2024년 1학기 경영통계와 의사결정

중간고사 대체과제

한국외국어대학교 경영대학원 이순희 교수

학번: 202368020 이름: 김재환

안내사항: 아래 3가지 질문에 대한 답변을 작성하되 다음 사항을 지키기 바랍니다. 제출 파일 형식은 pdf 형식으로 하기 바랍니다.

- 각 페이지 당 답변 한 개만을 작성.
- 다음 세 문제 중 두 문제를 선택.
- 총 4 페이지를 넘지 않도록 함.
- 모든 문제는 엑셀 또는 R을 이용하여 계산.
- 엑셀/R 작업 내용의 일부를 copy/paste하여 설명하고, 작업파일은 추가자료로 제출.
- 과제의 평가는 제출된 답변을 기준으로 할 것이므로, 엑셀/R을 이용한 작업 파일을 제출하더라도, 가능한 한 답변 시트에 필요한 사항을 모두 기입.

1. (데이터를 수집할 수 있는) 관심있는 주제를 정해서 데이터를 수집하고 R 또는 엑셀을 이용해서 확률분포를 만들고, 의미있다고 생각되는 통계량을 구한 뒤 결과물을 설명하시오. (다른 문헌을 참고시 인용)

한국갤럽 2023년 11월 2일 ~ 12월 4일 조사(갤럽리포트 G20240311)

다음은 2023년 10~12월 전화/ 온라인/ 면접조사를 통해 얻은 결과이다.

조사대상은 전국 만 19세이상 1550명이며 표본오차는 +- 2.5%(95%신뢰수준) 이다.

응답률은 29.8%(총 접촉 5,196명 중 1,550명 응답 완료)이다.

질문은 "만약 우리나라가 연루된 전쟁이 일어난다면 귀하는 나라를 위해 기꺼이 싸우시겠습니까, 싸우지 않겠습니까?" 이며

결과는 전체 찬성 46%, 반대 36%, 모름 및 응답 거절 18% 이다.

위의 층화표본추출 내용을 의도적으로 단순화하여 분석하기 쉽도록 모든 응답을 동일한 확률로 취급하여 이항분포 B(713, 385.02), p = 46%로 분석하고자 한다.(1. 반대와 응답거절 모두 반대로 가정한다. 2. 동전 던지기 시행으로 변환)

* 평균(기댓값) = n * size * p

* 분산 = n * size * p * (1 - p)

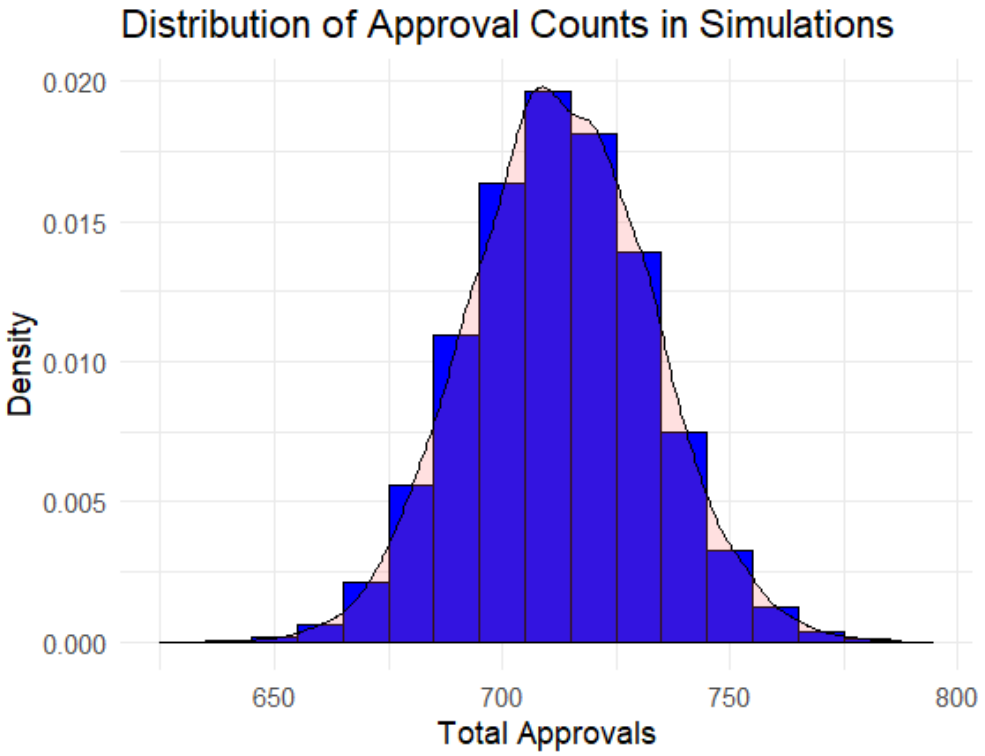
** n = 각 시뮬레이션에서의 시도 횟수, size = 각 시도에서의 성공최대 횟수, p = 성공확률

