단은 그 분포를 따른다고 하자 그렇지 않으면 기각한다.



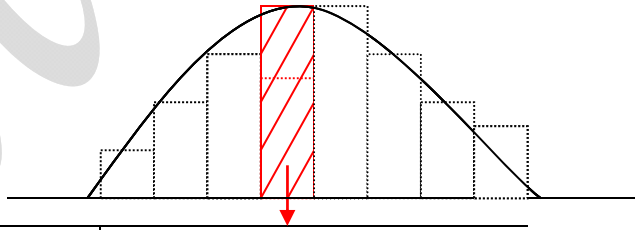

히스토그램 (histogram)

모집단(population) $f(x)$?



| 구간 | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ |
|---|---|---|---|---|---|---|---|---|
| 관측 도수 | $O_1$ | $O_2$ | $O_4$ | $O_4$ | $O_5$ | $O_6$ | $O_7$ | $O_8$ |

모집단의 분포가 $f(x)$가 정규분포를 (연속형 분포) 따를까?



| 구간 | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ |
|---|---|---|---|---|---|---|---|---|
| 기대 도수 | $E_1$ | $E_2$ | $E_4$ | $E_4$ | $E_5$ | $E_6$ | $E_7$ | $E_8$ |

표본 분포가 설정한 모집단 분포와 동일하다면 관측 도수와 (observed frequency) 기대 도수는 (expected frequency) 비슷한 값일 것이다. 즉 $O_1 \approx E_1$, $O_2 \approx E_2$, …, $O_k \approx E_k$ (위 예에서는 $k=8$)

검정통계량 (test statistics) ? $TS = \frac{\sum_{i=1}^{k}(O_i - E_i)^2}{E_i} \sim \chi^2(df = k - c - 1)$ c=모수 추정 개수

## 방법 2 : Probability Plot

(1) 두 데이터의 실증적 empirical 분포함수는 동일한가? (2) 이론적 분포함수와 데이터의 분포함수는 동일한가? 보여주는 시각적 그래프

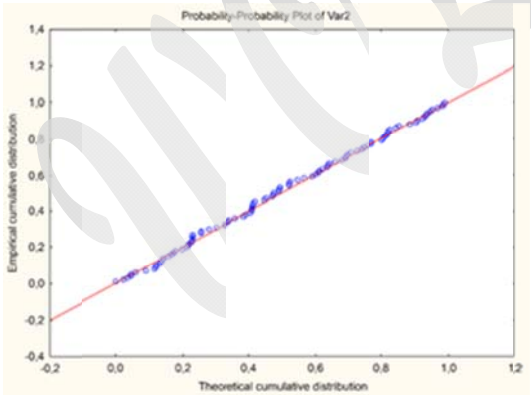
Probability-Probability plot, Quantile-Quantile plot

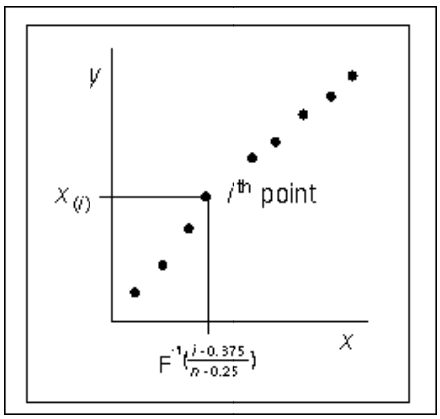
### Probability-Probability plot

두 데이터의 누적분포함수를 2 차원 그래프에 표현함, Y-축에는 한 데이터의 누적분포함수 $F(x_{(i)})$, Y-축에는 (1) 다른 데이터 (혹은 이론적 분포) 누적분포함수 $F(y_{(i)})$ (2) 이론적 분포의 $F(y_{(i)})$

### Quantile-Quantile plot

두 데이터의 누적분포함수를 2 차원 그래프에 표현함, X-축에는 한 데이터의 p-백분위 값 $F^{-1}(p_i = (i-0.5)/n)$, Y-축에는 (1) 다른 데이터 p-백분위 값 $F^{-1}(p)$, (2) 데이터의 순서통계량 $x_{(i)}$






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

## 이산형 확률분포

### ‖ 일반분포 ‖

주사위로 게임을 하려고 주사위를 하나 샀다. 이 주사위 각 면이 나올 확률이 동일한지 (fair) 알아보기 위하여 실험을 하기로 하였다. 주시위를 1,000 번 던져 다음 결과가 나왔다.

| 눈금 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 빈도 | 150 | 160 | 165 | 155 | 170 | 200 |

귀무가설: 각 눈금이 나올 확률은 모두 1/6 로 같다.

