Probability Density Function (확률분포함수) f(x)

확률변수 X가 가질 수 있는 값과 그에 대응하는 확률 값을 표, 그래프, 수식으로 표현한 함수 f(x)=P(X=x)

(예1) 당첨이 하나 있는 상품권 박스(5장)에서 5명이 하나씩 뽑는 실험에서, "나"는 몇 번째 뽑는 것이 유리할까? 확률밀도함수를 구하여 논리를 제시하라.

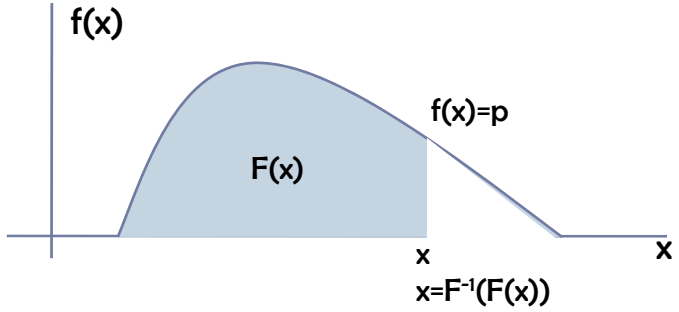
(예2) 취업률이 0.4인 학과 졸업생 10명을 임의 추출하여 조사하였을 때, 취업한 학생 수에 대한 확률밀도함수?

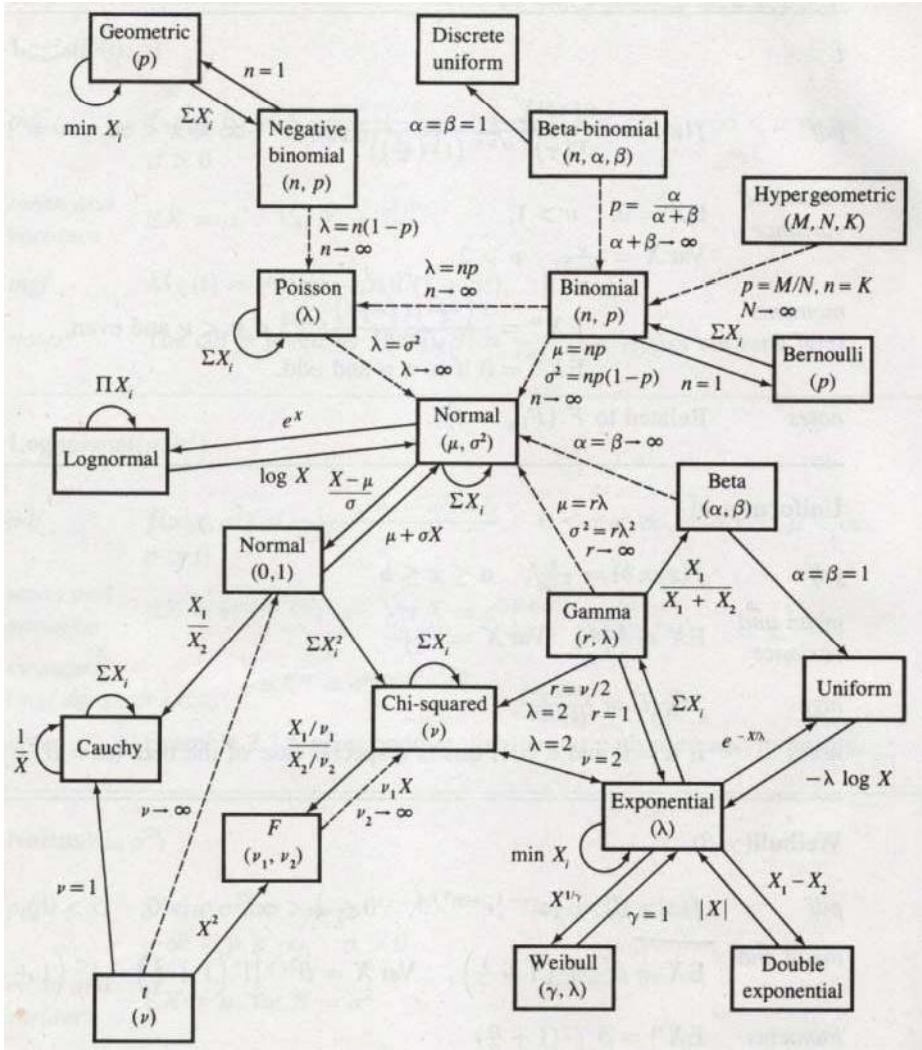
Cumulative Density Function (누적분포함수) F(x)

