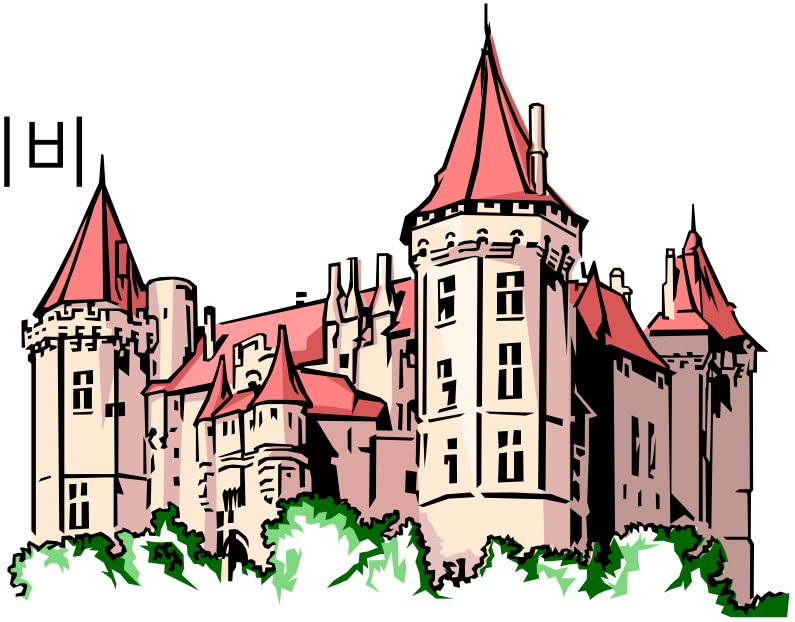
표본분포

예: St. Andrew's

St. Andrew's 대학은 매년 예비 대학생들로부터 900건의 지원(application)을 받고 있다.

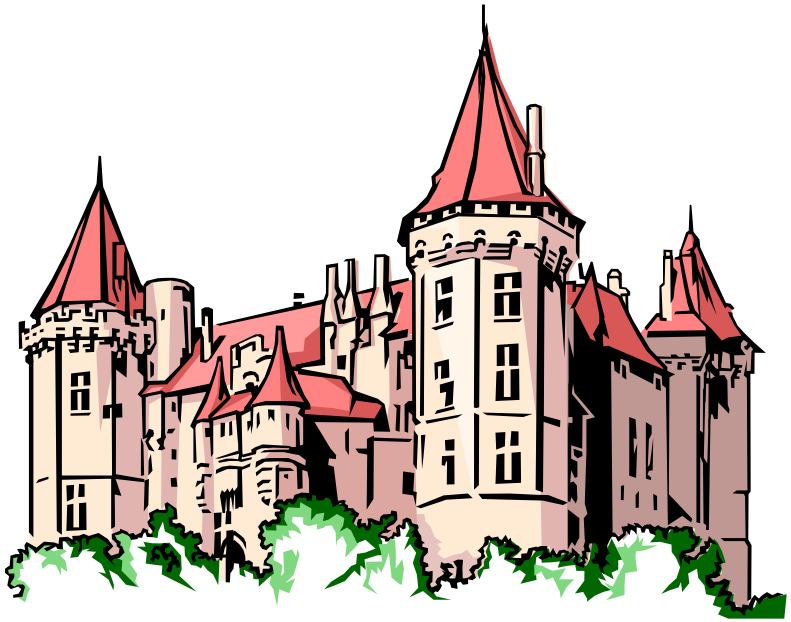


지원서에는 해당 학생의 SAT 성적과 기숙사 생활 희망여부를 담고 있다.

예: St. Andrew's

입학사정관은 다음과 같은 정보를 알고 싶어 한다:

- 900명의 응시자의 SAT 평균 성적
- 기숙사에 살고 싶어 하는 학생의 비율



전수조사(conducting a census)

- 만약 900명의 모든 지원자에 대한 자료가 학교 데이터베이스에 있다면, 제3장의 공식에 따라 관심의 대상이 되는 모수를 계산할 수 있을 것이다.

- 당분간 본 예에서는 전수조사가 가능하다고 가정한다.

전수 조사

모집단의 SAT성적 평균

$$\mu = \frac{\sum x_i}{900} = 1090$$

모집단의 SAT 성적에 대한 표준편차

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{900}} = 80$$

모집단의 기숙사 생활 희망 학생의 비중

$$p = \frac{648}{900} = 0.72$$

단순 무작위 표본 추출

번호	지원자	SAT 성적	기숙생활 희망여부
1	Conrad Harris	1025	Yes
2	Enrique Romero	950	Yes
3	Fabian Avante	1090	No
4	Lucila Cruz	1120	Yes
5	Chan Chiang	930	No
.	.	.	.
.	.	.	.
900	Emily Morse	1010	No

점추정

$\bar{x}$ 는 μ 의 점추정량 이다.

$$\bar{x} = \frac{\sum x_i}{30} = \frac{29910}{30} = 997$$

s 는 σ 의 점추정량이다.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{29}} = \sqrt{\frac{163996}{29}} = 75.2$$

$\hat{p}$ 는 p의 점추정량이다.

$$\hat{p} = \frac{20}{30} = 0.68$$

주목: 다른 표본을 선택하면 점추정값(point estimate)이 다를 수 있다.

$\hat{p}$의 표본분포

$\hat{p}$ 의 표본분포는 표본비율 $\hat{p}$ 의 모든 가능한 값의 확률분포이다.

