# Chapter 8 & 9
# Sample Distribution & Estimation

경영대학 재무금융학과
윤선중



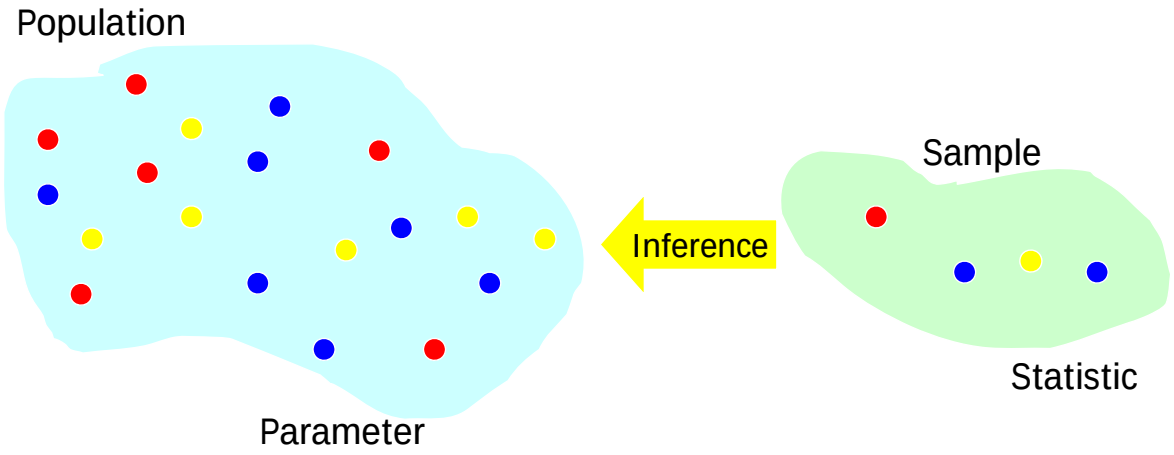
## Objectives

- 표본 분포 (Sampling Distribution) 의 정의
- 표본분포
  - 표본평균
  - 표본비율
  - 표본분산
  - 두 표본평균 차이의 표본분포
- 추정 (Estimation)
  - 추정의 개념
  - 모분산이 알려져 있을 때 모평균의 추정량 (estimator)
  - 모평균을 추정하기 위해 필요한 표본 크기 공식 소개

# I. Sampling Distribution



# Introduction



# Introduction

- 표본분포의 정의
  - 표본추출을 통하여 계산한 표본통계량이 가지는 확률분포
  - 예시: 표본평균의 확률분포, 표본표준편차의 확률 분포 등
- 표본분포의 필요성
  - 많은 경우 모집단의 특성을 모름
    - 표본추출을 통하여 표본평균 등과 같은 표본통계량에 기반하여 의사결정이 이루어져야함
    - 표본 통계량이 어떠한 확률분포를 가지는 지를 명확히 알아야 함
  - 예제
    - 상장기업에 대한 주식 투자 수익률 – N(5%, 10%^2)이라고 함
      - 임의로 하나의 종목을 선택하여 투자할 때, 손실이 나지 않을 확률은?
      - 임의로 9개의 기업을 선택하여 포트폴리오 투자를 한다고 할 때, 손실이 나지 않을 확률은?



# Sampling Distribution of Mean

- 표본평균의 기대값과 분산
  - $X \sim N(\mu, \sigma^2)$인 모집단에서 임의표본추출을 통하여 표본평균을 계산

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}, \Rightarrow \begin{cases} E(\bar{X}) = \mu \\ V(\bar{X}) = \sigma^2 / N \end{cases}$$

  - 주사위 평균의 표본분포

| x | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| P(x) | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