임의 값 x까지의 확률을 누적한 값, F(x)=P(X≤x)

(예1) F(-∞)=0 , F(∞)=1

(예2) z~표준정규분포, F(z=0)=Φ (z=0)=1/2





확률변수 생성(generating) in Software

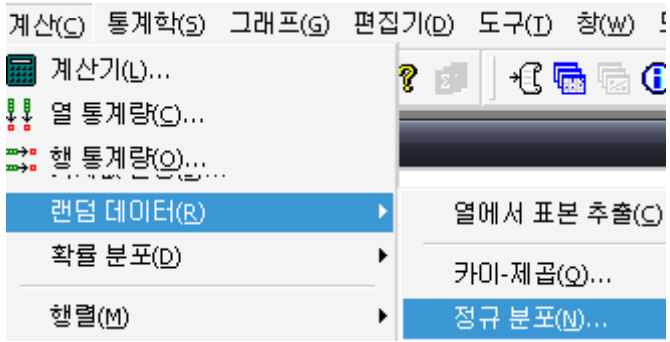
생성 데이터

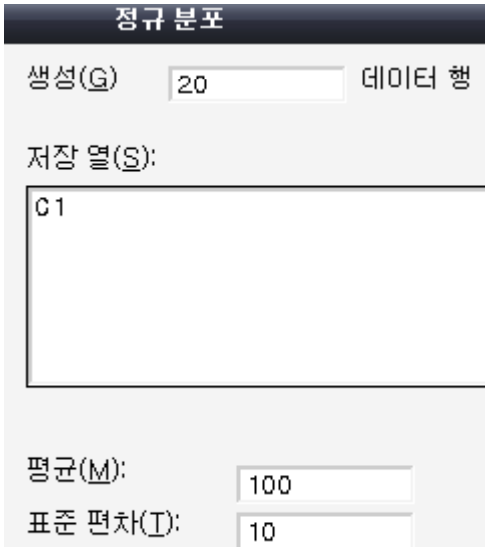
- X1 ~ Normal(μ=100, σ=10), n=20
- X2 ~ Exponential(μ=100), n=50

In SPSS

- 균일분포(uniform distribution)를 생성하여 이를 이용하여 난수를 생선해야 한다. ▶ tedious
  - In Excel: 데이터=>데이터분석=>난수생성 메뉴에서 가능하나 균일(일양), 정규, 이항, 포아송 정도만 가능

In Minitab





	C1
	정규분포
1	113.168
2	95.973
3	101.953
4	98.454
5	84.639
6	101.130
7	112.609

- 데이터 저장: *.mtw
- Output 저장: *.txt (세션 창) 그래프 저장 (*.mgf)

http://wolfpack.hannam.ac.kr
한남대학교 통계학과 권세혁교수

확률변수 생성(generating) in Software (2)

In R

```
> x<-rnorm(20,mean=100,sd=10)
> x
 [1]  91.37400 111.27882 100.40921 10
 [8] 105.63150 111.77338  97.82292  9
[15] 115.48399  91.49702  90.91295  9
> x[1:5]
[1]  91.37400 111.27882 100.40921 104

> x2<-rexp(50,rate=1/100)
> x
 [1]  91.37400 111.27882 100.40921
 [8] 105.63150 111.77338  97.82292
[15] 115.48399  91.49702  90.91295
```

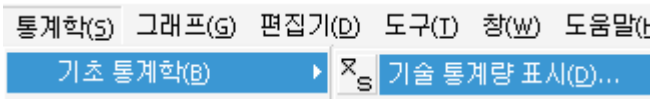
- rbeta, rchisq, rf, rt, rgamma, rgeom, rpoisson, rt, rweibull

확인 작업

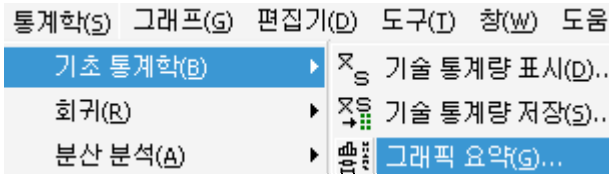
- 기초통계량 구하기
  - 평균과 표준편차 구하기
- 그래프 요약
  - 히스토그램, 나무상자 그림
- 적합성 검정 (Goodness of Fits)