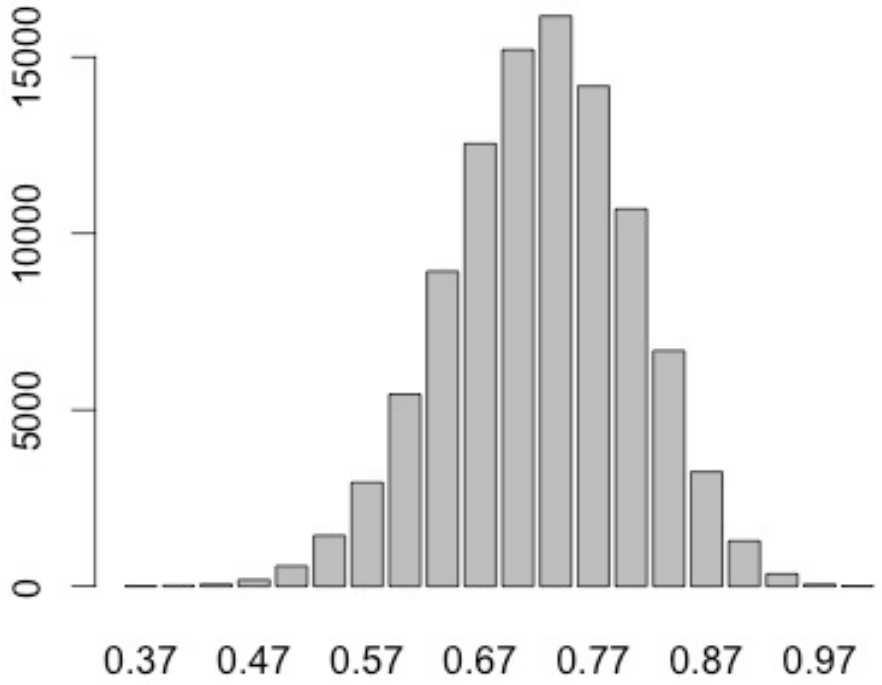
900명의 지원자중 30명을 무작위로 뽑는 행위를 여러 번 (예를 들면, 1000번) 반복했다고 **가정**해보자.

표본 (30 random numbers between 1 and 900)	기숙사 희망 학생의 비율
1,9,15,90,…	.68
3,22,102,132, …	.75
14,32,55,201, …	.70
…	…

1000개의 서로 다른 $\bar{p}$ 를 얻는다.

$\hat{p}$ 의 표본분포

1000개의 $\hat{p}$ 의 분포는 어떤 형태인가?




```
x=rbinom(100000,30,0.72)
phat=x/30
tb=table(phat)
barplot(tb,names.arg=round(as.numeric(names(tb)),2))
```

$\hat{p}$ 의 표본분포

$\hat{p}$ 의 표본분포는 표본비율 $\hat{p}$ 의 모든 가능한 값의 확률분포이다.

$\hat{p}$ 의 기대값 $E(\hat{p}) = p$

$\hat{p}$ 의 표준편차 (표준오차) $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

$\hat{p}$ 의 표본분포

- $\hat{p}$ 의 기대값과 표준편차는 어디서 왔나?

X: 30명 중 기숙사를 희망하는 학생 수
p: 기숙사 희망학생의 모집단 비율

X는 _____ 분포를 따른다. (n= ,p=)

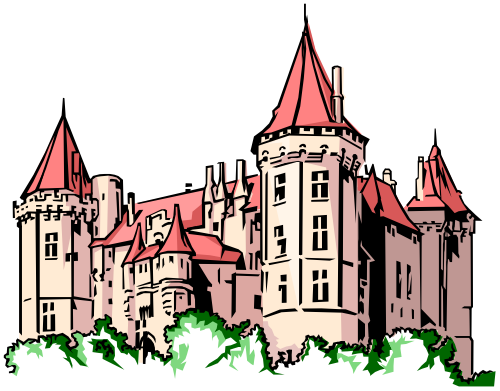
RECALL: 이항분포의 정규근사
시행횟수 n이 커지면 X 는
N(,) 에 근사한다.

따라서 X/n 는
N(,) 에 근사한다.

$\hat{p}$의 표본분포 형태

- 표본비율 $\hat{p}$ 의 분포는 표본의 규모가 커질수록 정규분포에 근사한다.
- 조건: $np \geq 5$ & $n(1-p) \geq 5$
- 모집단 비율이 0.5에 가깝다면, 표본 규모가 10 정도로 작다고 하더라도 정규분포를 이용할 수 있다
- 모집단 비율 p가 매우 크거나(1에 가까운경우) 또는 매우 작을 경우(0에 가까운 경우)에는, 매우 큰 표본이 필요하다.

$\hat{p}$의 표본분포



- 예: St. Andrew's College

앞의 예에서 St. Andrew's College에 지원한 72%의 학생이 기숙사 생활을 희망하는 것으로 조사 되었다.

30명의 단순 무작위 추출표본으로부터 (기숙사 생활을 희망하는) 모집단 비율 추정치가 실제 모집단 비율의 +/- . 05 이내에 있을 확률은 얼마인가?

$\bar{x}$ 의 표본분포

- $\bar{x}$ 의 표본분포는 표본평균 $\bar{x}$ 의 모든 가능한 값의 확률 분포이다.

$\bar{x}$ 의 기대값 $E(\bar{x}) = \mu$

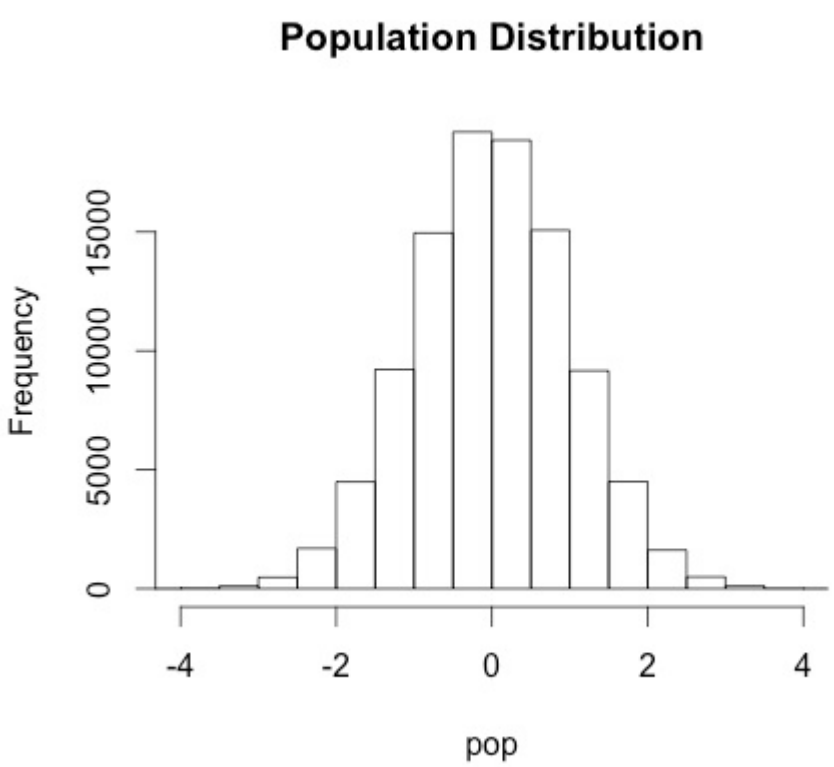
$\bar{x}$ 의 표준편차 (표준오차) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