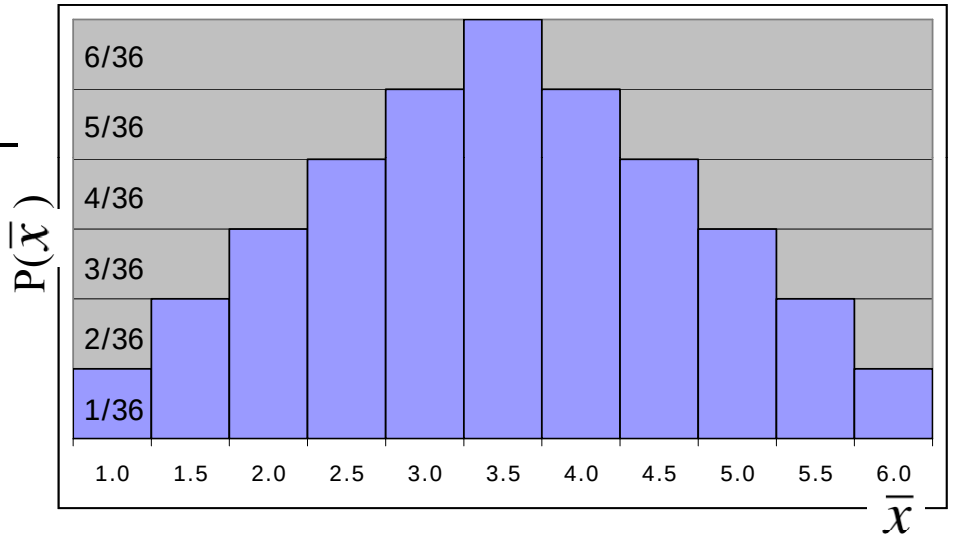
$$\mu = \sum xP(x) = 1(\tfrac{1}{6}) + 2(\tfrac{1}{6}) + \ldots + 6(\tfrac{1}{6}) = 3.5$$

$$\sigma^2 = \sum (x-\mu)^2 P(x) = (1-3.5)^2(\tfrac{1}{6}) + \ldots + (6-3.5)^2(\tfrac{1}{6}) = 2.92$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$$

# Sampling Distribution of Mean

| $\bar{x}$ | $\bar{x}$ |
|---|---|
| 1.0 | 1/36 |
| 1.5 | 2/36 |
| 2.0 | 3/36 |
| 2.5 | 4/36 |
| 3.0 | 5/36 |
| 3.5 | 6/36 |
| 4.0 | 5/36 |
| 4.5 | 4/36 |
| 5.0 | 3/36 |
| 5.5 | 2/36 |
| 6.0 | 1/36 |




$$\mu_{\bar{x}} = \sum \bar{x}P(\bar{x}) = 1.0(\tfrac{1}{36}) + 1.5(\tfrac{2}{36}) + \ldots + 6.0(\tfrac{1}{36}) = 3.5$$

$$\sigma_{\bar{x}}^2 = \sum (\bar{x} - \mu_{\bar{x}})^2 P(\bar{x}) = (1.0 - 3.5)^2(\tfrac{1}{36}) + \ldots + (6.0 - 3.5)^2(\tfrac{1}{36}) = 1.46$$

$$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}^2} = \sqrt{1.46} = 1.21$$



# Sampling Distribution of Mean

- 표본평균의 분포
  - 모집단이 정규분포인 경우
    - 표본평균도 정규분포
  - 모집단의 분포를 모르는 경우
    - 표본추출 시 표본의 크기가 충분히 크면 표본평균이 정규분포 (중심극한정리)
- 중심극한정리(Central Limit Theorem)
  - 임의의 모집단으로부터 추출된 표본평균의 분포는 표본의 크기가 충분히 큰 경우 _____분포로 수렴함
  - 일반적으로 요구되는 표본의 크기는 30개 이상

# Sampling Distribution of Mean

- 예제 8.1: 32온스 병의 내용물 중량
  - 실제 음료의 양: 평균 32.2온스 표준편차: 0.3온스의 정규분포
    - 임의 병이 32온스를 초과하는 음료를 가지고 있을 확률?

$$P(X > 32) = P\left(\frac{X-\mu}{\sigma} > \frac{32-32.2}{.3}\right) = P(Z > -.67) = 1 - .2514 = .7486$$

