ChatGPT를 활용한 실무 데이터 분석

- 5. 신뢰구간 -

presented by A.Iglue

오늘의 학습 목표

- 표본오차의 확률분포를 설명할 수 있다.
- 추론통계에서 신뢰구간의 역할을 설명하고, 추정통계량의 신뢰구간을 계산할 수 있다.

표본오차의 확률분포

- 표본오차: 표본평균 - 모평균
- 큰수의 법칙에 의해 기본적으로 표본크기가 크면, 모평균을 더욱 잘 대변하므로 표본오차의 크기가 줄어듦.
- 표본오차 확률분포: 표본의 평균값들이 모평균으로부터 얼마나 차이가 나는지, 그 차이의 분포가 어떤 확률로 나타나는지를 표현함.
- 중심극한정리에 의해 표본평균의 확률분포와 마찬가지로 표본오차의 확률분포는 정규분포를 따르며, 평균이 0, 표준편차는 $\frac{\sigma}{\sqrt{n}}$ (σ는 모표준편차)를 따름.

표준오차

■ 표준오차(Standard Error)

- ✓ 특정 통계량(ex. 평균)의 표본분포(확률분포)에서의 표준편차
- ✓ 중심극한정리에서 등장했던 표본평균의 확률분포에서 정규분포의 표준편차가 표준오차에 해당함.
- ✓ 정규분포의 평균 = 모집단의 평균(μ)
- ✓ 정규분포의 표준편차(표준오차) = $\frac{\sigma}{\sqrt{n}}$

신뢰구간

- 오차를 정량화하기 위해, 신뢰구간(confidence interval)을 도입함.
- 표준정규분포에서는 평균값(μ=0) ± 1.96*표준편차(s=1) 범위에 약 95%의 값을 포함함.
- 이 의미는 정규분포에서 무작위로 하나의 값을 꺼내면 약 95%의 확률로 그 범위 안에 포함된다는 의미임.
- 표본오차를 표준편차로 나누면, 표본평균의 표준정규분포가 됨.
- 표준정규분포의 95% 범위: $-1.96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96$
- 양변을 정리하면, $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$ 가 됨.
- 위 식이 표본크기가 적당히 클 때(n > 30) 표본평균분포가 정규분포에 근사한다는 가정 아래 설정한 모평균의 95% 신뢰구간이 됨.