X=주사위 눈금 ➔ $f(x)=1/6 \quad \text{for } x=1,2,...,6$

대립가설: 각 눈금이 나올 확률이 모두 1/6 은 아니다.

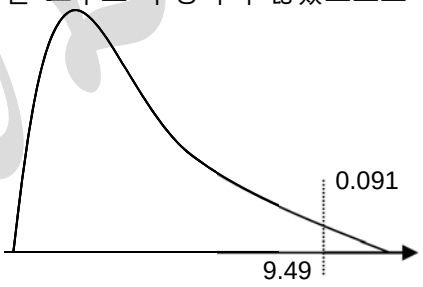
주사위 눈금 X 는 귀무가설의 확률 분포 $f(x)$ 를 따르지 않는다.

검정통계량

| 눈금 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 관측 빈도 $(O_i)$ | 150 | 160 | 165 | 155 | 170 | 200 |
| 기대 빈도 $(E_i)$ | 166.7 | 166.7 | 166.7 | 166.7 | 166.7 | 166.7 |

검정통계량: $T=\frac{(150-166.7)^2}{166.7}+\frac{(160-166.7)^2}{166.7}+...+\frac{(200-166.7)^2}{166.7}=9.49 \sim \chi^2(df=6-1)$

기대 빈도 계산을 위하여 어떤 모수도 추정하지 않았으므로 $c=0$ 이다.




결론: p-값이 0.091 이므로 귀무가설을 기각하지 못한다. (주사위는 공정하다)

```
x=c("1"=150,"2"=160,"3"=165,"4"=155,"5"=170,"6"=200)
H0=c(1/6,1/6,1/6,1/6,1/6,1/6)
x.chi=chisq.test(as.table(x),p=H0)
x.chi
x.chi$expected
```

```
        Chi-squared test for given probabilities

data:  as.table(x)
X-squared = 9.5, df = 5, p-value = 0.09071
```

**(빈도표 출력하기)**

```
> as.table(x)
  1   2   3   4   5   6
150 160 165 155 170 200
> prop.table(as.table(x))
    1     2     3     4     5     6
0.150 0.160 0.165 0.155 0.170 0.200
```

마케팅 활용

햄버거 시장 점유율이 M 사 50%, B 사 35%, L 사 15%라고 조사기관이 발표하였다. 이에 대한 검증을 위하여 한남대학교 학생 200 명을 무작위 조사한 결과 M 사 햄버거 80 명, B 사 80 명, L 사 40 명이 주로 사 먹는다고 응답하였다. 조사기관의 주장을 유의수준 5%에서 검정하시오.

공정 주사위

6 이 다른 숫자에 비해 2 배 잘 나오도록 디자인한 주사위가 있다. 이를 500 번 던지는 실험 데이터를 시뮬레이션 하시오. (rmultinom 함수 이용) 그리고 다음 스크립트 (이항분포 n=5, p=1/6 인 확률변수 데이터를 450 번 실험하여 얻은 데이터의 빈도 표를 얻었다)를 활용하여 디자인한 주사위대로 데이터가 시뮬레이션 되었는지 검증하시오.

```
> x=rbinom(450,5,1/6)
> xt=table(x)
> as.table(xt)
x
  0   1   2   3   4
175 189  68  15   3
```



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

## Ⅱ 이항분포 Ⅱ

남녀 출산 비율이 0.5 인지 알아보기 위하여 아이들이 3 명이 1,000 가구를 대상으로 남자 아이의 수를 조사하여 다음 표를 얻었다.

| 남자 아이 수 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 빈도 | 100 | 350 | 400 | 150 |

귀무가설: 남자 아이 수는 이항 분포(n=3, p=0.5)를 따른다.

