13 Basic statistical formulations

referring to the textbook 'Statistics for life science' by Myra.L.Samuels

---

Frequency interpretation of probability (gene frequencies)

The **frequency interpretation** of probability provides a link between probability and the real world by relating the probability of an event to a measurable quantity, namely, the long-run relative frequency of occurrence of the event.*

According to the frequency interpretation, the probability of an event E is meaningful only in relation to a chance operation which can in principle be repeated indefinitely often. Each time the chance operation is repeated, the event E either occurs or does not occur. *The probability $Pr\{E\}$ is interpreted as the relative frequency of occurrence of E in an indefinitely long series of repetitions of the chance operation.*

Specifically, suppose that the chance operation is repeated a large number of times, and that for each repetition the occurrence or nonoccurrence of E is noted. Then we may write

$$\Pr\{E\} \leftrightarrow \frac{\#\text{ of times } E \text{ occurs}}{\#\text{ of times chance operation is repeated}}$$

Conditional probability

Recall that the probability of an event predicts how often the event will occur. A conditional probability predicts how often an event will occur *under specified conditions.* The notation for a conditional probability is

$$\Pr\{E|C\}$$

which is read "probability of E, *given* C." When a conditional probability is estimated from observed data, the estimate is denoted by a hat ("$\widehat{\ }$"), thus:

$$\widehat{\Pr}\{E|C\}$$

The following example illustrates these ideas.

Observed and expected by Chi-squate test

In the snapdragon (*Antirrhinum majus*), individual plants can be red-flowered, pink-flowered, or white-flowered. According to a certain Mendelian genetic model, self-pollination of pink-flowered plants should produce progeny that are red, pink, and white in the ratio 1 : 2 : 1. This Mendelian prediction can be formulated as the following null hypothesis:

$$H_0: \quad \Pr\{\text{Red}\} = .25, \quad \Pr\{\text{Pink}\} = .50, \quad \Pr\{\text{White}\} = .25$$

This hypothesis asserts that each progeny plant has a 25% chance of being red, a 50% chance of being pink, and a 25% chance of being white. Equivalently, H_0 asserts that, in the conceptual population of all potential progeny, 25% of the individuals are red, 50% are pink, and 25% are white. ■

Formula for Chi-squate test

THE CHI-SQUARE STATISTIC

$$\chi_s^2 = \sum \frac{(O - E)^2}{E}$$

where the summation is over all the categories.

Linear regression

EQUATION OF THE REGRESSION LINE

The equation of a straight line can be written as

$$Y = b_0 + b_1 X \qquad \text{eller} \qquad Y = \bar{y} + b(X - \bar{x})$$

where b_0 is the intercept and b_1 is the slope of the line. The slope b_1 is the rate of change of Y with respect to X.

The fitted regression line of Y on X is the line whose slope and intercept are calculated from the data as follows:

REGRESSION LINE OF *Y* ON *X*

$$\text{Slope:} \quad b_1 = \frac{SP_{XY}}{SS_X}$$

$$\text{Intercept:} \quad b_0 = \bar{y} - b_1\bar{x} \qquad \text{eller} \quad (\bar{x}, \bar{y})$$

Correlation and linear regression

Sometimes a researcher wishes to describe how closely a set of data points (x, y) cluster about a straight line. In other words, he wants to describe the *tightness* of the linear relationship between X and Y. The residual standard deviation provides one such description. In this section we consider another descriptive statistic, the **correlation coefficient**.

DEFINITION OF *r*

The sample correlation coefficient is denoted r and is defined as follows:*

DEFINITION OF CORRELATION COEFFICIENT

$$r = \frac{SP_{XY}}{\sqrt{SS_X SS_Y}}$$

When both X and Y have the same variance then regression and correlation coefficients become equal as shown below

$$r = \frac{SP_{XY}}{\sqrt{SS_X SS_Y}} = b_1 = \frac{SP_{XY}}{SS_X}$$

This is true for instance for father - son or mother - daughter relationship
When the Sum of Products (SP) and the Sum of Squares (SS) are diveded with the degrees of freedom you get the Covariance and Variance.
They can be used in the formulas instead of SP and SS's