In Minitab

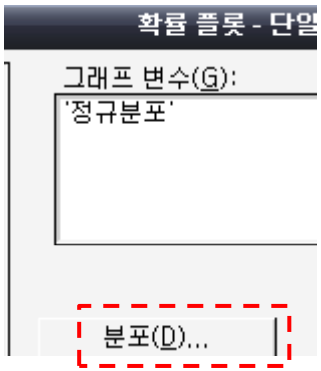
- 기초통계량



- 그래프 요약



- 정규성 검정

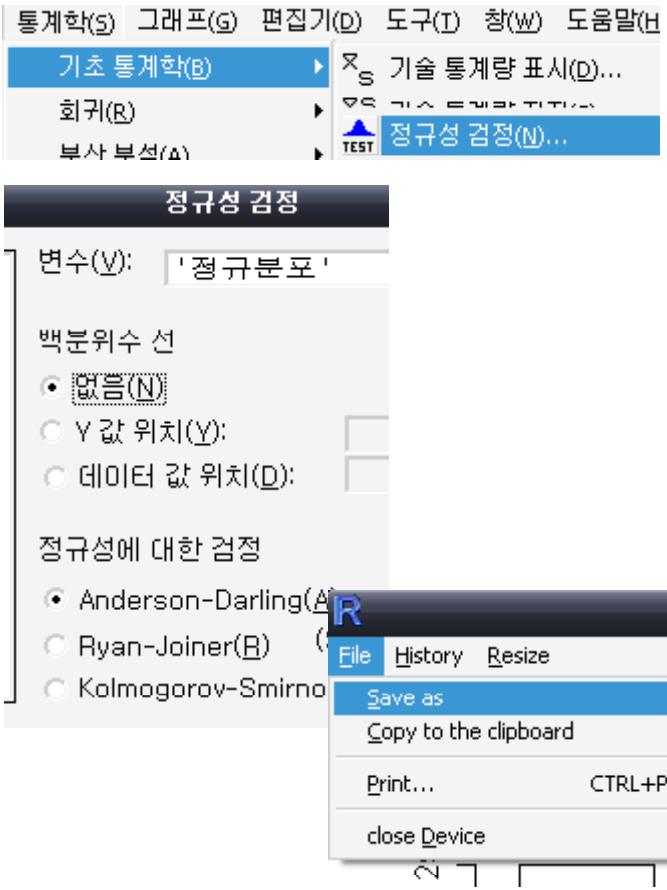


http://wolfpack.hannam.ac.kr
한남대학교 통계학과 권세혁교수

확률변수 생성(generating) in Software (3)

In Minitab (계속)

- 정규성 검정만



In R

- 기초 통계량

```
> summary(x2)
> mean(x2)
[1] 96.94138
> sd(x2)
[1] 98.17234
```

- 그래프 요약

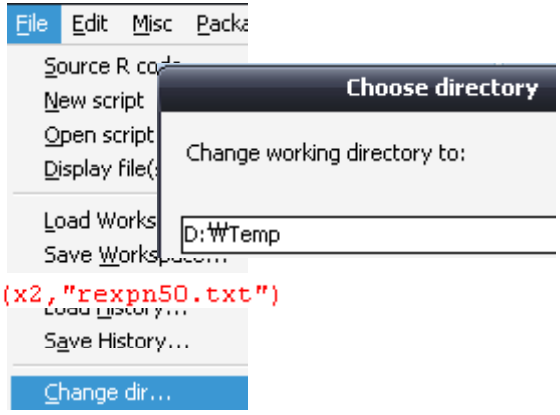
```
> hist(x2)
> boxplot(x2)
```

- Q-Q plot

```
> qqnorm(x2)
```

저장하기 in R

- Data file:



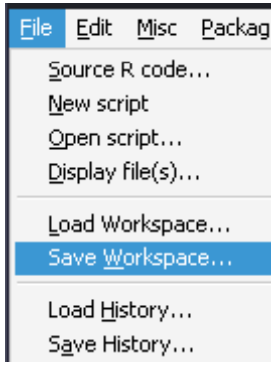

```
> write.table(x2,"rexpn50.txt")
```