X=남자 아이 수 ➔ $f(x)=\binom{3}{x}(0.5)^x(1-0.5)^{3-x}$ for $x=0,1,2,3$

대립가설: X 는 이항 분포를 따르지 않는다.

| 남자 아이 수 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 관측 빈도 $(O_i)$ | 100 | 350 | 400 | 150 |
| 기대 확률 | 0.125 | 0.375 | 0.375 | 0.125 |
| 기대 빈도 $(E_i)$ | 125 | 375 | 375 | 375 |

기대 빈도는 귀무가설이 맞다는 가정 하에서 계산한다.

$$f(x=0)=\binom{3}{0}(0.5)^0(1-0.5)^3=0.125\ ,\quad f(x=1)=\binom{3}{1}(0.5)^1(1-0.5)^3=0.375$$

검정통계량: $T=\frac{(100-125)^2}{125}+\ldots+\frac{(150-125)^2}{125}=13.3\sim\chi^2(df=4-1)$

기대 빈도 계산을 위하여 어떤 모수도 추정하지 않았으므로 $c=0$ 이다.

결론: p-값이 0.004 이므로 귀무가설을 기각한다. (혹은 검정 통계량 값 13.3 이 임계치 7.82 보다 크므로) 그러므로 남자 아이의 수는 성공 확률이 0.5 인 이항 분포를 따르지 않는다.

```
x=c("0"=100,"1"=350,"2"=400,"3"=150)
H0=c(dbinom(0,3,0.5),dbinom(1,3,0.5),dbinom(2,3,0.5),dbinom(3,3,0.5))
x.chi=chisq.test(as.table(x),p=H0)
x.chi
x.chi$expected
prop.table(as.table(x))
```

```
data:  as.table(x)
X-squared = 13.3333, df = 3, p-value = 0.003969

> x.chi$expected
  0   1   2   3
125 375 375 125
> prop.table(as.table(x))
   0    1    2    3
0.10 0.35 0.40 0.15
```

제품 검증 불량개수.csv

A 회사는 자사 제품의 불량률이 5%라 한다. 이 회사는 제품을 20 개 단위로 상자에 넣어 판매한다. 다음은 200 개 상자를 임의 추출하여 각 상자의 5 개 제품의 불량의 개수를 조사한 데이터이다. 제품의 불량률을 5%라 할 수 있나? 유의수준 5%에서 검정하시오.

## Ⅱ 포아송분포 Ⅱ

다음은 한남대학교 정문을 통과하는 차량의 수가 Poisson 분포를 따르는지 알아보기 위하여 1 분마다 차량 통과 회수를 300 회 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 모수를 ($\lambda$) 알아야 한다.

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

```
x=c("0"=20,"1"=54,"2"=74,"3"=67,"4"=45,"5"=25,"6"=11,">=7"=4)
n=c(0,1,2,3,4,5,6,7)
lm=sum(n*x)/sum(x)
> lm
[1] 2.673333
```

표본 자료로부터 모수 p 의 추정치를 ($\hat{\lambda}$) 구하면 $\hat{\lambda} = (0\times 20 + 1\times 54 + \ldots + 7\times 4)/300 = 2.67$

그러므로 기대 확률과 기대 빈도는 다음 Poisson 확률 분포에 의해 계산하면 된다.

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

| 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 | 0.019 |
| 기대 빈도 | 20.7 | 55.5 | 74.1 | 66 | 44.1 | 23.4 | 10.5 | 5.89 |

검정 통계량: $T = \frac{(20-20.7)^2}{20.7} + \frac{(54-55.5)^2}{55.5} + \ldots + \frac{(4-5.89)^2}{3.9} = 0.8107$

결론: 표로부터 임계치는 $\chi^2$(자유도=8-1-1=6, $\alpha = 0.05$) =12.59 이므로 귀무가설이 채택되고 이 자료는 Poisson 분포를 따른다고 할 수 있다. (자유도 계산 시 1 을 더 빼 주는 이유는 포아송 분포의 모수 $\lambda$ 를 알지 못하므로 자료를 이용하여 추정하였기 때문이다.

```
H0=c(dpois(0,lm),dpois(1,lm),dpois(2,lm),dpois(3,lm),
dpois(4,lm),dpois(5,lm),dpois(6,lm),1-ppois(6,lm))
x.chi=chisq.test(as.table(x),p=H0)
x.chi
```

```
data:  as.table(x)
X-squared = 0.8107, df = 7, p-value = 0.9973
```

