2.

(a) Since the gradient decent dimension is randomly chosen and one dimension is only chosen once in one epoch, we have

$$E\{\Delta \underline{w}(i)\} \xlongequal{LLN} \frac{\eta}{N}\sum_{i=1}^{N} \nabla_w J(w_i)$$

(b) For Batch-Gradient-Decent, we have:

$$J(w) = \sum_{i=1}^{N} J(w_i)$$

$$\text{So } \Delta W(i) = \eta \cdot \nabla J(w) = \eta \cdot \nabla \sum_{i=1}^{N} J(w_i)$$

For SGD. we have

$$E(\Delta W(i)) = (\Delta W_1 \cdot p_1 + \ldots + \Delta W_N p_N)\eta$$

since we pick with normal distribution, we have:

$$p_1 = p_2 = \cdots = p_n = \frac{1}{N}$$

$$\text{So } E(\Delta W(i)) = \frac{\eta}{N} \cdot (\Delta W_1(i) + \cdots + \Delta W_N(i))$$

$$= \frac{\eta}{N} \sum_{i=1}^{N} \nabla J(w_i)$$

Employ the linearity property of partial derivative: $\boxed{\sum \nabla f(x_i) = \nabla \sum f(x_i)}$

we have: $E(\Delta W(i)) = \frac{\eta}{N} \nabla \sum_{i=1}^{N} J(w_i) = \frac{1}{N} \cdot \Delta W(i)$, Q.E.D

Conclusion: $\Delta W(i)$ of BGD is N times of the $E\{\Delta W(i)\}$ of SGD

Explaination: BGD considers the loss of the whole data set and SGD considers one at one time. Since SGD data is normally random chosen, every data has the same probability to be chosen so it is somehow equavilant with considering the whole set. Meanwhile, it's obvious that the loss of SGD should be $\frac{1}{N}$ of the BGD for its randomness.