- 그래픽 파일

http://wolfpack.hannam.ac.kr
한남대학교 통계학과 권세혁교수

In R

- 데이터 저장하기
  - Work space에 저장



- 명령어들: history에 저장
- 파일 합쳐 저장하기

```
> d2<-list(x=m1,y=m2)
> d2
$x
[1]  1  2  3

$y
[1]  4  5  6

> hist(d2$y)
> write.table(d2,"2.txt")
```

실습

- 평균 60, 표준편차 10인 정규분포를 따르는 확률변수 데이터 100개 생성하여 Z에 저장하시오.
- 평균 60, 표준편차 10인 감마분포를 따르는 확률변수 데이터 100개 생성하여 G에 저장하시오. (α=shape β=scale 모수이며 평균은 $\alpha\beta$, 분산은 $\alpha\beta^2$이다)

In Minitab

- 기초통계량을 구하고 비교하시오.
- 그래프 요약을 그려 비교하시오.
- 적합성 검정을 하시오. (감마분포 따르나?)
  - 결과를 09182007이름.MPJ에 저장하시오.

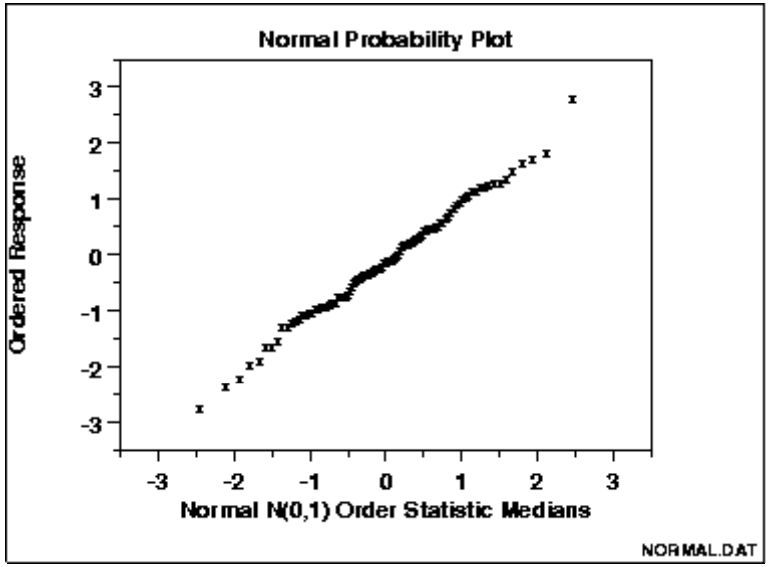
In R

- 기초통계량을 구하고 비교하시오.
- 그래프 요약을 그려 비교하시오.
  - 데이터를 합쳐(데이터 명 all) 09182007.txt에 저장하시오.
  - 결과를 09182007.rdata에 저장하시오.
  - 명령 요약을 09182007.rhistory에 저장하시오.

P-P Plot and Q-Q plot

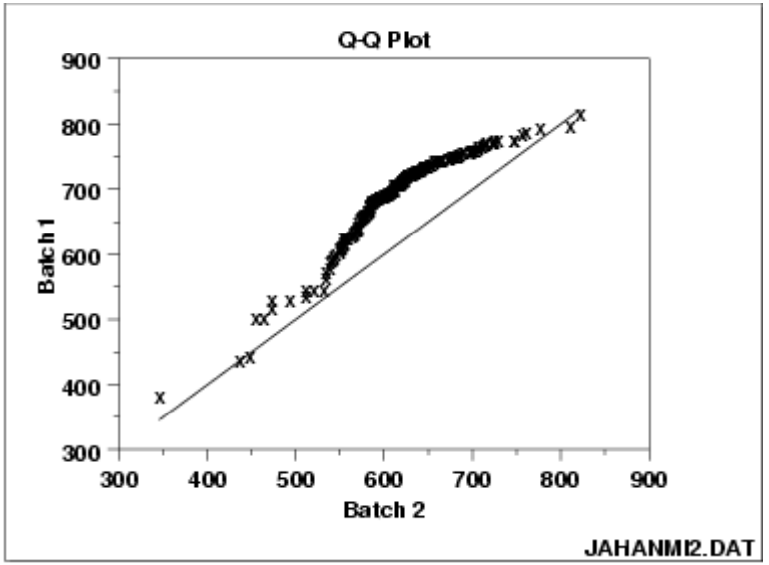
Probability-Probability Plot

- 표본 데이터의 분포 진단, eyeball method
- 검정통계량: Goodness of fits 적합성 검정 (χ^2 검정)
  - Y-축: (i-0.5)/n 혹은 i/(n+1), 혹은 $x_{(n)}$
  - X-축: $F^{-1}(i/(n+1))$: 이론분포의 관측치 값
  - (예) 최대값이 이상치(?), 중간에 높고 낮음이 있음 (확률분포함수 그래프는?)



