통계 분석 실무 QA

Q1. 상관관계 분석은 언제 피어슨(Pearson), 스피어만(Spearman)을 써야 하나요?
A1.
1. Pearson: 두 변수 모두 연속형이고, 정규성을 만족하며 선형 관계가 있는 경우 사용합니다.
2. Spearman: 변수 간 순위 상관관계를 보고자 할 때 사용하며, 비선형이거나 이상치가 있을 때 강건합니다.
3. 정규성/선형성 가정을 시각화(산점도, 히스토그램)와 정규성 검정(Shapiro-Wilk 등)으로 판단합니다.

Q2. 공분산과 상관계수는 어떻게 다르고, 실무에선 무엇을 더 많이 쓰나요?
A2.
1. 공분산은 두 변수의 방향성만 알려주며 단위에 의존적입니다.
2. 상관계수는 -1~1 사이 값으로 표준화되어 해석이 쉬워 실제 분석에서는 훨씬 더 많이 사용됩니다.
3. 보고서나 모델 입력에서는 보통 상관계수를 활용하고, 변수 선택 시 collinearity 판단에도 사용합니다.

Q3. 이항 로지스틱 회귀는 어떤 경우 쓰는 게 적절한가요?
A3.
1. 종속변수가 **이진값(0/1)**인 경우 사용하며, 예: 이탈 여부, 전환 여부, 구매 유무 등에 적합합니다.
2. 입력변수와의 관계는 선형성이 필요 없으며, 결과 해석은 오즈비(Odds Ratio) 기반으로 진행합니다.
3. 확률 기반 예측이 가능하므로 마케팅, 리스크 예측 등에서 자주 사용됩니다.

Q4. 회귀 분석에서 변수 간 다중공선성은 어떻게 확인하고 처리하나요?
A4.
1. **VIF(Variance Inflation Factor)**가 5 또는 10 이상이면 공선성이 있다고 판단합니다.
2. 높은 공선성 변수는 **삭제하거나 차원 축소(PCA, Lasso)**로 처리합니다.
3. 모델 성능에는 큰 영향이 없어도 해석이 왜곡될 수 있어, 특히 설명력이 중요한 분석에선 주의가 필요합니다.

Q5. 주성분분석(PCA)은 어떤 상황에서 적절한가요?
A5.
1. 다차원 데이터의 차원 축소, 시각화, 또는 노이즈 제거에 사용됩니다.
2. 각 변수 간 상관관계가 클수록 PCA 효과가 좋으며, 정규화(Standardization) 후 적용하는 것이 일반적입니다.
3. 예: 고객 행동 데이터, 이미지 데이터, 문서 벡터화 등에서 특징을 압축할 때 사용됩니다.

Q6. 독립표본 t-검정과 대응표본 t-검정은 어떻게 구분하나요?
A6.
1. 독립표본 t-검정: 두 그룹이 서로 다른 집단 (예: 남성과 여성, 신규 vs 기존 고객)
2. 대응표본 t-검정: 두 그룹이 쌍을 이루는 경우 (예: 같은 고객의 전후 비교, A/B 테스트에서 동일 대상 평가)
3. 잘못 사용하면 통계적 검정력 저하 및 오류 가능성이 높습니다.

Q7. 정규성 가정이 깨졌을 때 대체 검정 방법은?
A7.
1. t-검정 → Wilcoxon rank-sum test (독립), Wilcoxon signed-rank test (대응)
2. ANOVA → Kruskal-Wallis test
3. 분포가 왜곡된 경우 비모수 검정이 더 신뢰도 높은 결과를 줄 수 있습니다.

Q8. A/B 테스트에서 효과 검정 외에 고려해야 할 통계적 개념은?
A8.
1. 표본 크기 결정(Power Analysis): 효과를 잡기 위해 사전 계산 필수
2. 다중 비교 보정(Bonferroni, FDR): 여러 지표 평가 시 유의수준 조정
3. Test duration bias, novelty effect, early stopping 오류 등 실험 설계 주의 필요

Q9. 이상치 처리는 어떤 방식이 좋은가요?
A9.
1. 사전 탐색 시각화(Boxplot, Z-score 등) 후 도메인 전문가와 협의
2. 일반적으로 Z-score > 3, 또는 IQR 기준 외 범위값을 이상치로 판단
3. 삭제 외에도 대체(중앙값, 윈저라이징), 로그변환, 모델 기반 감지(Isolation Forest 등) 방법도 있음

Q10. 범주형 변수 간 연관성 검정에는 어떤 방법이 있나요?
A10.
1. 카이제곱 검정(Chi-square test): 교차표 기반으로 독립성 검정
2. 예시: 고객 성별과 상품 구매 여부 간 연관성
3. 셀 빈도 수가 적을 경우 Fisher의 정확 검정이 대안이 될 수 있습니다.

Q11. 회귀 계수 해석 시 표준화 회귀계수를 쓰는 이유는?
A11.
1. 변수 단위가 다를 경우 영향력 비교가 어려운데, 표준화 회귀계수는 표준편차 기준으로 비교 가능하게 합니다.
2. 예시: "변수 A는 변수 B보다 모델에 2배 더 영향을 준다"와 같은 상대적 중요도 판단에 사용됩니다.
3. 실무 리포트에서 변수 기여도 설명 시 유용합니다.

Q12. 로그변환은 언제, 왜 사용하는가요?
A12.
1. 변수 분포가 **비대칭적(우측 치우침)**일 때, 정규성 가정 만족을 위해 사용
2. 이상치 영향 완화, 선형 모델의 가정 충족, 해석의 직관성 향상 목적
3. 단, 0 이하 값이 있으면 로그 변환 전에 상수 추가 필요

Q13. 분산분석(ANOVA)은 어떤 경우에 적합한가요?
A13.
1. 두 개 이상의 그룹 간 평균 차이 검정이 필요할 때 사용
2. 예시: 3개 지점의 고객 만족도 평균 차이
3. 사후 분석(Post-hoc Test, Tukey 등)을 통해 어느 그룹 간 차이가 있는지도 분석 가능

Q14. 이항 데이터의 그룹 간 비율 비교는 어떻게 하나요?
A14.
1. Z-test for proportions 또는 Chi-square test를 사용합니다.
2. 예시: 남성과 여성의 전환율 차이 검정
3. 표본 수가 작을 경우 정확 검정(Fisher) 고려

Q15. 시계열에서 자기상관(autocorrelation)은 왜 중요하고 어떻게 확인하나요?
A15.
1. 자기상관은 시계열 데이터의 시간 간격별 상관성을 의미하며, 잔차 독립성 판단에 필수
2. ACF/PACF 플롯을 통해 확인하며, 잔차에 자기상관이 남아 있으면 모델 재조정 필요

3. 예시: ARIMA, SARIMA 모델 적합 전에 확인