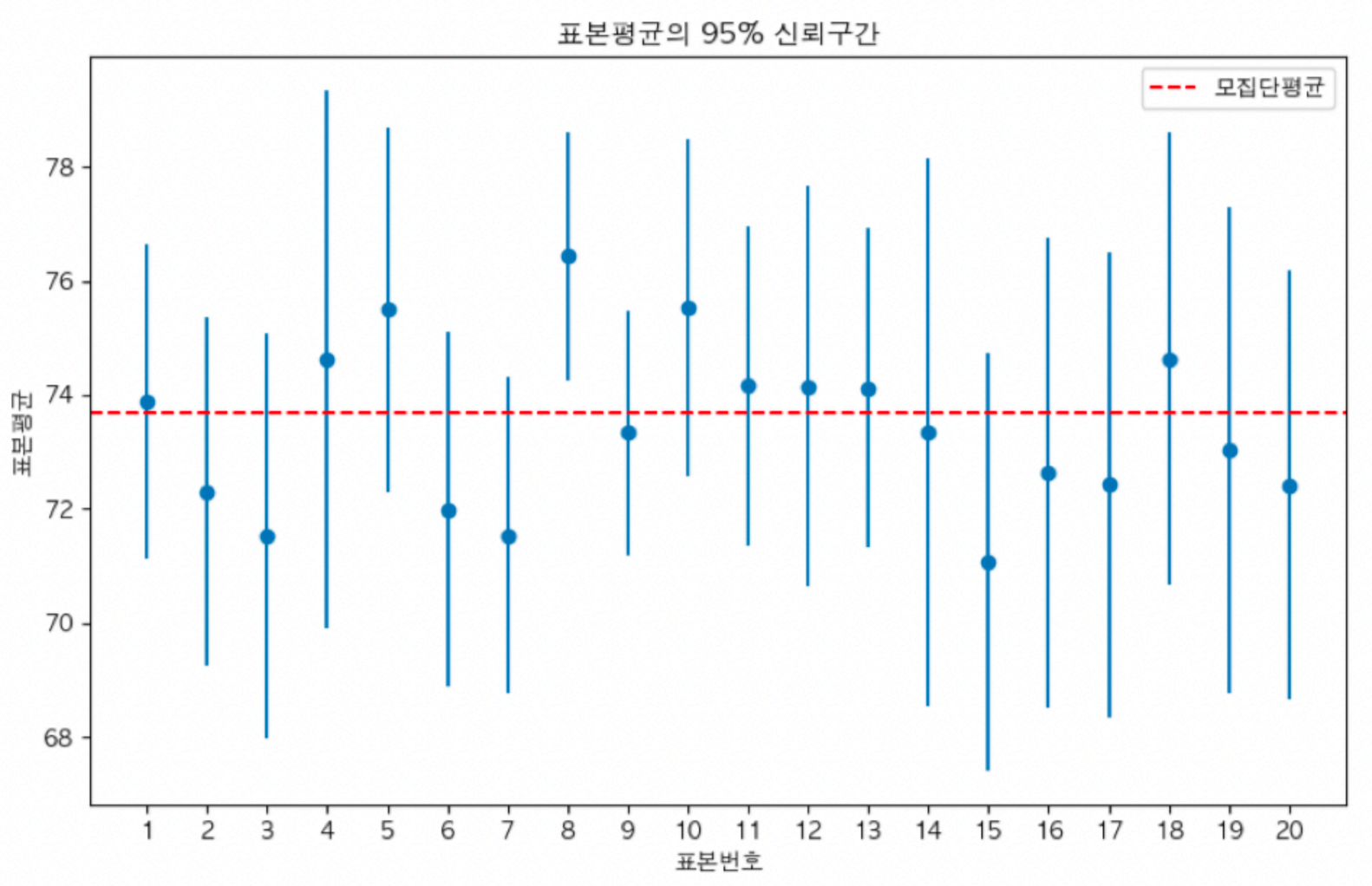
신뢰구간과 t-분포

- 만약, 모집단이 정규분포이긴 하나, 추출된 표본의 크기가 너무 작다면(n < 30), 표본평균분포는 정규분포와는 멀어짐.
- 이런 경우에는, 정규분포와는 비슷하나 표본크기에 따라 분포모양이 조금씩 달라지는 t-분포를 이용하여 신뢰구간을 설정할 수 있음.
- 신뢰구간을 설정하는 방법은 표준정규분포와 동일하나, 다만 퍼센트포인트의 위치가 달라짐.
- t-분포의 95% 범위: $\bar{x} - tvalue\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + tvalue\frac{\sigma}{\sqrt{n}}$
- cf) 표준정규분포의 95% 범위: $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$
- tvalue는 표본크기로부터 구한 자유도에 따라 조금씩 달라짐.

신뢰구간 해석하기

- “이 구간에 OO%의 확률로 모집단평균 μ가 있다.”
- 주의할 점은 여기서의 확률변수는 μ가 아니라 $\bar{x}$임.
- μ가 확률적으로 변하여 그 구간에 포함되는 것이 아닌, 모집단에서 표본을 추출하여 OO% 신뢰구간을 구하는 작업을 반복을 100번 반복했을 때, 평균적으로 그 구간에 μ가 포함되는 것이 OO번이라는 의미임.
- 하나의 표본에서 얻은 신뢰구간은 μ를 포함하거나 포함하지 않거나 둘 중 하나임.
- 신뢰구간은 표본에서 구한 모집단평균 μ의 추정값을 어느 정도 신뢰할 수 있는지를 나타내고, 신뢰구간이 좁다면 그만큼 더 신뢰할 수 있는 값이 됨.

신뢰구간 해석하기


표본평균의 95% 신뢰구간
모집단평균
78
76
74
72
70
68
표본평균
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
표본번호

신뢰구간 계산하기

■ 95% 신뢰구간: $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$

✓ 우리는 하나의 표본으로부터 이 문제를 해결해야 함.

✓ 지금까지 예시로 든 여러 개의 표본은 중심극한정리를 설명하기 위한 것일 뿐.

✓ 크기가 n인 표본으로부터 표본평균 $\bar{x}$를 구함.

✓ σ(모표준편차)를 구해야 하지만, 우리는 알 수 없음.

✓ 그 대안으로 불편표준편차(s)를 계산하여 σ를 대신함.