Quantile - Quantile plot

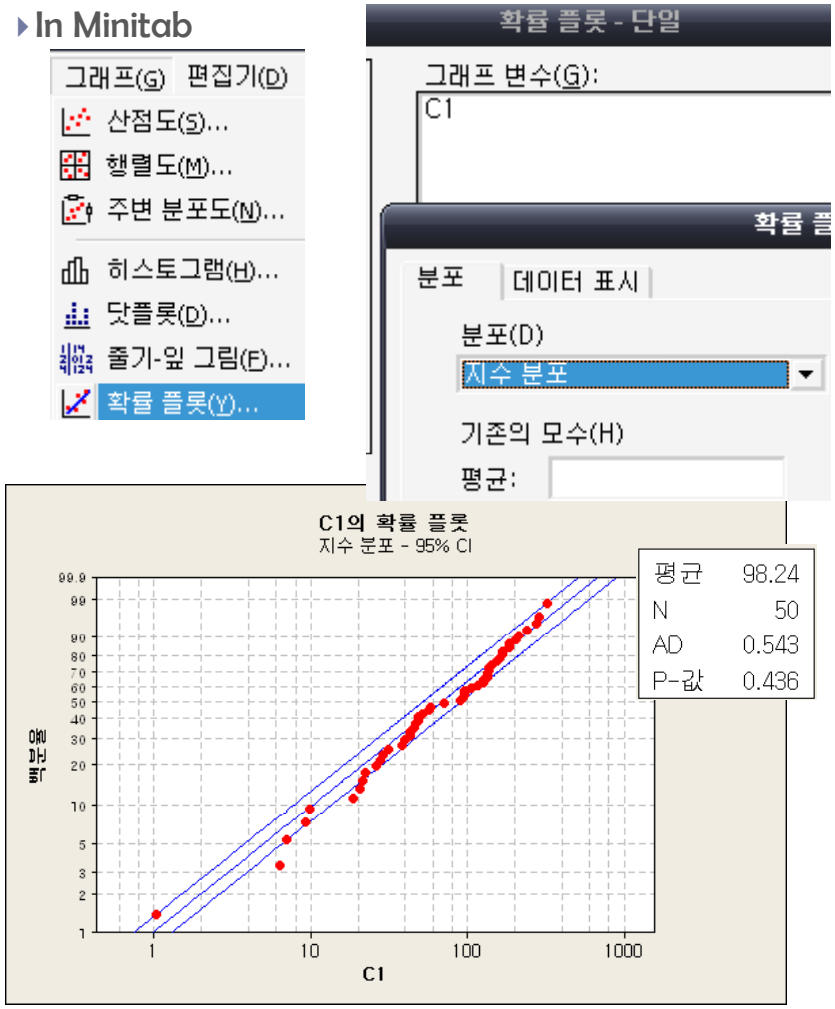
- 두 집단 분포 비교, x-y축에는 순서 통계량
- (예)중앙 이동, 분포모양 변화, 이상치 발생.



http://wolfpack.hannam.ac.kr
한남대학교 통계학과 권세혁교수

P-P Plot and Q-Q plot (in software)

