Statistical Interval Guide

I. Confidence Intervals

A. Confidence Interval for a Population Mean

Assumptions

1. There are n observations from a $N(\mu, \sigma^2)$ population (normality is not crucial if sample size is large enough: 30 or more).
2. σ^2 is unknown.

Formula A level L confidence interval for μ is $\overline{y} \pm t_{n-1,(1+L)/2} \frac{s}{\sqrt{n}}$, where $\overline{y}$ is the sample mean and s is the sample standard deviation.

B. An Approximate Score Confidence Interval for a Population Proportion

Assumptions

1. We are interested in estimating the proportion p having a certain characteristic in the target population.
2. y is the number having the characteristic in a random sample of size n taken from the population.

Formula

1. First, modify y and n as follows: $\tilde{y} = y + 0.5z^2_{(1+L)/2}$, $\tilde{n} = n + z^2_{(1+L)/2}$.
2. Next, create $\tilde{p} = \tilde{y}/\tilde{n}$.
3. Finally, an approximate score level L confidence interval for p is given by

$$\tilde{p} \pm z_{(1+L)/2}\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}.$$

C. Confidence Intervals for the Difference of Two Means

In all cases we assume the data are:

$$y_{1,1}, y_{1,2}, \ldots, y_{1,n_1} \sim N(\mu_1, \sigma_1^2), \text{ (population 1)}$$

$$y_{2,1}, y_{2,2}, \ldots, y_{2,n_2} \sim N(\mu_2, \sigma_2^2), \text{ (population 2).}$$

We compute level L confidence intervals for $\mu_1 - \mu_2$.

Case 1: Paired Data In this case, we take differences $d_i = y_{1,i} - y_{2,i}$. Then a confidence interval for the mean difference is also a confidence interval for $\mu_1 - \mu_2$.

Case 2: Independent Populations, Variances Assumed Equal

Assumption $\sigma_1^2 = \sigma_2^2$, and their value is unknown.

Formulas Estimate the population variance with the pooled variance estimator

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

where s_1^2 and s_2^2 are the sample standard deviations. The confidence interval is

$$\overline{y}_1 - \overline{y}_2 \pm t_{n_1+n_2-2,(1+L)/2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

Case 3: Independent Populations, Variances Not Assumed Equal The confidence interval is

$$\overline{y}_1 - \overline{y}_2 \pm t_{\nu,(1+L)/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where ν is taken as the largest integer less than or equal to $\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2 \Big/ \left[\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1} \right]$.

D. An Approximate Score Confidence Interval for the Difference of Two Proportions

Assumptions

(a) We are interested in estimating $p_1 - p_2$, where p_1 is the proportion having a certain characteristic in population 1 and p_2 is the proportion having a certain characteristic in population 2.

(b) y_1 is the number having the characteristic in a random sample of size n_1 taken from population 1 and y_2 is the number having the characteristic in a random sample of size n_2 taken from population 2.

Formula

(a) First, modify y_1, y_2, n_1 and n_2 as follows: $\tilde{y}_1 = y_1 + 0.25z^2_{(1+L)/2}$, $\tilde{n}_1 = n_1 + 0.5z^2_{(1+L)/2}$, $\tilde{y}_2 = y_2 + 0.25z^2_{(1+L)/2}$, $\tilde{n}_2 = n_2 + 0.5z^2_{(1+L)/2}$.

(b) Next, create $\tilde{p}_1 = \tilde{y}_1/\tilde{n}_1$ and $\tilde{p}_2 = \tilde{y}_2/\tilde{n}_2$.

(c) Finally, an approximate score level L confidence interval for $p_1 - p_2$ is given by

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{(1+L)/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{\tilde{n}_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{\tilde{n}_2}},$$

II. Prediction Interval

Assumptions

(a) The data are $y_1, y_2, \ldots, y_n$ where $y_j = \mu + \epsilon_j$.

(b) The ϵ_j are from a $N(0, \sigma^2)$ population.

(c) σ^2 is unknown.

Formulas

A level L prediction interval for a future observation is
$\left(\hat{y}_{new} - \hat{\sigma}(y_{new} - \hat{y}_{new})t_{n-1,\frac{1+L}{2}},\ \hat{y}_{new} + \hat{\sigma}(y_{new} - \hat{y}_{new})t_{n-1,\frac{1+L}{2}}\right)$, where
$\hat{\sigma}(y_{new} - \hat{y}_{new}) = S\sqrt{1 + 1/n}$.

III. Tolerance Interval

Assumptions

(a) The data are $y_1, y_2, \ldots, y_n$ where $y_j = \mu + \epsilon_j$.

(b) The ϵ_j are from a $N(0, \sigma^2)$ population.

(c) σ^2 is unknown.

Formulas A level L normal theory tolerance interval for a proportion γ of the population of measurements is

$$(\overline{y} - Ks,\ \overline{y} + Ks),$$

where $\overline{y}$ and s are the sample mean and standard deviation, and K is a mathematically-derived constant depending on L, γ and the number of observations, n. Values are obtained from a table or directly via computer algorithm.