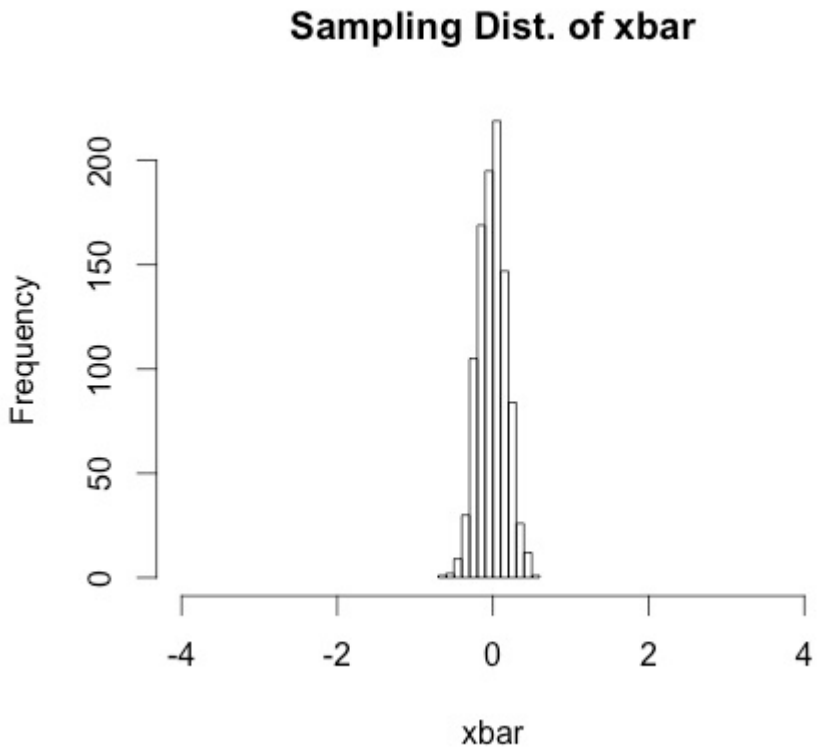
$\bar{x}$의 표본분포의 형태

- 만약 규모가 큰 ($n \geq 30$) 단순 무작위 표본을 사용한다면 **중심극한정리**에 의해 표본평균 $\bar{x}$의 분포는 정규분포에 근사한다.

- 단순 무작위 표본이 작을 경우 ($n < 30$) , 표본평균 $\bar{x}$ 의 분포는 모집단의 분포가 정규분포일 때만 정규분포를 따른다고 할 수 있다.

