7-Types of data

ENSY SILVER[1]

Friday 4th September, 2020

[1]Thanks to my family, my friend and freedom.

1 Introduction

A working knowledge of statistics requires that the subject be pursued at two different levels. On one level, attention needs to be paid to the mathematical properties inherent in the individual measurements. These are what might be thought of as the **micro** structure of statistics. What is the pdf of the Y_i? Do we know $E(Y_i)$ or $\text{Var}(Y_i)$? Are the Y_i independent?

Viewed collectively, though, every set of measurements also has a certain overall structure, or *design*. It will be those **macro** features that we focus on in this chapter. A number of issues need to be addressed.

2 Basic notions

2.1 Factors and levels

The word **factor** is used to denote any treatment or therapy "applied to" the subjects being measured or any relevant feature (age, sex, ethnicity, etc.) "characteristic" of those subjects. Different versions, extents, or aspects of a factor are referred to as **levels**.

2.2 Independent and dependent observations

Measurements collected for the purpose of comparing two and more factor levels for measurements can either be **dependent** or **independent**. Two or more observations are dependent if they share a particular commonality relevant to what is being measured. If there is no such linkage, the observations are independent.

2.3 Similar and dissimilar units

Units also play a role in a data set' s macrostructure. Two measurements are said to be **similar** if their units are the same and **dissimilar** otherwise.

Quantitative measurements and qualitative measurements

Measurements are considered **quantitative** if their possible values are numerical. Qualitative **measurements** have "values"that are either categories, characteristics, or conditions.

3 Types of data

3.1 Model equations

In describing experimental designs, the assumptions given for a set of measurements are often written in the form of a model equation, which, by definition, expresses the value of an aribitrary Yi as the sum of fixed and variable components.

3.2 One-sample data

The simplest of all experimental designs, the **one-sample data** design, consists of a single random sample of size n. Necessarily, the n observations reflect one particular set of conditions or one specific factor.

For one-sample data, the usual model equation is

$$Y_i = \mu + \epsilon_i$$

where ϵ is a normally distributed random variable with mean 0 and standard deviation σ.

3.3 Two-sample data

Two-sample data consist of two independent random samples of sizes m and n, each having quantitative, similar unit measurements. Each sample is associated with a different factor level.

If $X_1, X_2, \cdots, X_n$ denotes the first sample and $Y_1, Y_2, \cdots, Y_m$ the second, the usual model equation assumptions would be written

$$X_i = \mu_X + \epsilon_i$$

and

$$Y_i = \mu_Y + \epsilon'_j$$

where ϵ_i is normally distributed with mean 0 and standard deviation σ_X^2, and ϵ'_j is normally distributed with mean 0 and standard deviation σ_Y^2,

3.4 k-sample data

When more than two factor levels are being compared, and when the observations are quantitative, have similar units, and are independent, the measurements are said to be k-**sample data**.

The ith observation appearing in the jth factor level will be denoted Y_{ij}, so the model equations take the form

$$Y_{ij} = \mu_j + \epsilon_{ij}, \quad i = 1, 2, \cdots, n_j,\ j = 1, 2, \cdots, k$$

where n_j denotes the sample size associated with the j^{th} factor level, and ϵ_{ij} is a normally distributed random variable with mean 0 and the samestandard deviation σ for all i and j.

3.5 Paired data

In the two-sample and k-sample designs, factor levels are compared using independent samples. An alternative is to use dependent samples by grouping the subjects into n blocks. If only two factor levels are involved, the blocks are referred to as pairs, which gives the design its name.

The responses to factor levels X and Y in the i^{th} pair are recorded as X_i and Y_i, respectively. Whatever contributions to those values are due to the conditions prevailing in Pair i will be denoted P_i. The model equations, then, can be written

$$X_i = \mu_X + P_i + \epsilon_i$$

and

$$Y_i = \mu_Y + P_i + \epsilon_i'$$

where ϵ_i and ϵ_i' are independent normally distributed random variables with mean 0 and the same standard deviation σ^2. The fact that P_i is the same for both X_i and Y_i is what makes the samples dependent.

3.6 Randomized block data

Randomized block data have the same basic structure as paired data—quantitative measurements, similar units, and dependent samples; the only difference is that more than two factor levels are involved in randomized block data.

Suppose the data set consists of k factor levels, all of which are applied in each of b blocks. The model equation for Y_{ij} , the observation appearing in the ith block and receiving the j^{th} factor level, then becomes

$$Y_{ij} = \mu_j + B_i + \epsilon_{ij}, \quad i = 1, 2, \cdots, b;\ j = 1, 2, \cdots, k$$

where μ_j is the true average response associated with the j^{th} factor level, B_i is the portion of the value Y_{ij} that can be attributed to the net effect of all the conditions that characterize Block i, and ϵ_{ij} is a normally distributed random variable with mean 0 and the same standard deviation σ^2 for all i and j.

3.7 Regression data

All the experimental designs introduced up to this point share the property that their measurements have the same units. Moreover, each has had the same basic objective: to quantify or to compare the effects of one or more factor levels. In contrast, regression data typically consist of measurements with dissimilar units, and the objective with them is to study the functional relationship between the variables.

Regression data often have the form (x_i, Y_i), $i = 1, 2, \cdots, n$, where x_i is the value of an independent variable (typically preselected by the experimenter) and Y_i is a dependent random variable (usually having units different from those of x_i). A particularly important special case is the *simple linear model*,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i is assumed to be normally distributed with mean 0 and standard deviation σ^2. The simple linear model can be extended to include k independent variables. The result is a *multiple linear regression model*,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

3.8 Categorical data

Suppose two qualitative, dissimilar variables are observed on each of n subjects, where the first variable has R possible values and the second variable, C possible values. We call such measurements categorical data.

The number of times each value of one variable occurs with each value of the other variable is typically displayed in a **contingency table**, which necessarily has R rows and C columns. Whether the two variables are independent is the question that an experimenter can use categorical data to answer.