[R code]

```
# ggplot2 라이브러리 로드
library(ggplot2)

# 전체 응답자 수와 찬성 확률 설정
n_total <- 1550  # 전체 응답자 수
approval_prob <- 0.46  # 전체 찬성 비율

# 반복 횟수 설정
```

```
n_simulations <- 10000  # 시뮬레이션 횟수

# 반복적인 무작위 표본추출 시뮬레이션
set.seed(123)  # 결과의 재현성을 위한 시드 설정
simulated_approvals <- replicate(n_simulations, sum(rbinom(n_total, size=1, prob=approval_prob)))  #rbinom 함수를 활용하여 동전을 1번씩 던지는 행동을 10000번의 반복시행한다.

# 정규분포 형태로 시각화
simulated_data <- data.frame('TotalApprovals' = simulated_approvals)

# 시뮬레이션 결과의 히스토그램 그리기
ggplot(simulated_data, aes(x=TotalApprovals)) +
  geom_histogram(aes(y=..density..), binwidth = 10, fill="blue", color="black") +
  geom_density(alpha=.2, fill="#FF6666") +  # 밀도 그래프 추가
  labs(title="Distribution of Approval Counts in Simulations",
       x="Total Approvals",
       y="Density") +
  theme_minimal()

# 시뮬레이션 결과의 평균 계산
mean_approvals <- mean(simulated_approvals)

# 시뮬레이션 결과의 표준편차 계산
sd_approvals <- sd(simulated_approvals)

# 시뮬레이션 결과의 분산 계산
var_approvals <- var(simulated_approvals)

# 계산된 통계량 출력
cat("Mean of simulated approvals: ", mean_approvals, "\n")
cat("Standard deviation of simulated approvals: ", sd_approvals, "\n")
```

cat("Variance of simulated approvals: ", var_approvals, "\n")



실행 결과, TotalApprovals의 분포를 보여주는 히스토그램과 밀도 곡선이 나타납니다. 이는 전체 응답자 중 찬성을 표시할 것으로 예상되는 비율의 분포를 나타냅니다. 중심극한정리에 의해, n_simulations가 충분히 크면 시뮬레이션 결과는 정규분포에 가까워질 것으로 예상할 수 있습니다. 또한, size 가 충분히 커질수록 막대그래프의 모양은 선으로 연결된 밀도그래프와 가까워 지게됩니다.

시뮬레이션 결과의 평균은 712.8037으로 찬성 응답의 기대치를, 분산과 표준편차는 각각 398.8469와 19.97115을 나타내며 결과의 변동성을 나타냅니다.