✓ n > 30 이면, 표준정규분포를 가정하여 위 식을 그대로 사용하고,

✓ n < 30 이면, 표본의 자유도에 따른 t-분포를 가정하여 1.96을 수정해야 함.

표준정규분포 .vs. t-분포

표준정규분포의 95% 신뢰구간 공식	t-분포의 95% 신뢰구간 공식
$\bar{x} - 1.96 \times \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \times \frac{s}{\sqrt{n}}$	$\bar{x} - t\text{값} \times \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t\text{값} \times \frac{s}{\sqrt{n}}$

t-분포

■ t-값

✓ 표본오차를 표준오차로 나눈 값

✓ 추정통계량 중 하나

✓ t-값 = $\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

■ t-분포란 모집단이 정규분포라는 가정하에, 미지의 모표준편차 σ를 표본으로 계산한 비편향표준편차 s로 대용하고, 표본평균을 표준화한 값이 따르는 분포

■ 정규분포와 비슷하게 생겼지만 다름.

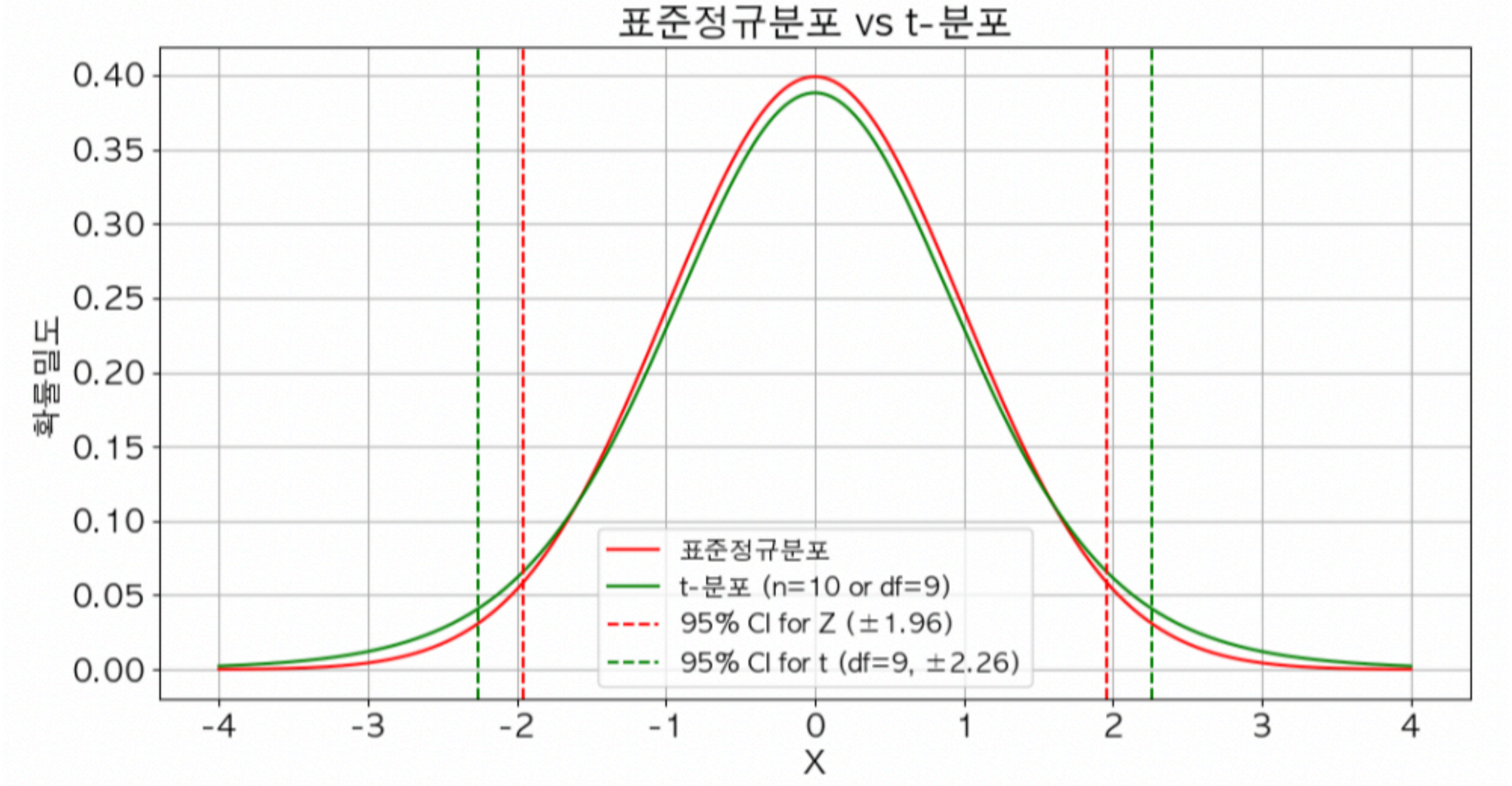
t-분포 들여다보기

- t-분포는 1908년 기네스 맥주에 근무하던 윌리엄 고셋에 의해 고안됨.
- 고셋은 맥주 효모 데이터를 분석할 때, 작은 표본으로도 모집단 전체를 추정할 수 있는 방법을 연구함.
- 큰수의 법칙이나 중심극한정리에서는 표본크기가 커질수록 모수에 근사적으로 접근하기 때문에, 실제 데이터 분석에서 작은 표본크기를 갖는 경우, 표본오차(또는 표본평균)가 정규분포를 따른다고 말할 수 없음.
  - ✓ 이런 경우, 정규분포를 가정할 수 없으므로 모수 추정이 어려워짐.
  - ✓ 이런 경우, t-분포를 사용하여 모수를 측정할 수 있음.
- t-분포는 모집단이 정규분포를 따른다는 가정 아래, 표본크기가 작은 경우 모수를 추정할 수 있는 확률 분포임.

t-분포의 특징

- t-분포는 표본크기(n)이 커짐에 따라 정규분포에 가까워짐.
- 이 말은 정규분포가 t-분포의 특수한 경우임을 의미함.
- 정규분포와 t-분포의 차이
