CONDITIONAL AND MARGINAL MEANS

I. Question: How are conditional means E(Y|X) and marginal means E(Y) related?

Simple example:

Population consisting of n_1 men, n_2 women.

(So size of entire population is $n_1 + n_2$.)

Y = height

X = sex

Categorical, two values: Male, Female

Conditional means:

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

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

Rearrange:

Sum of all men's heights = n_1E(Y|male)

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

Marginal mean E(Y) =

(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.*

X categorical with more than 2 values:
If the population consists 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., fish: pop_k = fish of age k, so

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

Rephrase: If the categorical variable X defines the subpopulations,

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

In words:

X Continuous:

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

$f_X(x)$ = 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

$$\overline{y} =$$

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

Consider E(Y|X) as a new random variable U:
Randomly pick an x from the distribution of X.
The new r.v. U has value E(Y|X = x).

Example: Y = height, X = sex

Randomly pick a person from the population in question.

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

What is the expected value of this new random variable?

Question: What might cause P(U = u) to be high?

Example: What is 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.)

For U discrete,

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

Example: U = E(height | sex). The values taken on by

U are and ,

with respective probabilities

and

So E(U) =

=

(from Part I)

Summary: E(E(ht|sex) =

This reasoning can be generalized to give:

The expected value of the conditional means is the weighted average of the conditional means, which by Part 1 is the marginal mean:

$$E(E(Y|X)) = E(Y)$$

III. CONDITIONAL AND MARGINAL VARIANCE

Marginal Variance: *Definition* of (population) (marginal) variance of a random variable Y:

$$\mathrm{Var}(Y) = E([Y - E(Y)]^2)$$

In words and pictures:

Useful *formula* for Var(Y):

$$\mathrm{Var}(Y) = E([Y - E(Y)]^2) = E(Y^2 - 2YE(Y) + [E(Y)]^2)$$

$$=$$

$$=$$

In words:

Conditional Variance: Similarly, we define the *conditional variance*

$$\mathrm{Var}(Y|X) = \text{Variance of } Y|X$$

$$= E([Y - E(Y|X)]^2 \mid X)$$

Note: *Both* expected values here are conditional expected values.

In words:

Additional formula for conditional variance:

$$\mathrm{Var}(Y|X) = E(Y^2|X) - [E(Y|X)]^2$$

(Derivation left as exercise for student.)

Conditional Variance as a Random Variable:

Var(Y|X) is a random variable.

Example: Y = height, X = sex for persons in a certain population

Var(height | sex) is the variable which assigns to each person in the population the variance of height for that person's sex.

IV. CONNECTING MEANS AND VARIANCES

Expected Value of the Conditional Variance: Since $Var(Y|X)$ is a random variable, it has an expected value.

$Var(Y|X) = E(Y^2|X) - [E(Y|X)]^2$ implies

$$E(Var(Y|X)) = E(E(Y^2|X) - [E(Y|X)]^2)$$

$$= E(E(Y^2|X)) - E([E(Y|X)]^2)$$

Recall: $E(E(Y|X)) = E(Y)$. Applying to Y^2:

$$(*) \quad E(Var(Y|X)) = E(Y^2) - E([E(Y|X)]^2)$$

Variance of the Conditional Expected Value: (Side calculation)

$$Var(E(Y|X)) = E([E(Y|X)]^2) - [E(E(Y|X))]^2$$

Since $E(E(Y|X)) = E(Y)$,

$$(**) Var(E(Y|X)) = E([E(Y|X)]^2) - [E(Y)]^2.$$

Putting It Together:

Adding (*) and (**):

$$E(Var(Y|X)) + Var(E(Y|X)) = E(Y^2) - [E(Y)]^2,$$
$$= Var(Y)$$

Rearranging:

$$(***) \qquad Var(Y) = E(Var(Y|X)) + Var(E(Y|X)).$$

In words: The marginal variance is the sum of the expected value of the conditional variance and the variance of the conditional means.

Consequences:

1.) The marginal (overall) variance has two components:
- the expected value of the conditional variance
- the variance of the conditional means.

Exercise …

2) (***) implies

$$\mathrm{Var}(Y) \geq E(\mathrm{Var}(Y|X)) \quad \text{(Why?)}$$

Moreover, $\mathrm{Var}(Y) = E(\mathrm{Var}(Y|X))$ if and only if $\mathrm{Var}(E(Y|X)) = 0$. This says:

3) Since $\mathrm{Var}(Y|X) \geq 0$, $E(\mathrm{Var}(Y|X))$ must also be ≥ 0. (Why?). Thus (***) implies

$$\mathrm{Var}(Y) \geq \mathrm{Var}(E(Y|X)).$$

Moreover, $\mathrm{Var}(Y) = \mathrm{Var}(E(Y|X))$ if and only if $E(\mathrm{Var}(Y|X)) = 0$.

Since $\mathrm{Var}(Y|X) \geq 0$, $E(\mathrm{Var}(Y|X)) = 0$ says that for each value of X, $\mathrm{Var}(Y|X) =$ ____.

This implies that for each value of X, Y|X is
______________ .

Thus the relationship between Y and X is
__________________.

4) Another perspective on (***)
(cf. Textbook, pp. 36 - 37)

i) E(Var(Y|X) is a weighted average of Var(Y|X)

ii) $\mathrm{Var}(E(Y|X) = E([E(Y|X) - E(E(Y|X))]^2)$

$$= E([E(Y|X) - E(Y)]^2),$$

which is a weighted average of $[E(Y|X)- (E(Y)]^2$

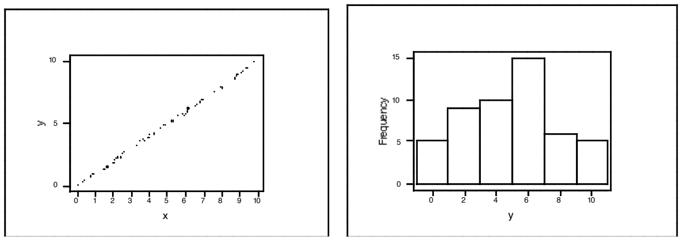
Thus, (***) says:

Var(Y) is a weighted mean of Var(Y|X) plus a weighted mean of $[E(Y|X) - E(Y)]^2$ - -

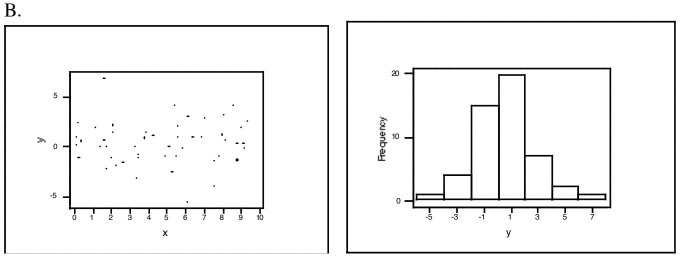
and is a weighted mean of Var(Y|X) if and only if all conditional expected values E(Y|X) are equal to the marginal expected value E(Y).)

WHAT CONTRIBUTES MOST TO VAR(Y): VAR(E(Y|X)) OR E(VAR(Y|X))?

A.



B.



C.