실제 평균을 추정한 정보가 없어 R 은 유의확률을 자유도 7 에 의해 계산하므로 아래와 같이 구해야 한다.

```
> 1-pchisq(x.chi$statistic,df=6)
X-squared
0.9917849

> x.chi$expected
       0        1        2        3        4        5        6      >=7
20.70653 55.35546 73.99180 65.93491 44.06650 23.56089 10.49768  5.88623
```

고객 수 콜회수.csv

일반적으로 은행 고객 콜 센터에 걸려오는 전화 회수는 1 시간당 평균 52 회가 오고 포아송분포를 따른다고 한다. 우리은행 고객 콜 센터도 동일한 패턴을 갖는지 알아보기 위하여 최근 80 시간 동안 걸려온 고객 전화회수를 (시간당) 조사한 데이터이다. 포아송분포를 따르는지 알아보고, 95% 신뢰구간을 구하시오.

## 연속형 확률변수 P-P plot

### Probability-Probability (P-P) Plot

- o 데이터의 분포가 임의의 이론적 분포를 따르는지 알아보는 시각적 도구
- o 이론적 분포에 따르면 직선을 보임
- o $(F(x_{(i)}), (i-0.5)/n)$ 산점도 표현

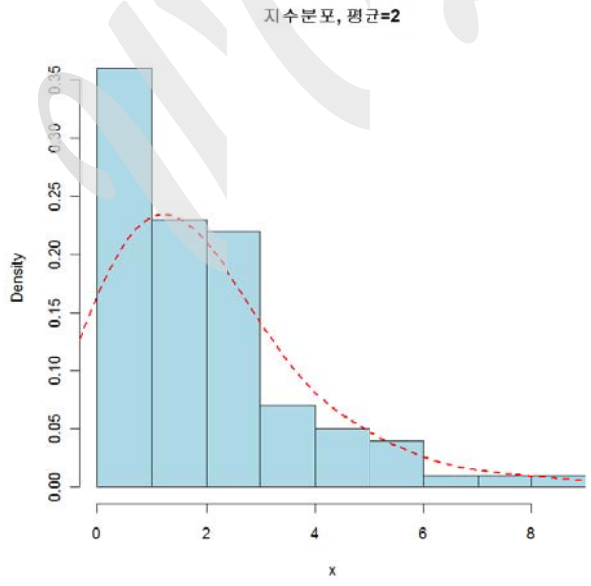
### Quantile-Quantile (QQ) Plot

- o 두 표본 데이터의 분포가 서로 동일한지 알아보는 시각적 도구
- o 동일한 분포인 경우 직선을 보임
- o $((x_{(i)},), F^{-1}((i-0.5)/n))$ 산점도 표현

### 확률분포함수

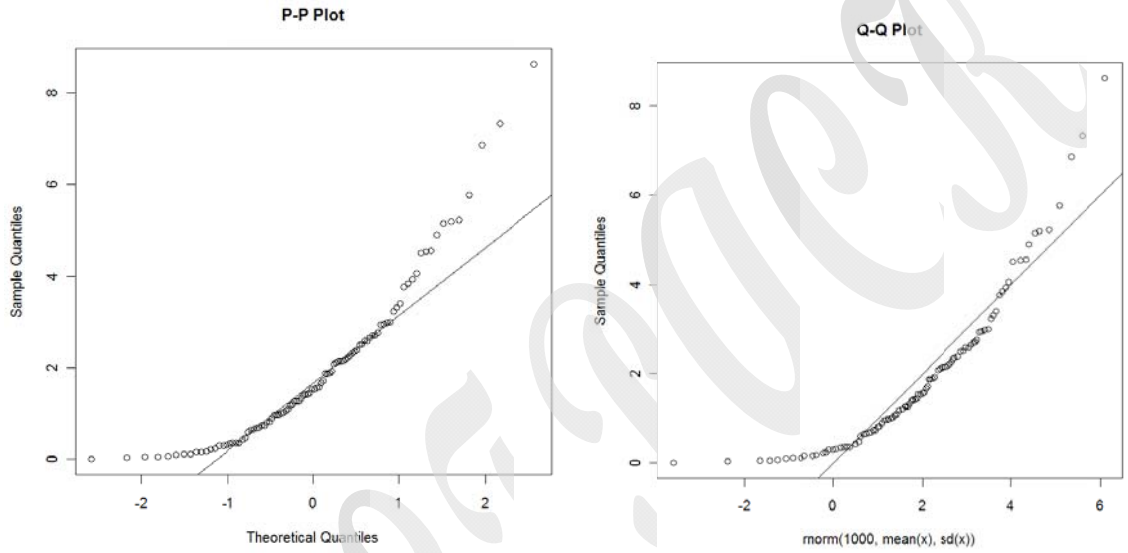
(평균이 2 인 지수분포 생성)

```
x=rexp(100,0.5)
hist(x,xlim=c(min(x),max(x)),
      probability=T, nclass=10,
     col='lightblue',
     main='지수분포, 평균=2')
lines(density(x,bw=1), col='red', lwd=2, lty=2)
```