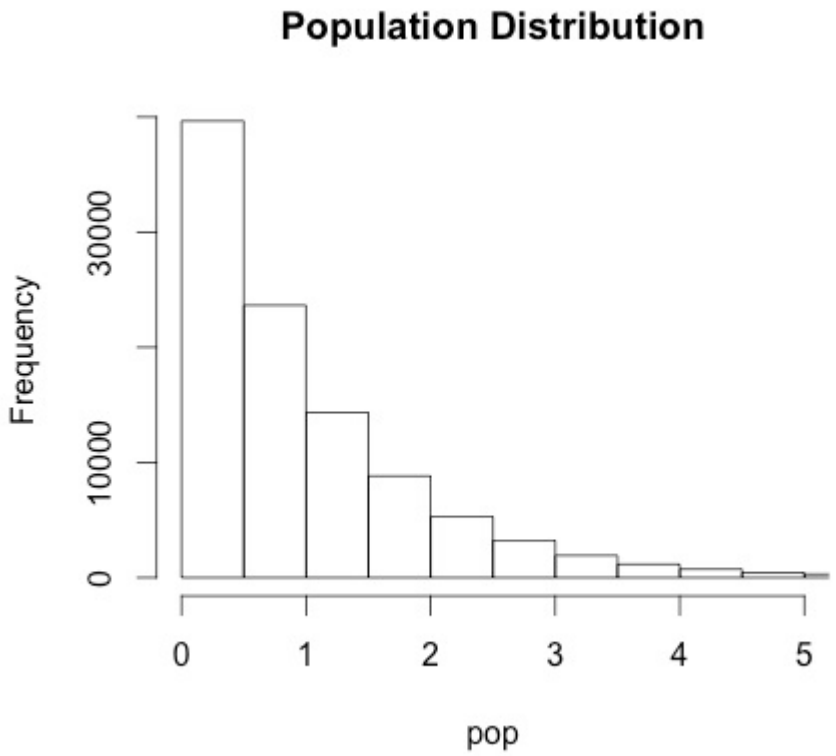
표본분포의 예: 정규분포

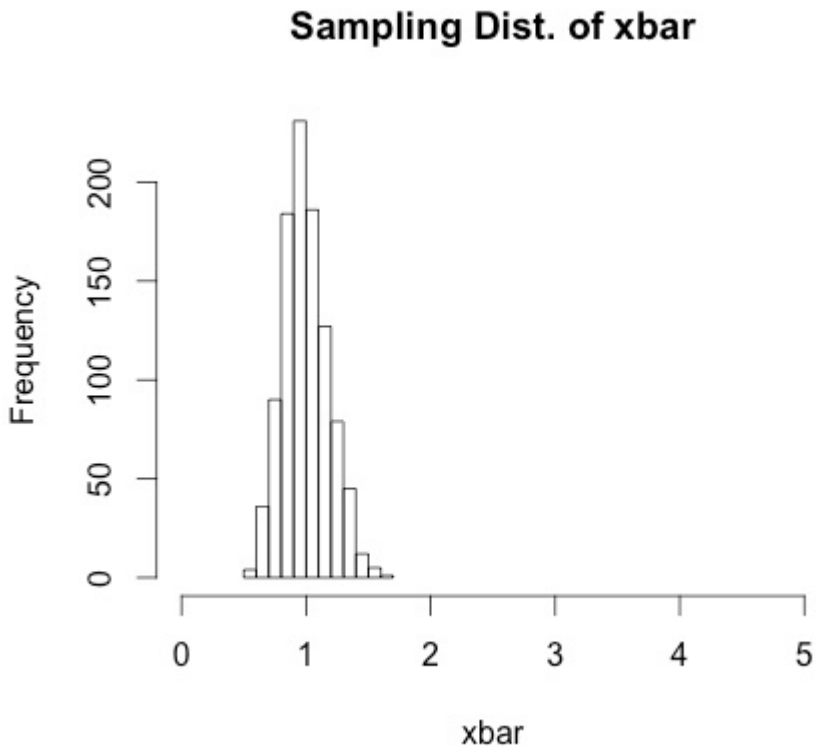





```
pop=rnorm(100000,0,1)
histxbar=c()
for (i in 1:1000){
  xbar[i]=mean(sample(pop,30))
}
hist(pop,main="Population Distribution",xlim=c(-4,4))
hist(xbar,main="Sampling Dist. of xbar",xlim=c(-4,4))
```

표본분포의 예: 정규분포


Population Distribution
Frequency
0
10000
30000
0
1
2
3
4
5
pop


Sampling Dist. of xbar
Frequency
0
50
100
150
200
0
1
2
3
4
5
xbar

SAT 성적에 대한 $\bar{x}$ 의 표본분포

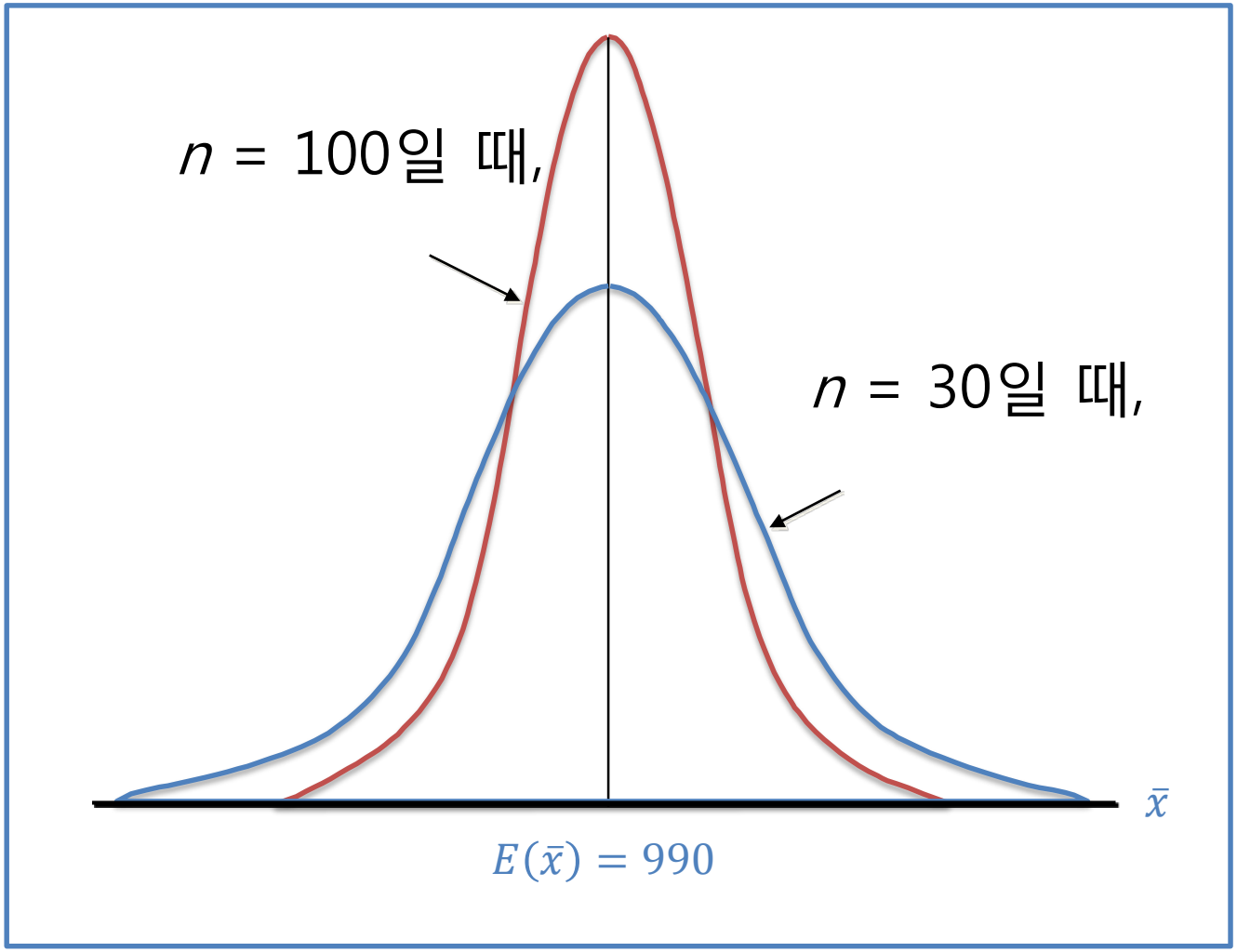
30명의 무작위 추출 표본에 대한 평균이 실제 모집단 평균의 +/- 10 안에 있을 확률은 얼마인가?
다시 말해서, $\bar{x}$가 980에서 1000 사이에 있을 확률은 얼마인가?

표본 규모(sample size)와 $\bar{x}$ 의 표본분포와의 관계

30명이 아닌 100명의 단순 무작위 표본 추출을 가정하자.

