Chinmayee Verraguntla
DS 3001: Introduction to ML

HW 3: EDA, Problem 1

1) Show $m(a+bX) = a+b \times m(X)$

$$m(a+bX) = \frac{1}{N}\sum_{i=1}^{N}(a+bx_i)$$
$$m(a+bX) = \frac{1}{N}\left(\sum_{i=1}^{N} a + \sum_{i=1}^{N} bx_i\right)$$
$$m(a+bX) = \frac{1}{N}\left(Na + b\sum_{i=1}^{N} x_i\right)$$
$$m(a+bX) = a + \frac{b}{N}\sum_{i=1}^{N} x_i =$$
$$m(a+bX) = a+b\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) =$$
$$\boxed{a+b\cdot m(X)}$$

2) Show that $cov(X, a+bY) = b \times cov(X,Y)$

$$cov(X, a+bY) = \frac{1}{N}\sum_{i=1}^{N}(x_i - m(X))((a+by_i) - m(a+bY)) =$$
$$\frac{1}{N}\sum_{i=1}^{N}(x_i - m(X))((a+by_i) - (a+b\cdot m(Y))) =$$
$$\frac{1}{N}\sum_{i=1}^{N}(x_i - m(X))(by_i - b * m(Y)) =$$
$$\frac{1}{N}\sum_{i=1}^{N}(x_i - m(X))\, b(y_i - m(Y)) =$$
$$b\cdot\frac{1}{N}\sum_{i=1}^{N}(x_i - m(X))(y_i - m(Y)) =$$
$$\boxed{b\cdot cov(X,Y)}$$

3) Show that $cov(a+bX, a+bX) = b^2 cov(X,X)$ & $cov(X,X) = s^2$

sub $Y = X$ & $a+bY = a+bX$ from (2); $cov(X, a+bY) = b\cdot cov(X,Y)$

$$cov(X, a+bX) = b\cdot cov(X,X)$$
$$cov(a+bX, X) = cov(X, a+bX)$$
$$cov(a+bX, a+bX) = \frac{1}{N}\sum_{i=1}^{N}((a+bx_i) - m(a+bX))((a+bx_i) - m(a+bX))$$
$$\frac{1}{N}\sum_{i=1}^{N}((a+bx_i) - (a+b\cdot m(X)))^2 =$$
$$\frac{1}{N}\sum_{i=1}^{N}(bx_i - b\cdot m(X))^2 =$$
$$\frac{1}{N}\sum_{i=1}^{N}(b(x_i - m(X)))^2 =$$
$$b^2\cdot\frac{1}{N}\sum_{i=1}^{N}(x_i - m(X))^2 = b^2\cdot s^2$$
$$cov(a+bX, a+bX) = b^2\cdot cov(X,X)$$
$$cov(X,X) = s^2$$

4) Yes, a non-decreasing transformation of the media is the median of the transformed behavior.

Explanation:

A non decreasing transformation g( ) preserves the order of the data points. The median is the value separating the lower half of the data from the upper half. Since g( ) won't change the relative ordering of the data points, the median before the transformation will be the median after the transformation.

This would apply to any quantiles, which are also defined by their position in the data. The ordering is preserved, so the value at any quantile will transform and remain at that quantile.

This answer does <u>not</u> apply to the range or IQR, which are measures of spread, not position. They're calculated based on the difference between values. While g( ) doesn't change ordering, it may change the actual values, and not necessarily by the same magnitude. So, this doesn't apply to IQR & range.

5) Is it always true that $m(g(X)) = g(m(X))$?

No, this is not always true. It is only true for linear transformations ($g = a + bx$). For non-linear transformations, however, it doesn't work.

If $X = \{1, 5\}$ and $g(x) = x^2$ then:

$m(X) = (1+5)/2 = 3$

$g(m(X)) = g(3) = 3^2 = 9$

$g(X) = \{1^2, 5^2\} = \{1, 25\}$

$m(g(X)) = (1+9)/2 = 5$

$3 \neq 5$, showing that this equality doesn't hold for non-linear transformations.