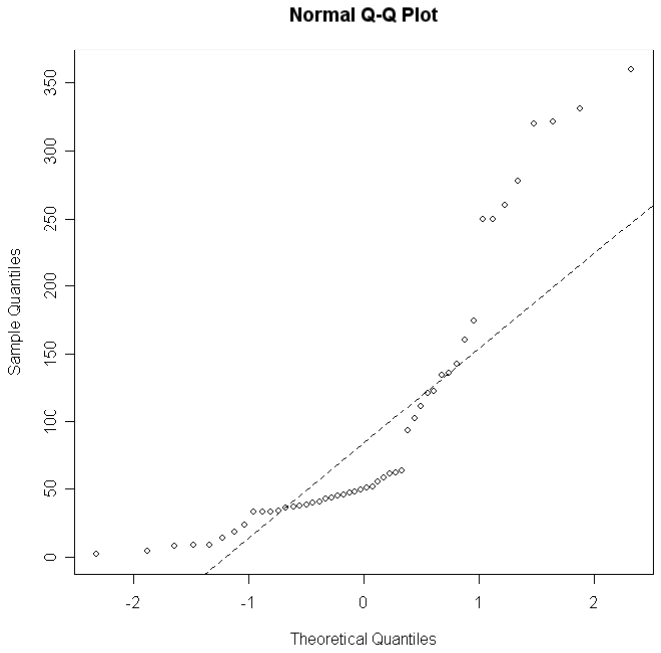
In Minitab



In R

- 정규분포 따르는가?

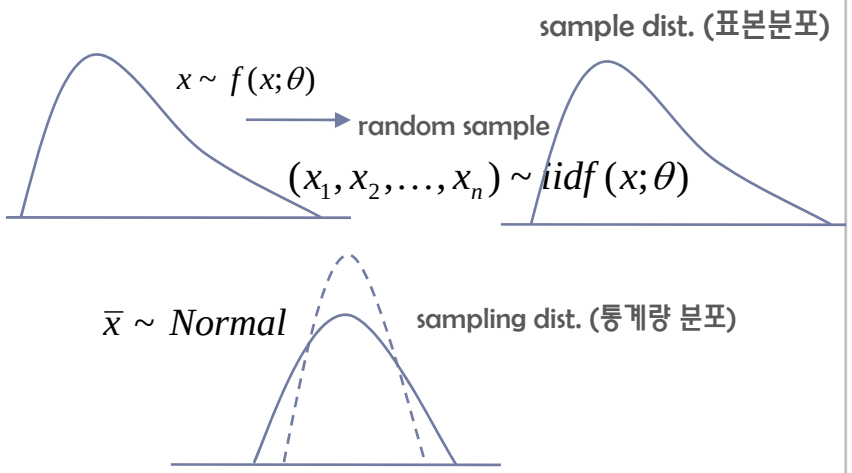
```
> qqnorm(x2)
> qqline(x2,lty=2)
```



- 정규분포 이외 분포는 다소 복잡 (skipped here)

Central Limit Theorem

모집단의 분포가 어떠하든지 표본의 크기(n)가 충분히 크다면 (n>20~30, 대표본, large sample) 표본평균 분포는 정규분포에 근사(approximation)한다.



표본 평균 $\overline{X}$ 에 대하여

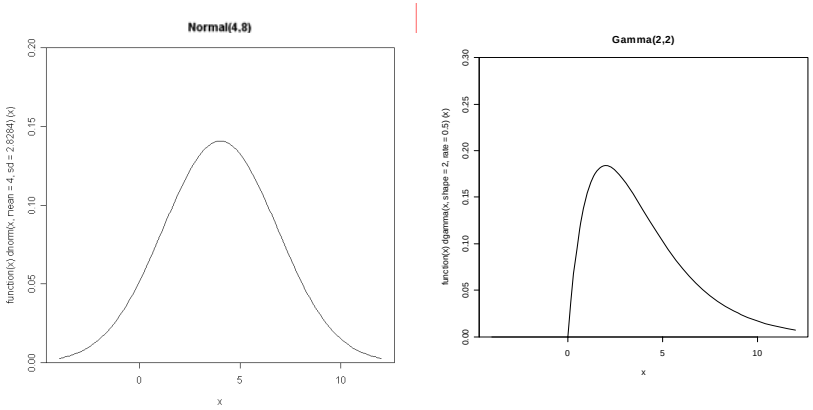
- 모집단 평균이 μ, 분산이 σ^2 인 경우
- 수리적 증명

$$E(\overline{X}) = \mu$$
$$V(\overline{X}) = \sigma^2 / n$$
$$S_{\overline{X}} = \sigma / \sqrt{n}$$

- CLT에 의해
  - 표본의 크기가 충분히 크면 $\overline{X} = \frac{\sum X_i}{n} \sim Normal$
- 표본평균 $\overline{X}$ 는 모집단 평균(μ)의 MVUE이다.
- 모집단이 정규분포이면
  - 표본평균은 표본의 크기에 상관없이 정규분포
- 모집단의 분포를 모르면
  - 대표본(n>20, 모집단의 치우침이 크면)일 경우 표본평균의 분포는 정규분포에 근사한다.