- 표본 규모에 상관 없이 $E(\bar{x})$ = _____ .
- 표본 규모가 증가하면 평균의 표준오차 $\sigma_{\bar{x}}$ 는 ____한다.
- 표본의 규모가 100으로 증가함에 따라, 평균의 표준 오차는 _____한다.

표본 규모와 $\bar{x}$ 의 표본분포와의 관계



표본규모와 $\bar{x}$의 표본분포와의 관계

n = 30일 때 , $P(980 \leq \bar{x} \leq 1000)$ =

n = 100 일 때, n = 30 일 경우의 단계를 적용하여

$P(980 \leq \bar{x} \leq 1000)$ 을 풀면,

$P(980 \leq \bar{x} \leq 1000)$ =

n = 100 일 경우 표본분포는 더 작은 표준오차를 가지기 때문에, $\bar{x}$의 값들은 n = 30일 때 보다 더 작은 변동성을 가지며 모집단 평균에 더 가까워 진다.

구간추정

오차한계와 구간추정값 (margin of error and the interval estimate)

- 점추정량은 모집단 모수의 정확한 값을 제공할 것으로 기대되지 않는다 .
- 구간추정값은 오차한계라 불리는 값을 점추정값에 더하고 뺌으로써 계산된다.

점추정값 +/– 오차한계

- 구간추정값의 목적은 점추정값이 모수에 얼마나 근사한지에 관한 정보를 제공하는 것이다 .

모집단 평균의 구간추정

모집단 표준편차 σ가 표본 추출전 알려지지 않은 경우,
표본의 표준편차 s을 σ의 추정치로 사용한다 .

이 때, μ의 구간 추정값은 **t 분포**에 기초한다.

(당분간 모집단이 정규분포라는 것을 가정한다.)

t 분포

- t 분포는 확률분포의 한 종류이다.
- 특정한 t 분포는 자유도에 따라 분포를 달리한다.
- 자유도는 s 를 계산하는데 사용되는 독립적인 정보의 수이다.
- 자유도의 수가 증가할수록 t 분포의 변동성은 낮다.
- 자유도의 수가 증가함에 따라, t 분포와 표준정규분포 간의 차이는 점점 더 줄어들게 된다.

모집단 평균의 구간추정

- 구간추정값

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

여기서: $1-\alpha$ = 신뢰계수

$t_{\alpha/2}$ = t 분포의 오른쪽 꼬리 $\alpha/2$ 에 해당하는 면적에 대한 자유도 $n-1$을 가지는 t 값

s = 표본 표준편차

모집단 비율의 구간추정

- 구간추정값

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

여기서: $1-\alpha$: 신뢰계수

$z_{\alpha/2}$: 정규분포 오른쪽 꼬리의 면적이 $\alpha/2$ 에 해당하는 z-값

$\hat{p}$: 표본비율

가설과 검정

가설검정

1. 귀무가설 (H_0)과 대립가설 (H_a) 설정
2. 알맞은 통계량과 분포 결정하고 가정을 만족하는지 체크
3. 요약된 값이 귀무가설 하에서 쉽게 발생할 수 있는 값인지 조사
   - 검정통계량과 p-value계산
4. 귀무가설 하에서 발생하기 힘든 값이면 귀무가설 기각, 충분히 발생할 수 있는 값이면 기각할 수 없다고 결론
   - 유의수준과 p-value 비교

가설설정

- 귀무가설 (Null Hypothesis; H_0)
  - 기각하고 싶은 기존의 가설
  - 언제나 “=“ 포함

- 대립가설 (Alternative Hypothesis; H_a)
  - 주장하고 싶은 새로운 이론
  - 양측검정: “다르다”, ≠ 포함
  - 단측검정: “~보다 크다 (작다)”, > 이나 < 포함

귀무가설과 대립가설

- 예: Metro EMS

어떤 서해안의 큰 도시가 세계에서 가장 종합적인 응급 의료 서비스를 제공하고 있다.

약 20개의 이동 의료장치(unit)를 가진 다중 의료 시스템을 작동시켜 , 응급상황 발생시 평균 12분 이내에 대처하는 것이 서비스의 목적이다.

귀무가설과 대립가설

- 예: Metro EMS

의료 책임자는 응급상황에서 12분 내에 대처하는지를 알기 위해, 응급상황에 대처하는 데 걸린 시간들을 표본으로 사용하여 가설검정을 하려 한다.

귀무가설과 대립가설

- $H_0: \mu \leq 12$

응급 서비스가 12분 이내에 제공 되었다; 사후조치가 필요 없다.

- $H_a: \mu > 12$

응급 서비스가 12분이 넘어서 제공 되었다; 사후조치가 필요하다.

여기서: μ = 응급서비스 요구에 모집단의 평균 대응시간

귀무가설 vs 대립가설

- 2000년에 조사한 중학생들의 평균 키는 170cm였다. 10년이 지난 2010년 식생활 패턴의 변화로 중학생들의 키가 커졌다고 생각된다.

- 기존에는 A약이 고혈압에 가장 효과가 있는 치료제였다. 유전공학을 이용해 새로 개발된 B약이 고혈압에 더 효과가 있다고 생각된다.

- Zyrtec의 특허가 끝난후 Walgreens가 Wyr-Zyr를 만들었는데 Zyrtec과 같은 성분인 Certirizine을 사용하여 제조하였으므로 효과가 같다고 생각된다.

검정통계량 (Test Statistics)

- 실험이나 관찰로부터 나온 데이터를 요약한 수치
- 문제와 가설에 따라 적절한 통계기법을 사용하여 검정통계량을 구한다.
  - Z-statistics, T-statistics, F-statistics, Chi-square statistics 등

1종 오류와 2종 오류

	모집단의 상황	
통계적 결론	H_0 참 ($\mu \le 12$)	H_0 거짓 ($\mu > 12$)
채택 H_0 (결론 $\mu \le 12$)	정확한 판단	2종 오류
기각 H_0 (결론 $\mu > 12$)	1종 오류	정확한 판단

1종 오류: H_0 가 참인데도 기각하는 경우 => 유의수준 (Level of Significance, α)
2종 오류: H_0 가 거짓인데 받아들이는 경우이다.

