(보안데이터분석) 연습문제_04

1. 다음 중 틀린 문장은?
   (단, 모든 변수는 이산형 난수 변수이며, X는 표본 데이터, μ는 모집단 평균, E(X)는 X의 기댓값을 나타낸다.)
   1. 표본 평균은 모집단 평균의 추정값이다.
   2. 기댓값은 모집단 평균과 같다. ❤
   3. 표본 평균은 표본 데이터의 합을 표본 크기로 나눈 값이다.
   4. 기댓값은 모든 가능한 값에 대한 확률 곱을 더한 값이다.
   5. 표본 크기가 커질수록 표본 평균은 기댓값에 가까워진다.
2. 다음 중 분산에 대하여 틀린 문장은?
   (단, 모든 변수는 이산형 난수 변수이며, X는 표본 데이터, μ는 모집단 평균, σ^2는 모집단 분산, s^2는 표본 분산을 나타낸다.)
   1. 분산은 평균으로부터 각 데이터 값의 거리를 평균한 값이다. ❤
   2. 표본 분산은 모집단 분산의 추정값이다.
   3. 분산은 표준편차의 제곱이다.
   4. 표준편차는 분산의 제곱근이다.
   5. 표본 크기가 커질수록 표본 분산은 모집단 분산에 가까워질 가능성이 커진 진다.
3. 다음 Permutation에 관한 것 중 틀린 문장은?
   (단, n개의 서로 다른 원소를 가진 집합에서 k개를 선택하여 순서대로 나열하는 경우의 수를 P(n, k)라고 한다.)
   1. P(n, k) = n! / (n-k)!
   2. P(n, n) = n!
   3. P(n, 1) = n
   4. P(n, 2) = n(n-1)
   5. P(n, k) = P(n, n-k) ❤
4. 다음 Combination에 관한 것 중 틀린 것은?
   (단, n개의 서로 다른 원소를 가진 집합에서 k개를 선택하는 경우의 수를 C(n, k)라고 한다.)
   1. C(n, k) = n! / (k!*(n-k)!)
   2. C(n, 0) = 1
   3. C(n, 1) = 1 ❤
   4. C(n, n) = 1
   5. C(n, k) = C(n, n-k)

5. 다음 정규분포에 관한 내용 중 틀린 문장은?
    1. 정규분포는 평균과 표준편차가 정해진 대칭적인 분포이다.
    2. 정규분포의 확률밀도함수는 종 모양의 곡선으로 나타낸다.
    3. 정규분포의 곡선 아래 면적은 확률을 나타낸다.
    4. 정규분포에서 평균에서 1 표준편차 떨어진 구간에 포함되는 데이터의 비율은 약 95%이다. ❤
    5. 정규분포에서 평균에서 2 표준편차 떨어진 구간에 포함되는 데이터의 비율은 약 95%이다.
6. 중심극한정리에 대한 다음 내용 중 틀린 문장은?
    1. 중심극한정리는 모집단이 정규분포를 따르지 않더라도 표본 크기가 충분히 크면 표본평균의 분포는 정규분포를 따른다는 정리.
    2. 중심극한정리는 표본 크기가 n 이상이면 성립. (n은 정해진 값) ❤
       : 일반적으로 샘플(표본)이 30개 이상이면 ($n \geq 30$)근사적으로 성립한다고 하지만 항상 그런 것은 아니다.
    3. 중심극한정리는 표본 평균의 평균은 모집단 평균과 같다.
       : 정확히는 표본 평균의 기대값이 모집단의 평균과 동일하다. $E(\bar{X}) = \mu$
    4. 중심극한정리는 표본 평균의 표준편차는 모집단 표준편차를 √n으로 나눈 값과 같다.
       : 모집단의 표준편차를 σ 라고 하면 $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
    5. 중심극한정리는 모집단의 분포가 정규분포라면 표본 평균의 분포도 정규분포이다.
7. 공분산에 관한 내용이다. 다음 중 틀린 문장은?
    1. 공분산은 두 변수의 **선형적** 관계를 나타내는 척도이다.
    2. 공분산은 두 변수의 평균값을 빼서 계산된다.
    3. 공분산은 양의 값, 음의 값, 또는 0이 될 수 있다.
    4. 공분산이 0이면 두 변수는 서로 **독립적**이다. ❤ : 아닐 수도 있다.
    5. 공분산의 단위는 두 변수의 단위의 곱이다.
8. 상관계수에 관한 내용이다. 다음 중 틀린 문장은?
    1. 상관계수는 두 변수의 **선형적** 관계의 강도를 나타내는 척도이다.
    2. 상관계수는 -1에서 1 사이의 값을 가지며, 0에 가까울수록 두 변수는 서로 **독립적**이다.
    3. 상관계수는 단위를 가지지 않는다.
    4. 상관계수는 공분산을 두 변수의 표준편차의 곱으로 나누어 계산된다.
    5. 상관계수가 0.5라면 두 변수는 완벽한 양의 상관관계를 가지는 것이다. ❤
9. T-검정에 대한 내용이다. 다음 중 틀린 문장은?
    1. T-검정은 두 모집단의 평균값이 서로 **동일**한지 검증하는 통계 검정 방법이다.
    2. T-검정은 모집단의 표준편차가 알려져 있지 않아도 사용할 수 있다.
    3. T-검정은 모집단이 정규분포를 따르는지 검증하는 방법이다. ❤