CLT 알아보기

- 모집단 (population)
  - 모집단1~Normal (μ=4, σ^2=8)
    - 왼쪽 그림
  - 모집단2~Gamma (α=2, β=2)
    - 오른쪽 그림



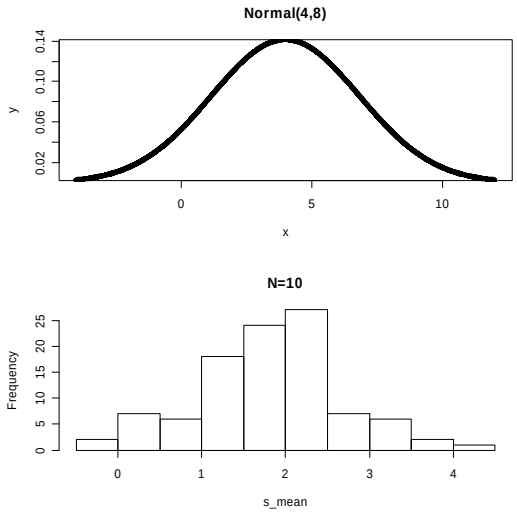
- 난수 생성 (random data generating)
  - 표본의 크기 n(=10, 20, 50)인 데이터 생성한다.
  - 평균을 구한다.
  - 이런 과정을 100번(히스토그램을 이용하여 분포 형태를 잘 알아보기 위하여) 반복한다.
  - 표본 평균의 분포를 알아보기 위하여 히스토그램을 그리고 정규성 검정을 실시한다.
  - 결과 예상
    - 모집단1의 결과는 표본의 크기에 관계 없이 표본평균의 분포는 정규분포에 따른다.
    - 모집단2의 결과: 표본의 크기가 작은 경우 표본평균의 분포는 치우쳐 있을 것이다. n이 커지면 정규분포에 근사할 것이다.

```
> plot(function(x) dnorm(x,mean=4,sd=2.8284),-4,12,ylim=c(0,0.3),main="Normal(4,8)")
> plot(function(x) dgamma(x,shape=2,rate=0.5),-4,12,ylim=c(0,0.3),main="Gamma(2,2)")
```


http://wolfpack.hannam.ac.kr
한남대학교 통계학과 권세혁교수

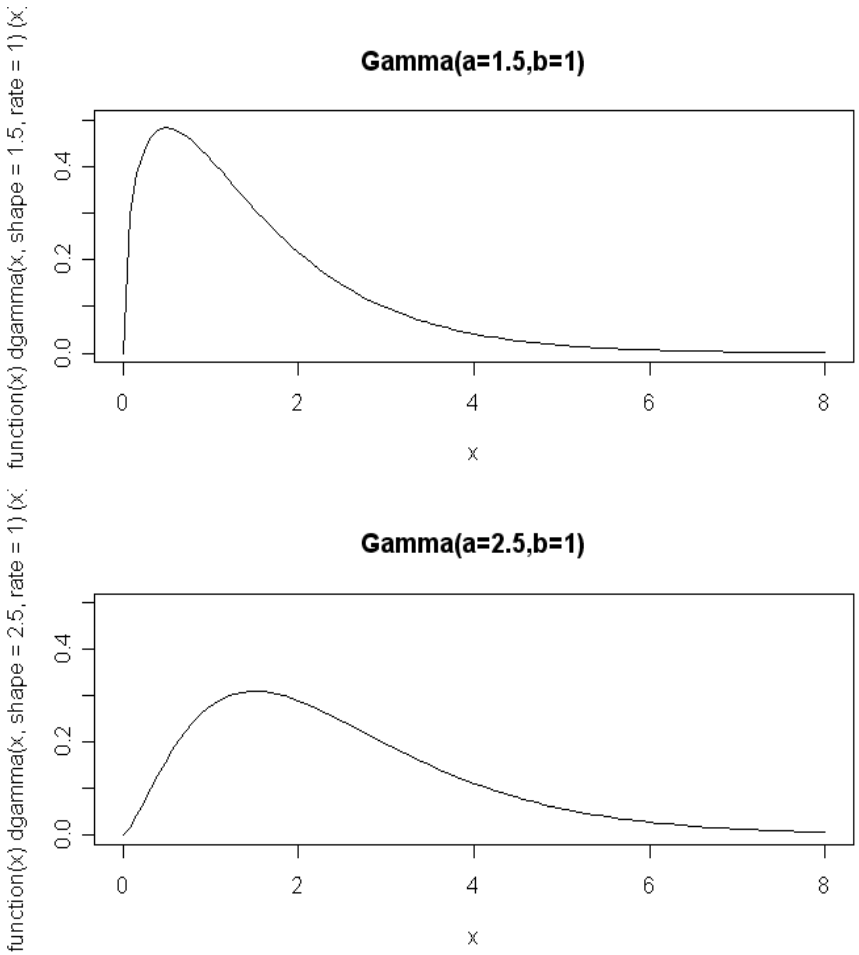
- In Minitab
  - 표본의 n=10인 난수를 100개 생성
  - 표본 100개 평균 구하기
  - 표본 평균의 히스토그램 그리고 정규성 검정
  - 결과 10022007이름 n10.spo에 저장
  - 같은 방법으로 n=20, 50인 경우 결과 저장
    - 10022007이름 n20.spo
    - 10022007이름 n50.spo
- In R
  - 중간 중간에 ls()를 입력하여 데이터 list 보자
  - 데이터 내용을 보고 싶다면 데이터 이름을 입력해 보자.
    - 그래프를 PDF 포멧으로 저장하자.
      - □10022007이름 R n10.pdf
      - □10022007이름 R n20.pdf
      - □10022007이름 R n50.pdf

```
> x<-seq(-4,12, by=0.01)
> y<-dnorm(x,mean=4,sd=2.8284)
> s_mean<-rep(1,100)
> for (i in 1:100) s_mean[i]<-mean(rnorm(10,mean=2,sd=2.8284))
> layout(1:2,2,1)
> plot(x,y,main="Normal(4,8)")
> hist(s_mean,main="N=10")
```