유의수준 (Level of Significance, α)

- 어느 정도 드문 현상을 유의하다고 판단하느냐의 기준
  - α=0.01: 귀무가설 하에서 관찰될 가능성이 1% 이하인 검정통계량을 유의하다고 결정하자고 약속
  - α=0.05는 α=0.01보다
    - 검정통계량이 유의하다고 결정할 가능성이 (높다, 낮다)
    - 귀무가설을 기각할 가능성이 (높다, 낮다)
- 일반적으로 0.01,0.05,0.1 중 선택

유의확률 (P-value)

- 검정통계량이 귀무가설을 지지하는 정도

- **P-value< α : 귀무가설 기각**

- **P-value> α : 귀무가설 기각하지 못함**

One-Sample T-Test

One-sample T-Test

- 모집단의 평균이 어떤 특정한 값과 같은지를 검증

예) 2006년 조사에 의하면 한국인의 1인1일 평균 알코올 섭취량은 8.1g이다. 2008년 대통령 선거로 알코올 섭취량이 달라졌는지 조사하기 위해 10명을 무작위로 뽑아서 조사한 결과 다음과 같은 데이터를 얻었다.

15.5,11.21,12.67,8.87,12.15,9.88,2.06,14.5,0,4.97

```
> rm(list=ls()) #workspace 지우기
> x=c(15.5,11.21,12.67,8.87,12.15,9.88,2.06,14.5,0,4.97)
```

1. 귀무가설 대립가설 설정

H_0:

H_a:

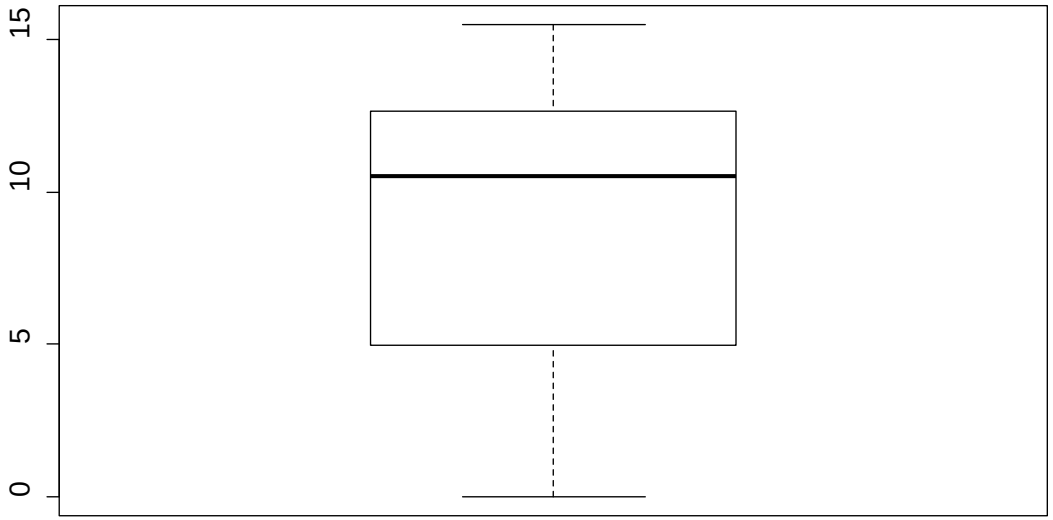
2. 가정체크

- **가정**
  - **자료가 정규분포를 따른다.**
  - **심하게 편중되거나 극단치를 포함한 경우 표본수가 50개 이상이다.**

- 조건을 만족하지 않으면 Wilcoxon signed-rank 이용

2.가정체크

- 히스토그램 or Boxplot



여기서 잠깐: 그림 내보내기 in R

```
jpeg("c:/figures/boxplot.jpg", width=480, height=480)
boxplot(x)
dev.off()
```

```
postscript("c:/figures/boxplot.eps",width=4,height=4,paper="special",h
orizontal=FALSE,onefile=FALSE )
boxplot(x)
dev.off()
```

2.가정체크

- Shapiro-Wilk normality test

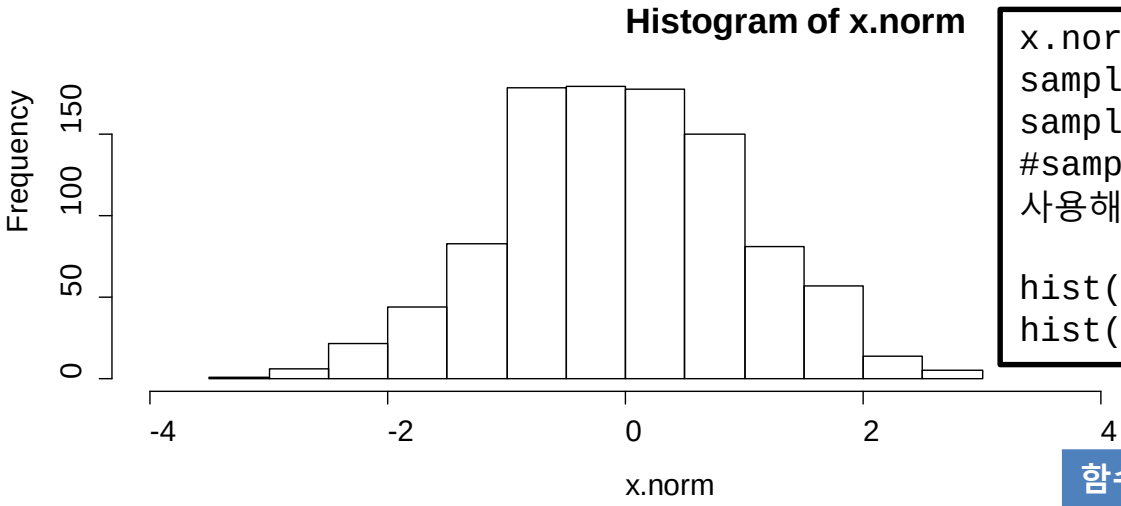
```
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9234, p-value = 0.3863
```

