

CONDITIONAL MEANS AND VARIANCES

M 384G/374G

I. Question: How are conditional means $E(y|x)$ and variances $\text{Var}(y|x)$ related to marginal means $E(y)$ and variances $\text{Var}(y)$?

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:

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

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

Then

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

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

The marginal mean is

$(\text{Sum of all heights})/(n_1 + n_2) =$

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

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

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

$$= (\text{proportion of males})(E(Y|\text{male})) + (\text{proportion of females})(E(Y|\text{female}))$$

$$= (\text{probability of male})(E(Y|\text{male})) + (\text{probability of female})(E(Y|\text{female}))$$

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 $\text{pop}_1, \text{pop}_2, \dots, \text{pop}_m$ (equivalently, if we are conditioning on a categorical variable with m values -- e.g., the age of a fish), then

$$E(Y) = \sum_{k=1}^m \Pr(\text{pop}_k) E(Y|\text{pop}_k)$$

e.g., for our fish, $\text{pop}_k = \text{fish of age } k$, and

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

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

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

Stated in words:

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 .

Note:

1. There are analogous results for conditioning on more than one variable.
2. The analogous result for *sample means* is

$$\bar{y} =$$

II. A second (related) relationship between marginal and conditional means for populations:

Consider $E(Y \mid X)$ as a new random variable U as follows:

Randomly pick an x from the distribution of X .

The new r.v. U has value $E(Y \mid X = x)$.

Example: $Y = \text{height}$, $X = \text{sex}$

Randomly pick a person from the population in question.

$$U = \begin{cases} \mu_f = E(Y \mid X = \text{female}) & \text{if the person is female} \\ \mu_m = E(Y \mid X = \text{male}) & \text{if the person is male} \end{cases}$$

Question to think about: What might cause $P(U = u)$ to be high? [Hint: There are two ways this might arise]

Consider the expected value of this new random variable. (e.g., the expected value of the mean height for the sex of a randomly selected person from the given population. In this case, we would expect $E(U)$ to depend on the proportion of the population which is of each sex.)

If U is discrete, then

$$E(U) = \sum_{\substack{\text{All possible} \\ \text{values of } U}} P(u)u \quad [\text{Why?}]$$

Example: For $U = E(\text{height} \mid \text{sex})$, the values taken on by U are

_____ and _____,

with respective probabilities _____ and _____,

so $E(U) =$ _____,

which from Part I is just _____.

In other words,

$$E(E(\text{ht} \mid \text{sex})) =$$

The same reasoning works in general, showing that:

The expected value of the conditional means is the weighted average of the conditional means marginal mean, which from Part 1 is just the marginal mean
i.e.,

$$E(E(Y \mid X)) = \text{weighted average of conditional means} \\ = E(Y)$$