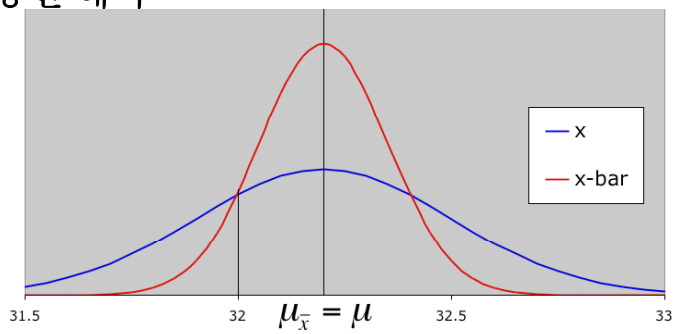
    - 한 고객이 4병 짜리 팩을 살경우 이 팩에 들어있는 탄산음료의 평균이 32온스를 초과할 확률?
      - X가 정규분포를 따르므로 샘플평균도 정규분포를 따름
      - 평균: $\mu_{\bar{x}} = \mu$
      - 표준편차: $\sigma_{\bar{x}} = \sigma/\sqrt{n} = .3/\sqrt{4} = .15$

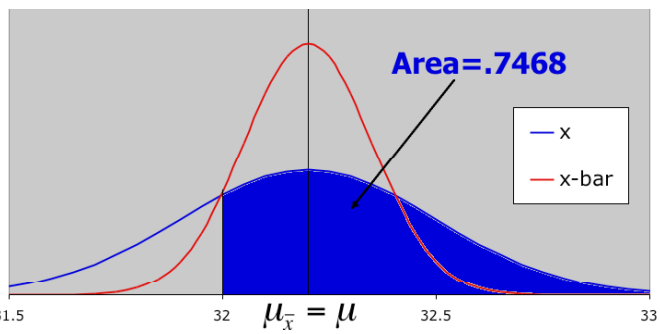
$$P(\bar{X} > 32) = P\left(\frac{\bar{X}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} > \frac{32-32.2}{.15}\right) = P(Z > -1.33) = .9082$$

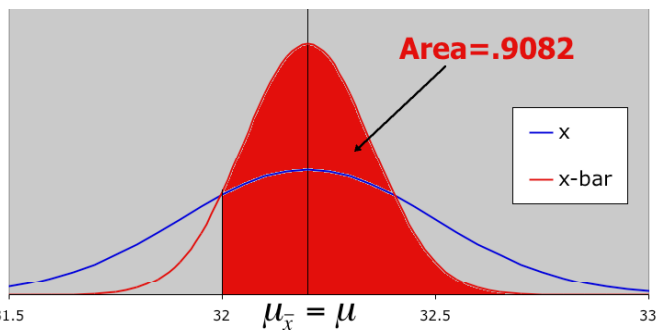


# Sampling Distribution of Mean

- 그래프를 통한 해석



# Sampling Distribution of Mean

- 예제
  - A 지역의 소방서가 화재신고를 받고 현장에 도착하는 시간
    - 평균 14분, 표준편차 4분인 정규분포를 따름
  - 이 소방서에 접수되는 화재신고 16건을 무작위로 추출
    - 16건의 평균 반응시간이 15분 이하일 확률은?
    - 16건의 평균반응시간이 12.5분과 15.5분 사이일 확률?
  - 반응시간이 정규분포라는가정이 없다면?
    - 위 문제의 확률은?
    - 16건이 아닌 64건을 무작위로 추출하였을 때의 결과는?



# Sampling Distribution of Proportion

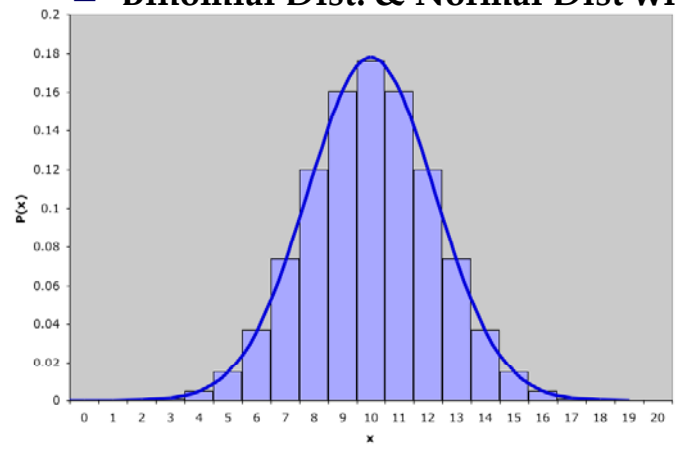
- 표본비율의 정의
  - 베르누이 시행을 n번 반복하는 이항실험
  - 이 실험에서 특정 결과가 발생할 비율
    - 동전을 100회 던져서 앞면이 나올 비율
    - 100개 종목에 주식투자를 하였을 때, 이익이 발생한 종목의 비율
- 표본비율의 표본분포 (정규분포로의 수렴)
  - 성공확률 p, 총 N개의 표본자료
  - 기대값과 분산

$$\hat{p} = \frac{X}{N} \Rightarrow \begin{cases} E(\hat{p}) = p \\ V(\hat{p}) = \dfrac{p(1-p)}{N} \end{cases}$$

