MA409: Statistical Data Analysis (SAS)

Assignment 1 (Feb 27 – Mar 19)

1.1 The sample standard deviation is defined as $S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}$. Why are we dividing $\sum_{i=1}^{n}(X_i - \bar{X})^2$ by $n-1$ instead of n? Please provide the corresponding mathematical justification. (10 points)

Solution: Dividing $\sum_{i=1}^{n}(X_i - \bar{X})^2$ by $n-1$, we are able to estimate the population standard deviation unbiasedly. The mathematical justification is shown below: (assuming $X_1, X_2, \cdots, X_n$ are i.i.d. with mean μ and variance σ^2)

$$E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = E\left(\sum_{i=1}^{n}[(X_i - \mu) - (\bar{X} - \mu)]^2\right)$$

$$= E\left(\sum_{i=1}^{n}[(X_i - \mu)^2 - 2(\bar{X} - \mu)(X_i - \mu) + (\bar{X} - \mu)^2]\right)$$

$$= \sum_{i=1}^{n} E(X_i - \mu)^2 - n(\bar{X} - \mu)^2 = n\sigma^2 - nE(\bar{X} - \mu)^2 \quad \cdots (1)$$

For $E(\bar{X} - \mu)^2$, we have

$$E(\bar{X} - \mu)^2 = \frac{1}{n^2} E\left(\sum_{i=1}^{n}(X_i - \mu)\right)^2 = \frac{1}{n^2} E\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - \frac{2}{n^2} E\left(\sum_{i=1}^{n}\sum_{j=i+1}^{n}(X_i - \mu)(X_j - \mu)\right)$$

The second term is zero due to independency so that $E(\bar{X} - \mu)^2 = \sigma^2/n$, plugging it into Equation (1) we are able to get:

$$E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = n\sigma^2 - \sigma^2 = (n-1)\sigma^2.$$

Therefore, dividing $\sum_{i=1}^{n}(X_i - \bar{X})^2$ by $n-1$ gives an unbiased estimate of σ^2.

1.2 Show that the population kurtosis of a normal distribution is 3. (10 points)

Solution: Let $X \sim N(\mu, \sigma^2)$, then $Y = (X - \mu)/\sigma \sim N(0,1)$ and $Z = Y^2 \sim \chi_1^2$, then the population kurtosis of X is computed as

$$\frac{E(X - \mu)^4}{\sigma^4} = EY^4 = EZ^2 = [E(Z)]^2 + Var(Z) = 1 + 2 = 3.$$

1.3 Let X and Y be two continuous variables. If the Pearson's correlation coefficient of X and Y is 0, then X and Y are "uncorrelated"; if the joint probability density of X and Y

equals the product of the densities of X and Y, i.e., $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, then X and Y are "independent". Show that if X and Y are independent, they must be uncorrelated. Then use a counter-example to show that if X and Y are uncorrelated, they are not necessarily independent. (10 points)

Solution: If X and Y are independent, then

$$E[(X-\mu_X)(Y-\mu_Y)] = E(X-\mu_X)E(Y-\mu_Y) = 0$$

Therefore, the Pearson's correlation coefficient

$$\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} = 0.$$

Suppose X follows a distribution that is symmetric about zero and $Y = X^2$, then $\mu_X = 0$

$$\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} = \frac{E[X(X^2-\mu_Y)]}{\sigma_X\sigma_Y} = \frac{EX^3}{\sigma_X\sigma_Y} = 0$$

This suggest that X and Y are uncorrelated, but Y is fully determined by X, they are not independent.

1.4 Consider a hypothetical clinical trial involving liver cirrhosis patients, the prothrombin index (a measure of liver function, higher value suggests better liver function) is recorded at study entry and 10 follow-up visits (one every 3 months). Not all the patients have 11 prothrombin index records, as some patients have missing records due to death of liver cirrhosis. Which type of missing is in the data? State your thoughts and rationale. (10 points)

Solution: The missingness in the data is highly likely to be missing not at random (MNAR). Patients with lower prothrombin index (indicating bad liver function) are more likely to die of liver cirrhosis. This suggests that the prothrombin index is more likely to be missing when its value is low. Therefore, the missingness is very likely to be missing not at random.

The solution for 1.5 is provided in the *Assignment1.sas* program file.