(경영통계와 의사결정 과제 1.R 참조)

2. 여론조사 뉴스에서 다음과 같은 기사가 있다고 하자. 지난 0월 0일까지 전국 유권자 3600명을 대상으로 조사한 결과, A정당의 지지율은 42%이고, B정당 지지율은 37%를 기록하였다. 이번조사는 무작위 추출하였고, 전화 조사원 인터뷰 방식으로 진행하였다. 표본 오차는 95%신뢰수준에서 +/- 1.6% 포인트 응답률은 14%였다. 각 정당 지지율에 대한 신뢰구간을 구하고, 신뢰구간의 의미를 설명하시오. 신뢰구간을 개선시킬 수 있는 방법들에는 무엇이 있는지 한가지 이상 논하시오.

[답안]

신뢰수준 : 95% (유의수준 α = 5%)

z값(엑셀이용 / NORM.S.INV(0.975)) = 1.959963985 = 1.96

조사대상 유권자 : $\frac{x}{(1-14\%)} = 3600$, x = 4187명

x를 표본의 성공의 횟수, n은 표본크기라고 할때, 모비율의 점 추정량은

$\hat{p}_A = \frac{x}{n}$ = 42% , $\hat{p}_B = \frac{x}{n}$ = 37% 이다.

응답자 3600명 중 A정당 지지자는 1,523명 이고, B정당 지지자는 1,332명 이다.

$\hat{p}$의 표본분포는 평균이 p이고, 분산이 $\frac{p(1-p)}{n}$인 정규분포를 근사적으로 따른다.

$A \sim P(0.42, 0.00006767)$, B ~ P(0.37, 0.00006475)

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

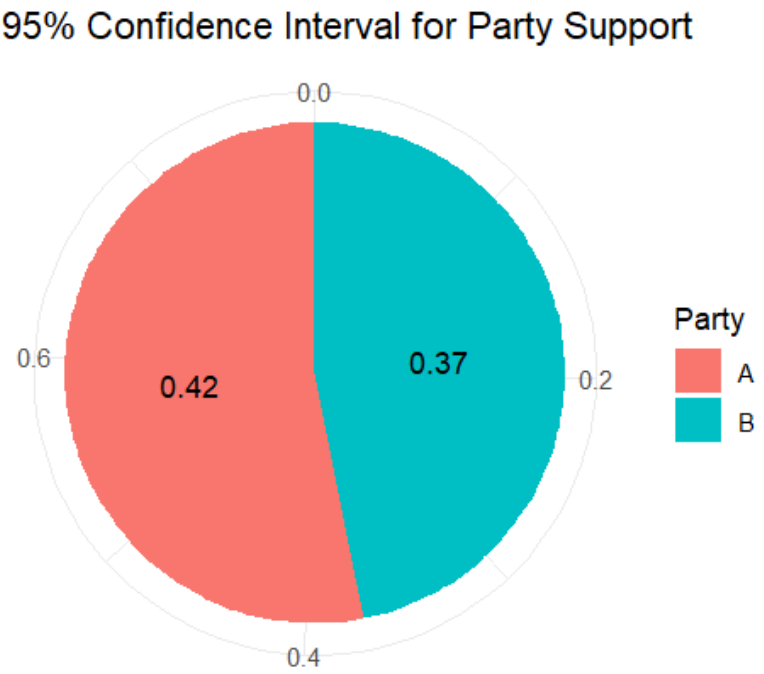
A~LCL = $\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ = 0.42 - 1.96×0.008225975 = 0.403877

A~UCL = $\hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ = 0.42 + 1.96×0.008225975 = 0.436123

B~LCL = $\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ = 0.42 - 1.96×0.008046738 = 0.354228

B~UCL = $\hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ = 0.42 + 1.96×0.008046738 = 0.385772

신뢰구간의 의미 : 우리가 같은 방법으로 많은 수의 여론조사를 실시한다면, 그 결과로 나온 지지율의 신뢰구간이 실제 지지율을 포함할 확률이 95%라는 것입니다. 다시 말해, 신뢰구간은 측정된 비율의 불확실성을 수치적으로 표현한 것이며, 실제 지지율이 이 범위 안에 있을 가능성을 제공합니다



신뢰구간을 개선시킬 방안 :

1. 표본 크기 증가: 표본의 크기를 늘리면 표준 오차가 줄어들고, 결과적으로 신뢰구간이 좁아집니다. 이는 신뢰구간의 정밀도를 향상시키는 가장 확실한 방법입니다.