  - 표본분포
    - 표본의 크기가 충분히 크면 _____분포로 수렴

# Sampling Distribution of Proportion

- **Binomial Dist. & Normal Dist with** $\mu = 10, \sigma = 2.24$




$$\mu = np$$

$$\sigma^2 = np(1-p)$$

$$\sigma = \sqrt{np(1-p)}$$

이항분포가 잘 근사화 될 수 있는 조건

1) $np \geq 5$

2) $n(1-p) \geq 5$

$P(X = 10) \approx P(9.5 < Y < 10.5)$

$P(X = 10) = .176$

vs

$P(9.5 < Y < 10.5) = .1742$



# Sampling Distribution of Proportion

- 표본 비율의 표본 분포 근사

$$E(\hat{P}) = p$$

$$V(\hat{P}) = \sigma_{\hat{p}}^2 = \frac{p(1-p)}{n} \qquad Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

- 예제 8.2: 정치 여론 조사
  - 지난 선거에서 52%지지, 임의의 300명의 표본에서 p>1/2일 확률은?

$P(\hat{P} > .50)$

$$\sqrt{p(1-p)/n} = \sqrt{(.52)(1-.52)/300} = .0288$$

$$P(\hat{P} > .50)$$

$$= P\left(\frac{\hat{P} - p}{\sqrt{p(1-p)/n}} > \frac{.50 - .52}{.0288}\right)$$

$$= P(Z > -.69)$$

$$= .7549$$

# Sampling Distribution of Differences b/w Means

- 표본평균 차이의 기대값과 분산
  - 이고 서로 독립이라고 하면,

$$\bar{X}_1 = \frac{\sum_{i=1}^{N_1} X_{1i}}{N_1}, \bar{X}_2 = \frac{\sum_{i=1}^{N_2} X_{2i}}{N_2} \Rightarrow \quad \mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- 표본평균 차이의 분포
  - 두 모집단이 정규분포인 경우
    - 표본평균의 차이도 정규분포
  - 두 모집단의 분포를 모르는 경우
    - 표본의 크기가 충분히 크면 표본평균의 차이는 _____분포
    - 표본의 크기가 충분하지 않으면 분포를 알 수 없음



# Sampling Distribution of Differences b/w Means

- 예제
  - 중국 시장 상장기업 주식의 투자수익률은 평균 10%, 분산 25%
  - 일본 시장 상장기업 주식의 투자수익률은 평균 5%, 분산 9%
  - (1) 두 시장의 투자수익률은 모두 정규분포를 따른다고 가정
    - 임의로 중국기업하나와 일본기업 하나에 투자하였을 때, 중국기업의 투자 수익률이 더 높을 확률은?
    - 중국 16개 기업, 일본 9개 기업에 각각 포트폴리오 투자를 하였다면, 중국 기업 포트폴리오의 수익률이 더 높을 확률은?
  - (2) 두 시장의 투자수익률 분포가 알려져 있지 않다고 가정
    - 위 (1)의 두 질문에 대한 답은?
    - 중국 100개 기업, 일본 49개 기업에 각각 포트폴리오 투자를 하였다면, 중국 기업 포트폴리오의 수익률이 더 높을 확률은?