CLT 보기

- 치우친 정도에 따른 n 크기 영향
  - Gamma(α=1.5. β=1)
    - More right skewed
  - Gamma(α=2.5. β=1)
  - n=(10, 20, 50)의 영향보기
    - 각 표본에서 표본 평균을 구하고
    - 이런 작업을 100번 하여 표본평균 히스토그램을 그린다.
  - In R
    - n=10, 20, 30 결과를 한 그래프에 그리시오.
      - □그래프 저장하기: 10092007R.pdf
  - In Minitab
    - 각 그림을 Word 파일에 저장하기
    - 그래프가 의미하는 것 적기
      - □워드 파일 이름: 10092007Minitab.doc



http://wolfpack.hannam.ac.kr
한남대학교 통계학과 권세혁교수

Normal Approximation 2 Binomial

- CLT 응용

  이항분포(Binomial (n,p))에서 실험 회수 n이 크면 CLT에 의해 이항분포는 정규분포에 근사한다.

  X~Binomial(n, p) => 평균=np, 분산=npq

  As $n \to \infty$, X~Binomial(n, p) $\to$ X~Normal(np. npq)

- Why?
  - 이항분포에서 확률표본 Xi는 0 혹은 1(성공)이다. $(x_1, x_2, .., x_n)$
  - 이항분포의 X는 n 시험에서 성공의 회수이다. $X = \sum x_i$
  - 그러므로 CLT에 의해 n이 크면 이항분포 변수 X~Normal

- Continuity Correction (연속 보정)
  - 이항분포를 정규분포 근사로 확률을 구하는 경우 고려해야 하는 요인 (X=k에서 이산형은 확률 존재, 그러나 연속형은 0)
  - (예) P(X≤5|X~Binomial)=P(X≤5.5|X~Normal)
  - (예) P(X≥5|X~Binomial)=P(X ≥ 4.5|X~Normal)

다음을 In Minitab, In R에서 실행하시오.

- 다음 확률을 계산하시오.
  - P(1≤X≤2|X~Binomial (20, 0.1))
    - P(???≤X≤???|X~Normal (?, ?))
  - P(5≤X≤10|X~Binomial (100, 0.1))
    - P(???≤X≤???|X~Normal (?, ?))

- 이항분포의 정규분포 근사를 Histogram이나 확률밀도함수 그리기를 이용하여 보이시오.
  - (n=10, 0.1)과 (n=100, p=0.1)

결과를 워드 문서에 저장(capture)하여 제출하시오.

10112007Minitab.doc

10112007R.doc

http://wolfpack.hannam.ac.kr
한남대학교 통계학과 권세혁교수