    4. T-검정은 표본 크기가 작은 경우에도 검정력은 낮아지지만 사용할 수 있다.
    5. T-검정은 유의수준과 p-값을 사용하여 결과를 판단한다.
10. ANOVA(분산분석)에 관하여 다음 중 틀린 문장은?
    1. ANOVA는 두 개 이상의 모집단 평균값을 비교하는 통계 검정 방법이다.
    2. ANOVA는 모집단의 분포가 정규분포를 따르는지 검증하는 방법이다. ❤
    3. ANOVA는 각 모집단에서 오는 오차는 동일하다고 가정.
    4. ANOVA는 F-통계량을 사용하여 결과를 판단이다.
    5. ANOVA는 다양한 설계에 따라 사용될 수 있다.
11. 카이제곱 검정에 관하여 다음 중 틀린 문장은?
    1. 카이제곱 검정은 두 범주형 변수 간의 독립성을 검증하는 통계 검정 방법이다.
    2. 카이제곱 검정은 기댓값과 관측값의 차이를 제곱하여 계산한다.
    3. 카이제곱 검정은 카이제곱 분포를 사용하여 결과를 판단한다.
    4. 카이제곱 검정은 표본 크기가 작은 경우에도 사용할 수 있다. ❤
    5. 카이제곱 검정은 유의수준과 p-값을 사용하여 결과를 판단한다.
12. 다음 중 이항분포와 정규분포의 관계에 대한 설명으로 틀린 것은?
    1. 이항분포는 성공 확률이 p인 독립적인 시행을 n번 반복했을 때의 성공 횟수를 나타내는 분포이다.
    2. n이 충분히 크고 p가 0.5에 가까울수록 이항분포는 정규분포에 가까워진다.
    3. 이항분포를 정규분포로 근사할 때, 정규분포의 평균은 np이고 분산은 np(1-p)이다.
    4. 이항분포를 정규분포로 근사하면 복잡한 이항분포 계산을 더 간단하게 할 수 있다.
    5. 이항분포를 정규분포로 근사할 때, n이 작을수록 근사의 정확도는 높아진다. ❤
13. 다음 중 **회귀분석에서 상관관계와 인과관계의 차이**에 대한 설명으로 **틀린 것**은? ***
    1. 회귀분석에서 독립변수와 종속변수 사이에 높은 상관관계가 있다고 해서 반드시 인과관계가 성립하는 것은 아니다.
    2. 상관관계는 두 변수 간의 선형적인 관계를 나타내지만, 인과관계는 한 변수가 다른 변수에 영향을 미치는 관계를 의미한다.
    3. 회귀분석을 통해 변수 간의 관계를 분석하면 인과관계를 확정할 수 있다. ❤