## P-P plot / Q-Q plot

```
#P-P Plot
 qqnorm(x,main="P-P Plot")
 qqline(x)

#Q-Q Plot
qqplot(rnorm(1000,mean(x),sd(x)), x, main="Q-Q Plot",
   ylab="Sample Quantiles", xlab="teoretical Q")
abline(0,1)
```




## (정규성 검정)

```
> ad.test(x)

        Anderson-Darling normality test

data:  x
A = 3.0329, p-value = 1.146e-07
```

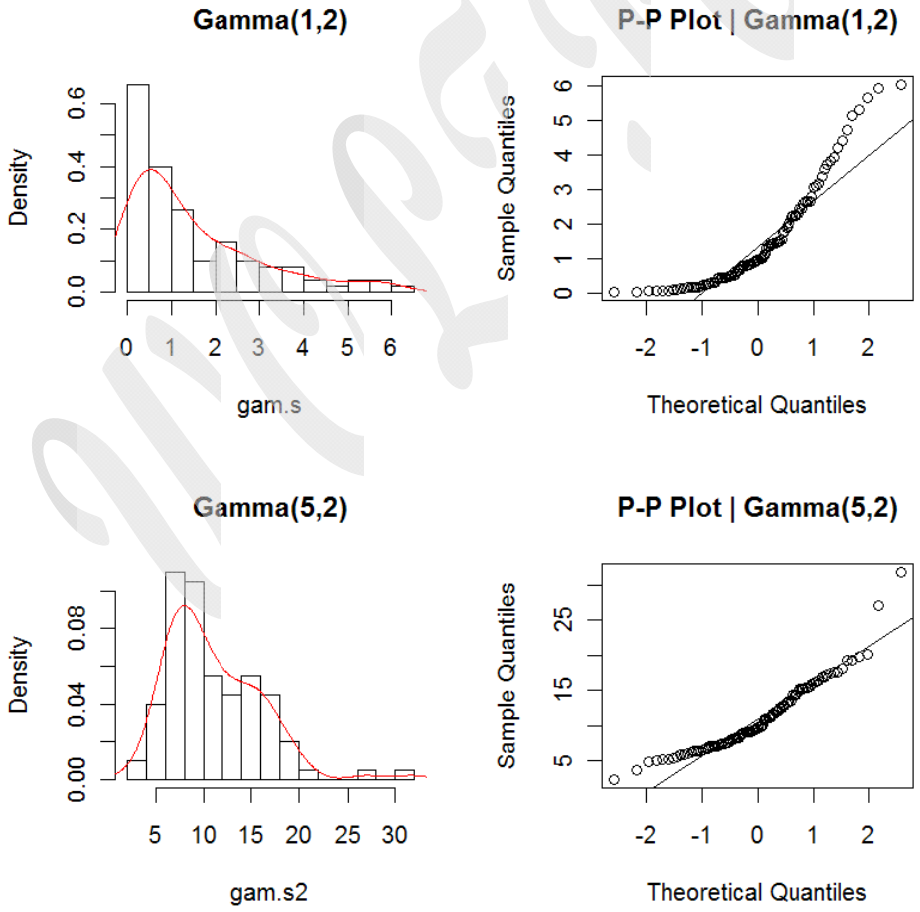
P-P plot 그리기 (Cancer 데이터)

1. (blad, lung, kid, leuk) 데이터가 정규분포를 따르는지 P-P plot 을 그려 시각적으로 확인하자. (물론 정규성 검정도 하시오) 두 함수 모두 사용
2. (blad, lung) 두 데이터의 분포가 동일한지 Q-Q plot 을 그려 시각적으로 판단해 보자.