# Sampling Distribution of Proportion

- 표본분포 & 통계적 추론

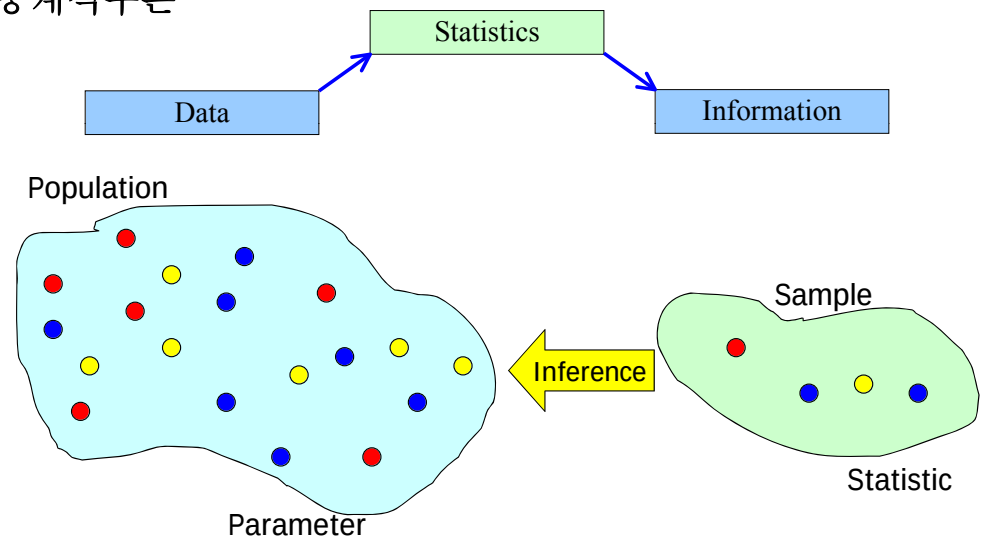
| | | |
|---|---|---|
| • 모집단과 모수 | 확률분포 → | 확률변수의 개별확률 |
| • 모집단과 모수 | 표본분포 → | 통계량의 확률 |
| • 통계량 | 추정 → | 모수 |



# II. Estimation

# Introduction

- 통계적추론






# Introduction

- 추론
  - 앞선 수업에서 이항, 포아송, 정규분포는 모집단의 특성에 대한 분석을 돕는다.
  - 그러나 대부분의 현실 상황에서 모집단의 특성이 알려져 있지 않다.
  - 따라서 표본을 통해 모집단에 대한 특성을 추정 해야함.
  - 이항분포 p; 포아송 μ; 정규분포 μ & σ
- 통계적 추론
  - 추정(Estimation) – 이번 시간에 살펴볼 내용
    - 표본통계량에 근거하여 모수(parameter)의 근사값을 결정하는 것
  - 가설검정(Hypothesis test)
    - 모수에 관한 가설이 데이터에 의해 지지되는지를 검증하는것

# Introduction

- 기본 용어
  - **추정량 (estimator)**
    - 모수를 추정하기 위하여 사용한 표본 통계량
    - 표본평균은 모평균의 _____
  - **추정치 (Estimate)**
    - 추정량이 가지는 구체적인 값
    - 표본평균의 값은 모평균의 _____
  - **점 (point) 추정**
    - 모수의 참값으로 생각되는 단 하나의 값을 추정치로 선택
  - **구간(interval) 추정**
    - 단, 하나의 값이 아니라 모수의 참 값이 속할 것으로 기대되는 범위를 선택



# Desirable Properties of an Estimator

- **불편성 (unbiasness)**
  - **추정량의 기대값이 모수와 같은 경우**
  - **표본평균, 표본분산은 모두 불편 추정량** $\bar{X} = \frac{\sum X_i}{n}, s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$
  - **불편 추정량 (unbiased estimator) vs. 편의 추정량 (biased estimator)**
- **일치성 (Consistency)**
  - **표본의 크기가 커짐에 따라 추정량과 모수의 차이가 더 줄어드는특성**
  - **표본평균의 경우?**
- **(상대적) 효율성 (Efficiency)**
  - **한 모수에 대해 두개의 불편 추정량이 존재한다고 할 때, 분산이 더 작은 추정량을 효율적이라고 말함**
  - **표본평균 vs. 표본 중앙값**

# Example

- 백화점 신용카드 사용 예
  - 한림 백화점 고객들의 한달 평균 지출을 알아보고자 함
  - 백화점 고객: 약 150,000 명
  - 400명만 무작위로 추출하여 조사하기로 함
  - 400명을 대상으로 조사한 결과,표본평균은45,000원으로 나타남
  - 관련협회의 통계에 의하면 백화점 고객들의한달 지출액의 표준편차는 약 100,000원이라고 알려짐
  - 한림 백화점 고객의 지출액 표준편차도 100,000일 것으로 가정
- 문제
  - (1)한림 백화점 고객들의 한달 평균 지출액은 얼마인가?
  - (2) 90% 정확도로 이야기 한다면, 한림 백화점 고객들의 한달 평균 지출액은 어느 범위에 속하는가?