       : 회귀분석 모델(회귀식)만으로는 상관관계를 분석할뿐이며, 인과관계를 확정 지을 수 없다. 따라서 실험 또는 연구적 통제나 연구모형을 통해서 인과관계에 대한 해설적 가설이 필요하다.
    4. 외생 변수(숨은 변수)가 존재할 경우, 두 변수 간에 높은 상관관계가 있어도 직접적인 인과관계는 아닐 수 있다.

    5. 실험적 연구(예: 무작위 대조 실험)를 통해 인과관계를 보다 확실하게 검증할 수 있다.

14. 다음 중 결정 계수(R-squared: R^2)에 대한 설명으로 틀린 것은?
    1. 결정 계수는 회귀 모델이 종속 변수의 분산을 얼마나 잘 설명하는지 나타내는 지표이다.
    2. 결정 계수는 0과 1 사이의 값을 가지며, 1에 가까울수록 모델의 설명력이 높다.
    3. 결정 계수는 독립 변수의 개수가 증가할수록 감소하는 경향이 있다. ❤ : 일반적으로 설명 변수가 증가하면 데이터 대한 설명력이 증가하기 때문에 결정계수는 증가한다. (overfitting을 생각해 보자)
    4. 결정 계수가 높다고 해서 반드시 모델이 데이터를 완벽하게 설명하는 것은 아니다.
    5. 결정 계수는 상관 계수의 제곱과 같다.

15. 다음 중 L1, L2 정규화와 Lasso, Ridge 선형 모델에 대한 설명으로 틀린 것은?
    1. L1 정규화는 Lasso 회귀 모델에서 사용되며, 일부 가중치를 정확히 0으로 만들어 변수 선택의 효과를 가진다.
    2. L2 정규화는 Ridge 회귀 모델에서 사용되며, 가중치를 0에 가깝게 줄이지만 완전히 0으로 만들지는 않는다.
    3. Lasso 모델은 Ridge 모델에 비해 모델의 복잡성을 줄이고, 특성 선택에 유리하다.
    4. L1 정규화는 가중치의 제곱합을 최소화하고, L2 정규화는 가중치의 절대값 합을 최소화한다. ❤
    5. 정규화는 선형 회귀 모델의 과적합(Overfitting) 문제를 해결하기 위해 사용된다.

16. 다음 중 L1, L2 정규화와 예측 오차 또는 이상치 입력에 대한 설명으로 틀린 것은?
    1. L1 정규화는 이상치에 민감하게 반응하여 예측 오차를 크게 증가시킬 수 있다.
    2. L2 정규화는 이상치에 덜 민감하며, 예측 오차의 변동성을 줄이는 데 효과적이다.
    3. L1 정규화는 일부 가중치를 0으로 만들어 변수 선택의 효과를 가지므로, 이상치에 영향을 받는 변수를 제거할 수 있다.
    4. L2 정규화는 모든 가중치를 0에 가깝게 줄여 이상치의 영향을 분산시키므로, 예측 오차를 안정화하는 데 도움이 된다. : 제거하는 것은 아니다.
    5. 일반적으로 이상치가 많은 데이터셋에서는 L2 정규화보다 L1 정규화를 사용하는 것이 예측 성능 향상에 더 유리하다. ❤ : 변수를 제거(가지치기)하는 것이 아니라 측정에 대한 오류가 있어서 이를 고려하여야 하는 상황에서는 L2를 사용하는 것이 변수를 제거하지는 않아 설명력을 지키면서도 이상치에 안정적일 수 있다.