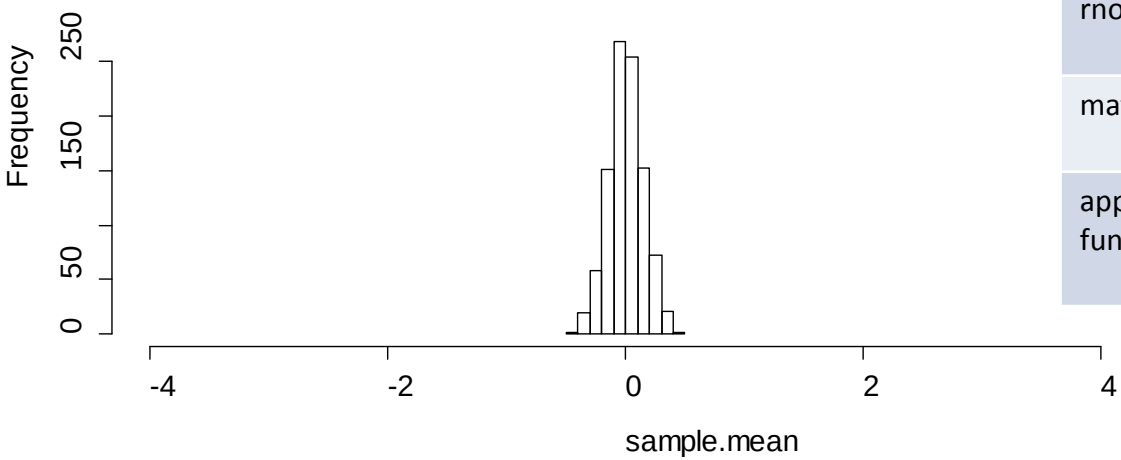
H_0: 데이터가 정규분포를 따른다.
H_a: 데이터가 정규분포를 따르지 않는다.

- P-value>0.05 ➔ 데이터가 정규분포를 따른다고 결정 => t-test 진행
- P-value<0.05 ➔ 데이터가 정규분포를 따르지 않는다 => 관측수가 충분히 크지 않으면 (>50), Wilcoxon signed rank test 이용

3. 검정통계량 계산: 정규분포, 표본분포

자료가 평균 μ, 분산 σ^2인 정규분포를 따른다고 가정하자. $N(\mu, \sigma^2)$
표본평균 $\bar{x}$ 는 $N(\mu, \sigma^2/n)$을 따른다.






```
x.norm=rnorm(1000)
sample=matrix(rnorm(50000),50,1000)
sample.mean=apply(sample,2,mean)
#sample.mean=colMeans(sample)을
사용해도  무방

hist(x.norm,xlim=c(-4,4))
hist(sample.mean,xlim=c(-4,4))
```

함수명	내용
rnorm(n,m,s)	평균 m, 표준편차 s인 정규분포를 따르는 n개 수
matrix(x,a,b)	벡터 x로 a행 b열의 행렬생성
apply(matrix,margin, fun)	Matrix의 행(mar=1)이나 열(mar=2)에 fun의 함수를 적용

3. 검정통계량 계산: T-statistics

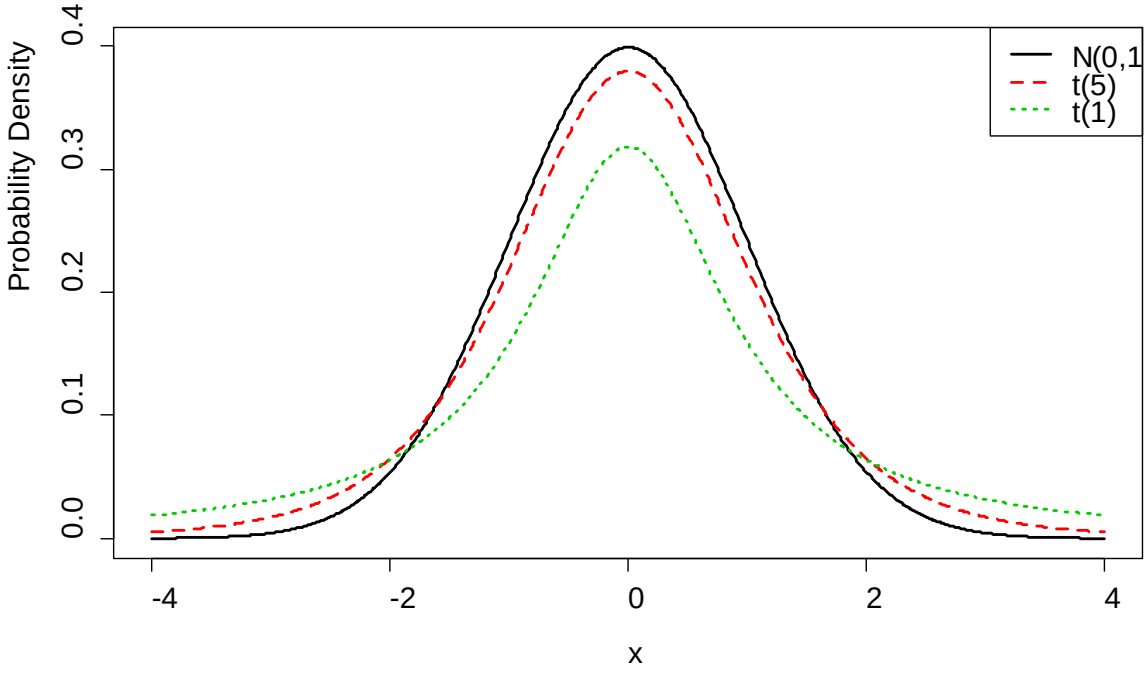
$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- Z는 평균 0, 표준편차 1인 표준정규분포를 따른다.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{x} - \mu)}{s} \sim t(n-1)$$

- 모분산(σ^2)이 알려진 경우 거의 없다.
- 모분산을 표본분산 (s^2)으로 대체하면 자유도 (t-1)인 t분포를 따른다.