# Example

- (1) 점추청
  - 표본평균=450,000
  - 표본표준편차(표준오차)=100,000/20=5,000
- (2) 구간 추정
  - 표본의 크기가 충분히 클 때,

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

  - 문제:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

$$P\left( z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < z_{\alpha/2} \right) = 1 - \alpha$$

$$P\left( \mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

$$P\left( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

# Confidence Interval with known Variances

- 신뢰수준
  - 구간이 실제로 모수(i.e. 모평균)를 포함하고 있을 확률; $1-\alpha$
- 신뢰하한
  - $\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$
- 신뢰상한
  - $\left(\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$
- 신뢰수준과 z

| 1−α | α | α/2 | z |
|---|---|---|---|
| 0.9 | 0.1 | 0.05 | 1.645 |
| 0.95 | 0.05 | 0.025 | 1.96 |
| 0.98 | 0.02 | 0.01 | 2.33 |
| 0.99 | 0.01 | 0.005 | 2.575 |



# Example

- 한림백화점 예제의 해답

$$\frac{\bar{X} - x_h}{\sigma/\sqrt{n}} = -1.645, \frac{\bar{X} - x_l}{\sigma/\sqrt{n}} = 1.645$$

$$\Rightarrow \frac{450,000 - x_h}{100,000/\sqrt{400}} = -1.645, \frac{450,000 - x_l}{100,000/\sqrt{400}} = 1.645$$

$$x_h = 458,225\text{원}\ x_l = 441,775$$

# Exercise

- 한림대학교 학생 용돈
  - 한림대학교 학생들의 한달 용돈은 모르며, 다만 표준편차는 5만원
  - 임의로 선택한 100명의 학생들로부터 조사한 표본평균은 20만원
  - 한림대학교 학생들의 한달 평균 용돈에 대한 95% 신뢰구간은?
- 주가상승률
  - 유가증권시장 개별 종목들의 주가 상승률 평균은 알지 못하며, 표준편차는 10%라고 함
  - 임의로 선택한 100개 종목의 평균 수익률이 8%라고 할 때, 유가증권시장의 평균 주가상승률에 대한 99% 신뢰구간은?
- 위 두 예제에서 신뢰구간의 의미는?
  - 위와 같이 표본평균을 계산하여 신뢰구간을 계산하였을 때, 해당 구간 내에 모수가 위치할 확률이 95% 혹은 99%
  - 위에서 계산한 신뢰구간의 범위 자체에 모수가 위치할 확률이 아님에 유의!!



# Confidence Interval with Unknown Variances

- 분산이 알려져 있는 경우
  - 신뢰수준 $1-\alpha$ 에 대한 신뢰구간

$$\left( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

- 분산이 알려져 있지 않은 경우
  - 신뢰수준 $1-\alpha$ 에 대한 신뢰구간

$$\left( \bar{X} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \right)$$

  - 모표준편차를 표본표준편차 s로 대체하고 검정통계량은 William S. Gosset이 창안한 t-통계량을 사용
  - 자유도 n-1을 가지는 t 통계량 (제 7장에서 소개)

# Exercise

- 한림대학교 학생용돈
  - 한림대학교 학생들의 한달 용돈은 모르며 표준편차도 모름
  - 임의로 선택한 100명의 학생들로부터 조사한 표본평균은 20만원, 표본표준편차는 10만원
  - 한림대학교 학생들의 한달 평균 용돈에 대한 95% 신뢰구간은?
  - 만약, 조사한 학생의 수가 16명이었다면?
- 신뢰구간의 길이에 영향을 미치는 요인(**)
  - 신뢰수준
    - 추정의 정확성을 반영
  - 모 표준편차
    - 원 데이터의 산포도를 반영
  - 표본의 크기
    - 추정에 소요되는 비용과 연관



