University of Waterloo W. H. Cherry

RESPONSE MODELS IN STAT 231: Definitions of Symbols

This Statistical Highlight #72 summarizes definitions of the symbols in the common STAT 231 response models; some terminology and notation has been slightly modified from that in the Course Notes – EPS denotes equiprobable (or 'random') selecting.

Model 1: $Y_j = \mu + R_j, \quad j = 1, 2, ..., \mathrm{n}; \quad R_j \sim G(0, \sigma);$ independent; EPS.

Y_j is a random variable whose distribution represents the possible values of the measured response variate for the jth unit in the sample of n units selected equiprobably from the respondent population, if the selecting and measuring processes were to be repeated over and over.

μ is a model parameter which represents the *average* of the measured response variate of the units of the respondent population.

R_j is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the measured value of the response variate for the jth unit in the sample of n units selected equiprobably from the respondent population, if the selecting and measuring processes were to be repeated over and over.

σ the (probabilistic) *standard deviation* of the Gaussian model for the distribution of the residual, is a model parameter which represents the (data) *standard deviation* of the measured response variate of the units of the respondent population; this (data) standard deviation (and, hence, σ) *quantifies* the *variation* of the measured response variate over the units of the respondent population – as this variation increases, so does the respondent population (data) standard deviation (and, hence, so does σ).

Model 1 is useful for a Question with a *descriptive* aspect investigated with a Plan which involves *equiprobable* selecting and a *calibrated* measuring process.
The Question usually involves the values of μ and/or σ.

Model 1a: ${}_{\mathrm{M}}Y_j = \tau + \delta + R_j, \quad j = 1, 2, ..., \mathrm{m}; \quad R_j \sim G(0, \sigma);$ independent; EPS.

${}_{\mathrm{M}}Y_j$ is a random variable whose distribution represents the possible values of the jth measurement of the response variate of a unit, if the measuring process were to be repeated over and over on this unit.

τ is a model parameter which represents the *true value* of the response variate of the unit measured m times independently.

δ is a model parameter (called the *bias*) which represents the ***in**accuracy* of the measuring process; the value of δ *quantifies* the inaccuracy of the measuring process – as inaccuracy *in*creases (*i.e.*, as accuracy *de*creases), δ *in*creases.

R_j is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the value of the jth measurement of the response variate of the unit measured m times independently, if the measuring process were to be repeated over and over on this unit.

σ the (probabilistic) *standard deviation* of the Gaussian model for the distribution of the residual, is a model parameter (called the *variability*) which represents the ***im**precision* of the measuring process and describes measuring variation if the measuring process were to be repeated over and over on a unit; the value of σ *quantifies* the imprecision of the measuring process – as imprecision *in*creases (*i.e.*, as precision *de*creases), σ *in*creases.

Model 1a is useful for a Question involving assessing the *inaccuracy* and *imprecision* of a measuring process with a Plan which involves measuring m times independently the response variate of a unit whose true value is *known*.
The Question usually involves the values of δ and σ.

If we take the response variate as $Y_j = {}_{\mathrm{M}}Y_j - \tau$, the *difference* between the *measured* value and the *true* value,

Model 1b: $Y_j = \delta + R_j, \quad j = 1, 2, ..., \mathrm{m}; \quad R_j \sim G(0, \sigma);$ independent; EPS.

thus, Model 1a rewritten as Model 1b is equivalent to Model 1, except the structural component is δ instead of μ.

Model 2: $Y_{ij} = \mu_{\mathrm{i}} + R_{ij}, \quad \mathrm{i} = 1, 2, ..., \mathrm{q}, \; j = 1, 2, ..., \mathrm{n_i}; \quad R_{ij} \sim G(0, \sigma);$ independent; EPS.

Y_{ij} is a random variable whose distribution represents the possible values of the measured response variate for the jth unit in the *sample* of $\mathrm{n_i}$ units selected equiprobably from respondent population i, if the selecting and measuring processes were to be repeated over and over.

μ_{i} is a model parameter which represents the *average* of the measured response variate for the units of respondent population i.

R_{ij} is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the measured value of the response variate for the jth unit in the sample of $\mathrm{n_i}$ units selected equiprobably from respondent population i, if the selecting and measuring processes were to be repeated over and over.

σ the (probabilistic) *standard deviation* of the Gaussian model for the distribution of the residual, is a model parameter which represents the (data) *standard deviation* of the measured response variate of the units of *each* of the q respondent populations; this (data) standard deviation (and, hence, σ) quantifies the *variation* of the measured response variate over the units of each of the q respondent populations – as this variation increases, so does each (data) standard deviation (and, hence, so does σ).

Model 2 is useful for a Question with a *causative* aspect investigated using a Plan with*out* blocking or matching.
When q = 2, the Question usually involves the value of the difference $\mu_1 - \mu_2$.

