CONDITIONAL MEANS AND VARIANCES - PART 1 M 384G/374G

**Question**: How are conditional means E(y|x) and variances Var(y|x) related to marginal means E(y) and variances Var(y)?

**Simple example**:

Population consisting of $n_1$ men, $n_2$ women.

Y = height

X = sex

Categorical, two values: Male, Female

So there are two conditional means:

E(Y|male) = (Sum of all men's heights)/$n_1$

E(Y|female) = (Sum of all women's heights)/$n_2$

Then

Sum of all men's heights = $n_1$E(Y|male)

Sum of all women's heights = $n_2$ E(Y| female)

The marginal mean is

(Sum of all heights)/($n_1$ + $n_2$) =

$$\frac{(Sum\ of\ \ men's\ heights) + (Sum\ of\ \ women's\ heights)}{n_1 + n_2}$$

$$= \frac{n_1 E(Y \mid male) + n_2 E(Y \mid female)}{n_1 + n_2}$$

$$= \frac{n_1}{n_1 + n_2} \text{E(Y| male)} + \frac{n_2}{n_1 + n_2} \text{E(Y| female)}$$

= (proportion of males)( E(Y| male) + (proportion of females)( E(Y| female)

= (probability of male)( E(Y| male) + (probability of female)( E(Y| female)

Thus: The marginal mean is the weighted average of the conditional means, with weights equal to the probability of being in the subgroup determined by the corresponding value of the conditioning variable.

Similar calculations show: If we have a population made up of m subpopulations $pop_1$, $pop_2$, …, $pop_m$ (equivalently, if we are conditioning on a categorical variable with m values -- e.g., the age of a fish), then

$$E(Y) = \sum_{k=1}^{m} \Pr(pop_k) E(Y \mid pop_k)$$

e.g., for our fish,

$$E(\text{length}) = \sum_{k=1}^{6} \Pr(Age = k) E(Length \mid Age = k)$$

Rephrasing in terms of the categorical variable X defining the subpopulations,

$$E(Y) = \sum_{all\ values\ x\ of\ X} \Pr(x) E(Y \mid X = x)$$

The analogue for conditioning on a continuous variable X is:

$$E(Y) = \int_{-\infty}^{\infty} f_X(x) E(Y \mid x) dx,$$

where $f_X(x)$ is the probability density function (pdf) of X.