  - ✓ 정규분포의 95% 신뢰구간은 정확하게는 표준오차의 1.96(z-값)배임.
  - ✓ 표본크기(n=10)인 경우 t-분포의 95% 신뢰구간은 정규분포보다 조금 더 넓어져 표준오차의 2.26(t-값)배가 됨.
  - ✓ 정규분포에서는 95%의 신뢰구간이라는 단서만 있어도 z-값을 구하지만, t-분포에서는 자유도(표본크기와 관련)에 따라서 분포가 달라지기 때문에 t-값을 구하기 위해서는 자유도도 반드시 필요함.
- t-분포를 활용하려면, 모집단이 반드시 정규분포이면서 표본크기가 작아야 함.
- 만약 표본크기가 크다면, 모집단이 정규분포가 아니라고 하더라도 중심극한정리에 의해 표본평균분포를 정규분포로 근사할 수 있음.

표준정규분포 .vs. t-분포



추측통계의 정밀도

- 추측통계의 정밀도를 높이는 방법
- 신뢰 가능한 평균값을 추정하고 싶다면, 오차분포의 너비를 나타내는 표준오차 $\frac{s}{\sqrt{n}}$를 작게 만드는 게 중요함.
- 비편향표준편차 s를 작게 하거나, 분모인 표본크기 n을 크게하는 방안이 있음.
- s는 모집단 데이터 퍼짐이라는 모집단 그 자체의 성질에서 유래하기에 작게 만들기란 어렵지만, 측정한 데이터 퍼짐(변동) 정도를 줄일 수는 있음.
- 표본크기 n을 크게 하면 좋지만, 비용과 전수조사의 어려움 때문에 한계가 있음.
- 또한 2배의 정밀도를 높이기 위해서는 4배의 표본크기가 커져야 함

신뢰구간 연습문제

■ 표본크기가 10인 몸무게 데이터로부터, 표본평균, 표본불편표준편차, 표준오차, 95% 신뢰구간을 구하시오.

✓ 표본 데이터단위: kg): [78.82, 72.00, 74.89, 81.20, 79.34, 65.11, 74.75, 69.24, 69.48, 72.05]

✓ 표본크기(Sample Size): 10

✓ 표본평균(Sample Mean): 약 73.69 kg

✓ 표본 불편 표준편차(Sample Unbiased Std): 약 5.10 kg

✓ 표준오차(Standard Error): 약 1.61 kg

✓ 95% 신뢰구간(95% Confidence Interval): 약 70.52 kg ~ 76.84 kg

오늘의 실습

- 표준정규분포를 이용한 신뢰구간 계산하기
- t-분포를 이용하여 신뢰구간 계산하기