(*continued overleaf*)

Model 3: $Y_{ij} = \mu_i + \gamma_j + R_{ij}$, $i = 1, 2$, $j = 1, 2, ..., n$; $R_{ij} \sim G(0, \sigma)$; independent; EPS.

γ_j is a model parameter (called *the effect for block j*) which represents the amount by which the *average* of the measured response variate of the units in block j differs from the average of the measured response variate for the units of respondent population i; the effect for block j is assumed to be the *same* when i = 1 and i = 2.

Taking the response variate as $Y_j = Y_{1j} - Y_{2j}$, the intrapair *difference*, Model 3 becomes:

Model 3a: $Y_j = \mu_d + R_j$, $j = 1, 2, ..., n$; $R_j \sim G(0, \sigma_d)$; independent; EPS. Model 3a is Model 1 with parameters μ_d and σ_d.

Y_j is a random variable whose distribution represents the possible values of the difference in the measured response variate for the jth unit in the sample of n units selected equiprobably from the respondent population when i = 1 and i = 2, if the selecting and measuring processes were to be repeated over and over.

$\mu_d = \mu_1 - \mu_2$ is a model parameter which represents the *difference* between the *averages* of the measured response variate for the units of the respondent population when i = 1 and i = 2.

$R_j = R_{1j} - R_{2j}$ is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the difference in the measured response variate for the jth unit in the sample of n units selected equiprobably from the respondent population when i = 1 and i = 2, if the selecting and measuring processes were to be repeated over and over.

σ_d the (probabilistic) *standard deviation* of the Gaussian model for the distribution of the residual R_j, is a model parameter which represents the (data) *standard deviation* of the measured difference in the value of the response variate of the units of the respondent population when i = 1 and i = 2; this (data) standard deviation (and, hence, σ_d) quantifies the *variation* of the measured difference in the response variate over the units of the respondent population when i = 1 and i = 2 – as this variation increases, so does σ_d.

Model 3 is useful for a Question with a *causative* aspect investigated using a Plan *with* blocking or matching.

The Question usually involves the value of the difference $\mu_d = \mu_1 - \mu_2$.

In comparative investigating using a Plan with blocking or matching, there are *two* respondent populations corresponding to the units available for investigating with the *two* values of the focal variate; when using Model 3, the definitions of the symbols become too cumbersome unless we denote these two respondent populations as *one* population with i = 1 and i = 2, as in the definitions of Y_j, μ_d, R_j and σ_d above.

Model 4: $Y_j = \alpha + \beta_1(x_j - \overline{x}) + R_j$, $j = 1, 2, ..., n$; $R_j \sim G(0, \sigma)$; independent; EPS.

Y_j is a random variable whose distribution represents the possible values of the measured response variate for the jth unit in the sample of n units selected equiprobably from the respondent population, if the selecting and measuring processes were to be repeated over and over.

α is a model parameter which represents the *average* of y for the units of the respondent population whose value of x is $\overline{x}$, the *sample* average; it can be convenient to think of α as an 'intercept' – the ordinate of the point on the straight-line model for the relationship between x and the average of y when $x = \overline{x}$.

$\beta_0 = \alpha - \beta_1\overline{x}$ is a model parameter which represents the *y intercept* of the straight-line model for the relationship between x and the average of y in the respondent population; *i.e.*, the ordinate of this straight line when x = 0.

β_1 is a model parameter which represents the *slope* of the straight-line model for the relationship between x and the average of y in the respondent population;
i.e., the change in the measured average of y for unit change in x over the units of the respondent population.

x_j is the value of the explanatory variate x for the jth unit in the sample of n units selected equiprobably from the respondent population.

$\overline{x}$ is the average value of the explanatory variate x over the n units of the *sample*.

R_j is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the measured value of the response variate for the jth unit in the sample of n units selected equiprobably from the respondent population, if the selecting and measuring processes were to be repeated over and over.

σ the (probabilistic) *standard deviation* of the Gaussian model for the distribution of the residual, is a model parameter which represents the (data) *standard deviation* of the measured response variate of the units of the respondent population with value x_j for explanatory variate x; this (data) standard deviation (and, hence, σ) quantifies the *variation* of the measured response variate of the units of the respondent population with value x_j for explanatory variate x – as this variation increases, so does σ.

Model 4 is useful for a Question with a *causative* aspect investigated with a Plan in which values are available for an explanatory variate x that has a relationship to the average of y that can be modelled by a straight line.

The Question usually involves the value of the slope parameter β_1 and sometimes the intercept parameters α or β_0 of the model for the straight-line relationship between x and the average of y in the respondent population.

- Numbering the Models (1, 2, 3, 4) is only for convenience in this Highlight #72 and does *not* carry over to the Course Notes.
- When using the definitions from this Highlight #72 in a specific Question context, the generic 'response variate' should be replaced by the relevant description of the *actual* response variate for the Question context.