# Excel Application

- 예제 9.1 Doll Computer Company; Xm 09-01
  - 리드타임 동안의 수요

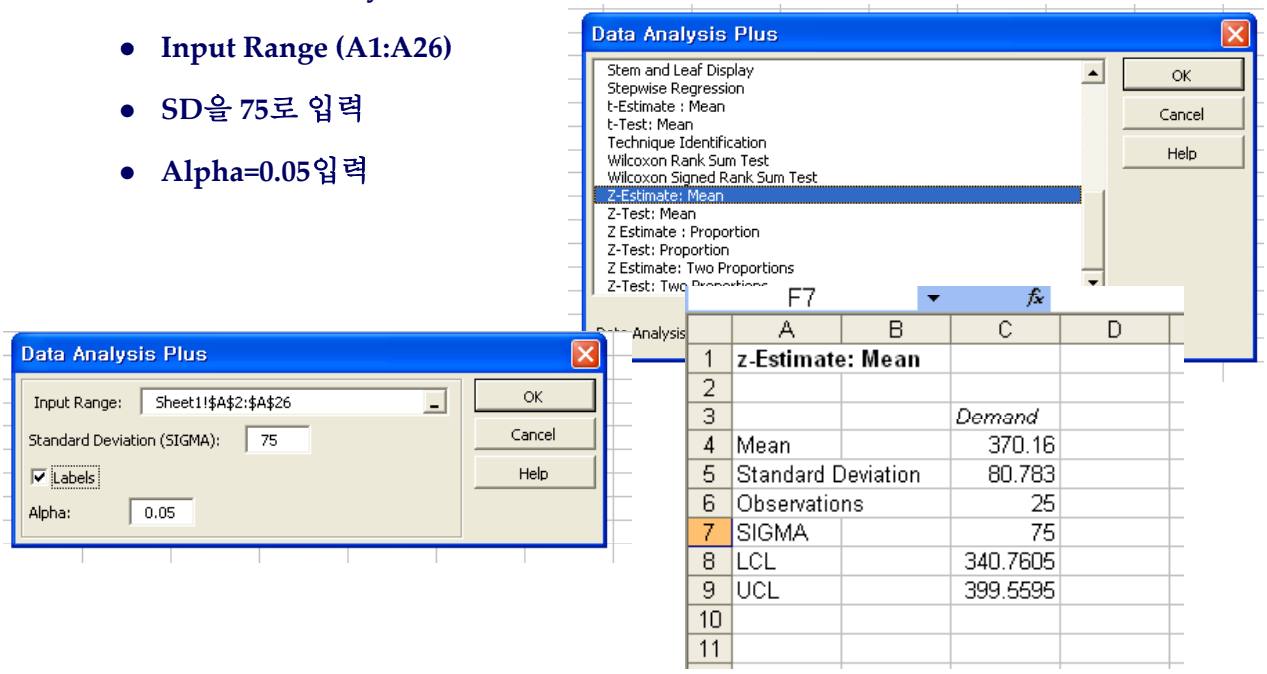
| | A |
|---|---|
| 1 | Demand |
| 2 | 235 |
| 3 | 421 |
| 4 | 394 |
| 5 | 261 |
| 6 | 386 |
| 7 | 374 |
| 8 | 361 |
| 9 | 439 |
| 10 | 374 |
| 11 | 316 |
| 12 | 309 |
| 13 | 514 |
| 14 | 348 |
| 15 | 302 |
| 16 | 296 |
| 17 | 499 |
| 18 | 462 |
| 19 | 344 |
| 20 | 466 |
| 21 | 332 |
| 22 | 253 |
| 23 | 369 |
| 24 | 330 |
| 25 | 535 |
| 26 | 334 |

**(1) 리드타임 동안의 평균 수요에 대한 95% 신뢰구간추정치를 구하라 표준편차는 오랜 경험으로 부터 75대라는 것을 알고 있다.**

**(2) 표준편차를 모를 때의 구간은?**

# Excel Application

- Data Analysis Plus 이용 –Z estimates
  - 도구-Data Analysis Plus – Z-estimate: Mean
  - Input Range (A1:A26)
  - SD을 75로 입력
  - Alpha=0.05입력






# Excel Application

- t-estimate