3. 검정통계량 계산: T 분포와 표준정규분포



함수명	내용
dnorm(x,m,s)	평균 m, 표준편차 s인 정규분포의 x에서의 분포함수값
dt(x,df)	자유도 df인 t 분포의 x에서의 분포함수값
matplot(x, matrix)	X와 matrix의 각 열의 그래프를 동시에 그림
legend	여러 그래프를 한 그림에 그렸을 때 범례삽입

```
y=seq(-4,4,0.01)
matplot(y,cbind(dnorm(y),dt(y,5),dt(y,1)),'l',xlab="x",ylab="Probability Density",lwd=2)
legend('topright',c("N(0,1)","t(5)","t(1)"),col=1:3,lty=1:3,lwd=2)
```

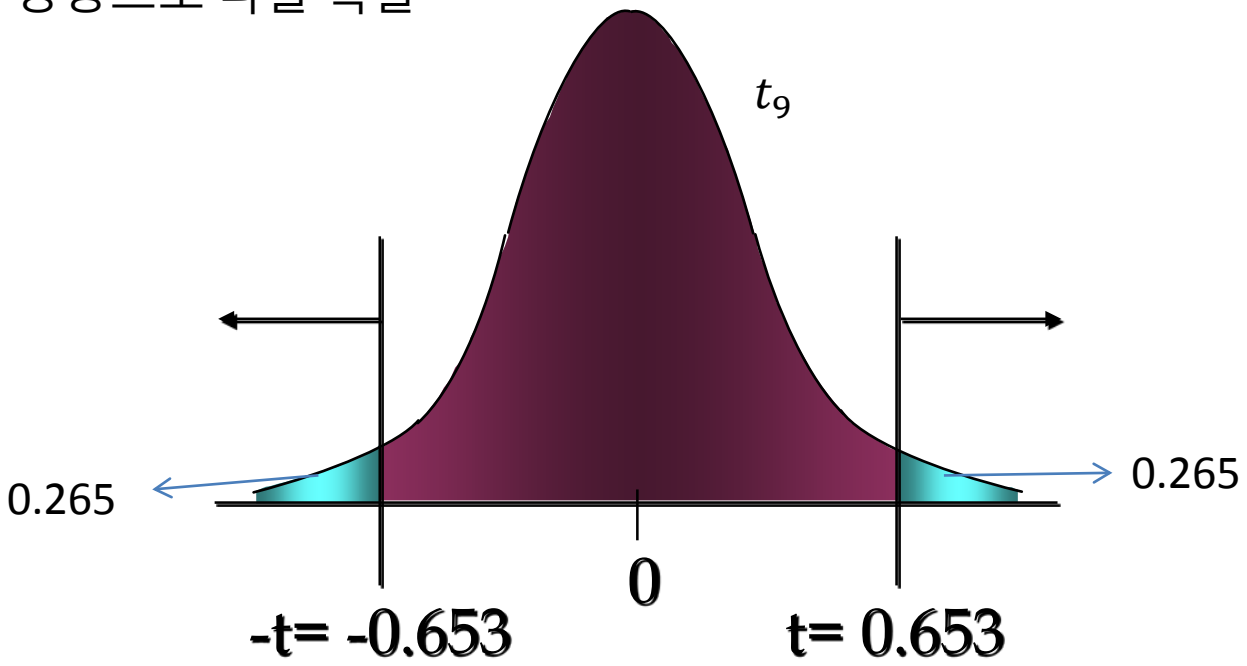
3. 검정통계량 계산: T-statistics (예)

- *t.test 명령어를 사용하면 자동으로 t-statistics과 p-value를 계산*
- *여기서는 연습으로 한번 직접 구해본다.*
- $\mu = 8.1$

```
> sqrt(length(x))*(mean(x)-8.1)/sd(x)
[1] 0.6529981
```

3. P-value 계산

모집단에서 표본추출을 반복했을 때 검정통계량들이 0.653보다 더 대립가설을 지지하는 방향으로 나올 확률



$$\Pr(|t_9| > 0.653) = 2 \times \{1 - \Pr(t_9 < 0.653)\}$$

```
> 2*(1-pt(0.653, df=9)) #p-value
[1] 0.5300818
```

3. 검정통계량과 P-value 계산 (t.test)

```
> t.test(x,mu=8.1)

        One Sample t-test

data:  x
t = 0.653, df = 9, p-value = 0.5301
alternative hypothesis: true mean is not equal to 8.1
95 percent confidence interval:
 5.436132 12.925868
sample estimates:
mean of x
    9.181
```

T-statistics

P-value
: α보다 작으면
귀무가설 기각

대립가설

95% 신뢰구간
: 실제 알콜섭취 평균량이 이 구간 안에 포함될 것을 95% 만큼 확신한다.

점추정량
:실제 알콜섭취 평균량을 9.181로 추정한다.

4. 결론

- P-value=0.5301 > 0.05

➔귀무가설을 기각하지 못한다.

대선 후 알콜섭취 평균량이 평상시 평균량과 다르다고 할 수 없다.

양측검정 vs 단측검정

- $H_0: \mu = 8.1 \; vs \; H_a: \mu > 8.1$

```
> t.test(x,mu=8.1,alter="greater")

        One Sample t-test

data:  x
t = 0.653, df = 9, p-value = 0.265
alternative hypothesis: true mean is greater than 8.1
95 percent confidence interval:
 6.146389      Inf
sample estimates:
mean of x
    9.181
```

대립가설

- $H_0: \mu = 8.1 \; vs \; H_a: \mu < 8.1$

```
> t.test(x,mu=8.1,alter="less")
```

유의수준 (α) 조정

- default: $\alpha = 0.05$

```
> t.test(x,mu=8.1,conf.level=0.99)

        One Sample t-test

data:  x
t = 0.653, df = 9, p-value = 0.5301
alternative hypothesis: true mean is not equal to 8.1
99 percent confidence interval:
  3.801088 14.560912
sample estimates:
mean of x
    9.181
```

99% 신뢰구간
: 실제 알콜섭취
평균량이 이 구간 안에
포함될 것을 99% 만큼
확신한다.