2. 표본 설계 개선: 무작위 추출을 더 철저히 하거나, 세대별 계층화 표본추출과 같은 방법을 사용해 대표성을 개선할 수 있습니다. 더 대표적인 표본은 신뢰구간의 타당성을 높여 줍니다.

3. 응답률 향상: 낮은 응답률은 비응답 편향을 초래할 수 있습니다. 응답률을 높임으로써, 더 정확한 표본을 얻을 수 있고, 이는 신뢰구간의 정확도를 개선할 수 있습니다.

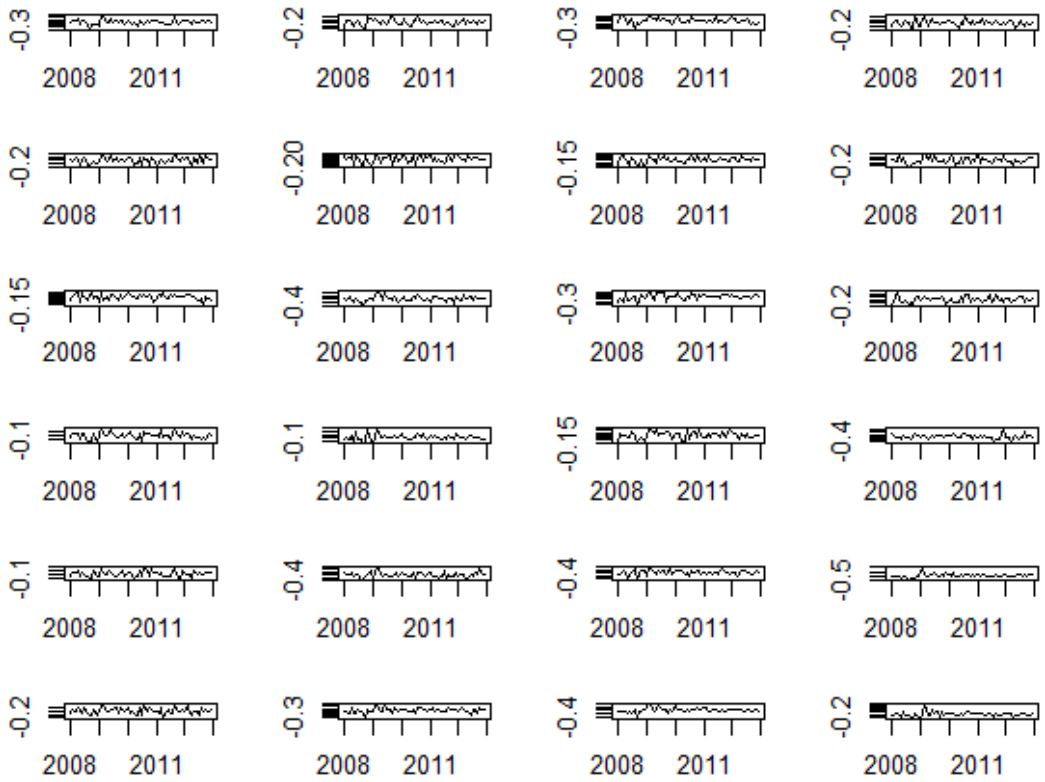
4. 데이터 수집 방법 다양화: 전화 인터뷰 방식만 사용하는 대신 온라인 조사나 우편 조사 등 다양한 방식을 병행하면 다른 집단의 의견을 더 잘 반영할 수 있어 신뢰구간의 정확도가 향상될 수 있습니다.