## Example : Normal Plot

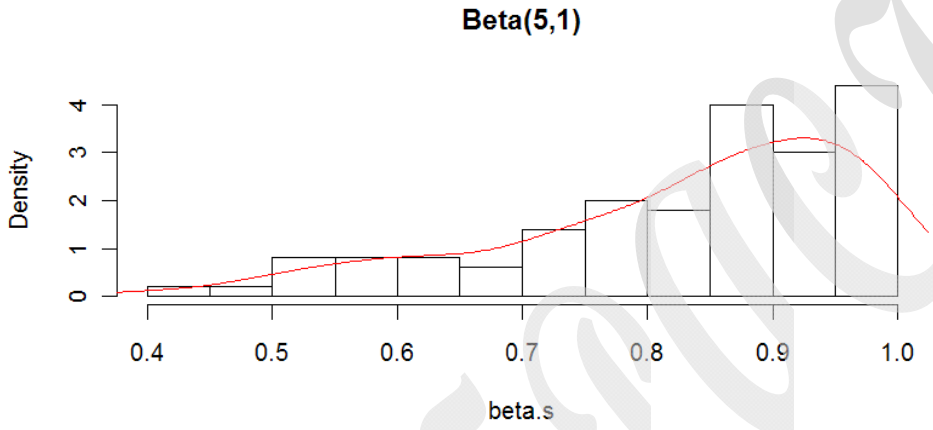
### 우로 치우침 (긴 꼬리 / 짧은 꼬리)

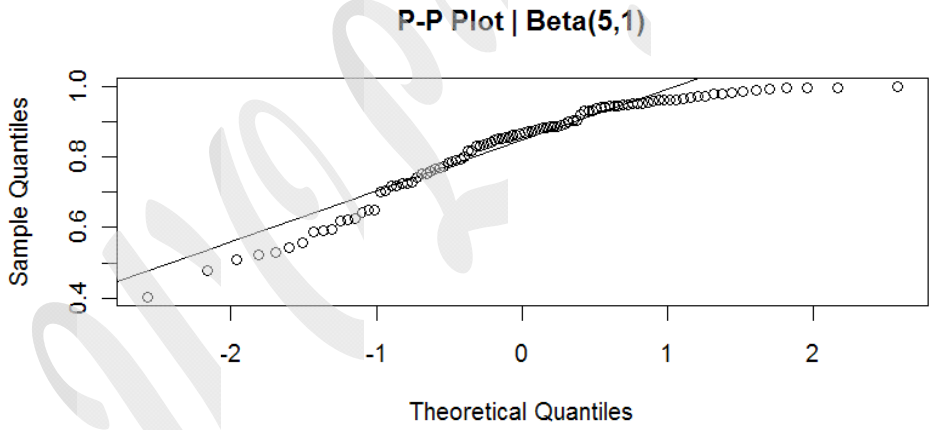
```
#Gamma(1,2)
gam.s=rgamma(100,1,0.5)
split.screen(c(2,2))
screen(1)
hist(gam.s,nclass=12, probability=T, main="Gamma(1,2)")
lines(density(gam.s), col='red')
screen(2)
qqnorm(gam.s,main="P-P Plot | Gamma(1,2)")
qqline(gam.s)

#Gamma(5,2)
gam.s2=rgamma(100,5,0.5)
screen(3)
hist(gam.s2,nclass=12, probability=T, main="Gamma(5,2)")
lines(density(gam.s2), col='red')
screen(4)
qqnorm(gam.s2,main="P-P Plot | Gamma(5,2)")
qqline(gam.s2)
```



## 좌로 치우침

```
#Beta(5,1)
beta.s=rbeta(100,5,1)
split.screen(c(2,1))
screen(1)
hist(beta.s,nclass=12, probability=T, main="Beta(5,1)")
lines(density(beta.s), col='red')
screen(2)
qqnorm(beta.s,main="P-P Plot | Beta(5,1)")
qqline(beta.s)
```






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