(경영통계와 의사결정 과제2.R, 경영통계 계산.xlsx 참조)

3. 4주차의 NASDAQ.csv 에서 (투자가치가 있다고 생각되는) 3가지 이상 종목으로 포트폴리오를 구성하여 기대수익률과 분산을 추정하시오. (답변에 기입해야 하는 사항: 투자가치가 있다고 선택한 기준에 대한 설명, 선택한 종목, 포트폴리오를 구성한 %를 최소 2가지 이상 비교, 기대수익률/분산을 계산한 방식에 대한 설명)

   주식 포트폴리오 최적화를 위하여 시계열 데이터인 NASDAQ.csv 자료를 이용하여 R과 엑셀의 기능을 활용하고 몇가지 가정에 의하여, 3가지 종목을 고르고 그에 대한 포트폴리오 분석을 실시하였습니다.

   먼저 R을 이용하여 ggplot2와 forecast 패키지를 이용하고, ts함수를 활용하여 시계열 자료를 시각화 하였습니다.



   두번째로 예측값을 위와 같이 시각화 하여 육안으로 예측값을 시각적으로 찾는 것은 매우 어렵기 때문에 기존 ts데이터를 auto.arima 함수를 활용하여 ARIMA함수로 변환하고, forecast 함수를 이용하여 주어진 자료의 마지막으로 부터 1개월의 예측치를 구하고 행렬로 저장하였습니다.

   이후 예측값을 정렬하고 가장 높은 주식 3개를 추출하였습니다.

NFLX(넷플릭스)	:	0.04051900
AMZN(아마존)	:	0.03748198
SIRI(시리우스XM 홀딩스)	:	0.02874538

(경영통계와 의사결정 과제3.R, hwdata.csv 참조)

[방법1]

이후 3개의 주식 시계열 데이터를 추출하여 엑셀 파일로 만들고 파이썬 프로그램을 활용하여 각 자산에 대한 수익률 최대화에 따른 포트폴리오 최적화를 진행하였습니다.

모듈은 pandas와 numpy, scipy.optimize.minimize를 사용하였으며, data.resample()

함수를 통하여 주어진 시계열 데이터를 연도별로 재샘플링하고, 각 연도의 합산 값을 연도의 첫번째 값으로 나누어서 연간 수익률을 계산 하였습니다.

objective(weights)함수를 활용하여 기대수익률을 정의 하였습니다.

minimize() 함수를 활용하여 최적화를 실행하였습니다.

아래는 파이썬 프로그램의 결과값 입니다. (hwdata2.csv ,optimize_portfolio.py 참조)

Optimal weight for AMZN: 6.125100426856989e-13

Optimal weight for NFLX: 0.9999999999992848

Optimal weight for SIRI: 1.0269562977782698e-13

따라서 $E_p(R)$ = 0.04052 , σ_p = 0.19096 입니다.

(hwdata2-1.xlsx_방법1 시트 참조)

[방법2]

3개의 주식 데이터들의 평균수익률(E(r)) 과 표준편차(σ), 그리고 상관계수를 구하고 각 가중치에 따라 표준편차를 구하여 위험이 최소가 되도록 엑셀로 계산 하였습니다.

우선 3가지 주식의 가중치 조합 테이블을 만들기위하여 파이썬을 이용하여 테이블을 만들었으며(table.py, result.xlsx 참조), 주어진 테이블을 바탕으로 각 주식들의 평균, 분산, 표준편차, 공분산, 상관계수 들을 구하여, 주어진 가지수의 포트폴리오의 표준편차(위험)와 기대수익을 계산 하였습니다.

마코위츠 모형에 따라 MVP (Minimum Variance Portfolio) 포트폴리오가 주어진 가지수에서 최소 분산에 가중치를 배분하였습니다. 아래는 결과 입니다.

AMZN : 0.8 , NFLX : 0.15 , SIRI : 0.05 , $E_p(R)$ = 0.02837 , σ_p = 0.0999

(hwdata2-1.xlsx_방법2 시트 참조)