

Percentile Estimation

- Percentile Point Estimation:

Recall that the $(100p)^{\text{th}}$ percentile of the distribution of the random variable X is given by the smallest π_p such that

$$\Pr\{X \leq \pi_p\} = p.$$

Let Y_i denote the i^{th} order statistic of the vector of i.i.d. observations $\mathbf{X} = X_1, X_2, \dots, X_n$.

To estimate π_p , use

$$\hat{\pi}_p = \begin{cases} Y_{(n+1)p} & \text{if } (n+1)p \in \{1, 2, \dots, n\} \\ w_L Y_L + w_U Y_U & \text{otherwise} \end{cases},$$

where Y_L and Y_U are consecutive order statistics such that

$$L < (n+1)p < U, \text{ and} \\ w_L = 1 - [(n+1)p - L] = 1 - w_U.$$

Note that the choice of the weights w_L and w_U , while justified by intuition, is not the only selection that could be made.

- Non-Parametric Confidence Intervals

Let the X_i be continuous random variables, and, for each observation X_i , let the event $X_i \leq \pi_p$ be called a “success”.

Now consider the probability

$$\begin{aligned} & \Pr\{Y_i \leq \pi_p < Y_j\} \\ &= \Pr\{\text{at least } i \text{ successes, and at most } j-1 \text{ successes}\} \\ &= \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned}$$

To construct a confidence interval for π_p with confidence level of at least $1 - \alpha$, it suffices to find a pair of order statistics Y_i, Y_j such that

$$\Pr\{Y_i \leq \pi_p < Y_j\} \geq 1 - \alpha.$$

Note that the order statistics Y_i, Y_j will generally not be unique.

Fitting Distributions

- For a vector of i.i.d. observations, $\mathbf{X} = X_1, X_2, \dots, X_n$, the empirical distribution function is given by

$$F_n(x) = \frac{\text{number of } X_i \leq x}{n}, \text{ for } x \in \mathbb{R}.$$

Note that $F_n(x)$ is a non-decreasing step function with a jump of size $\frac{1}{n}$ at each observation X_i .

(If the X_i are discrete random variables, then there may be a jump of $\frac{2}{n}, \frac{3}{n}$, etc. if there are ties among the observations.)

- For any partition of the real line, $c_0 < c_1 < c_2 < \dots < c_k$, where $c_0 < \text{Min}\{X_i\}$ and $c_k > \text{Max}\{X_i\}$, the piecewise linear function $H(x)$ that joins the points

$$(c_0, 0), (c_1, F_n(c_1)), (c_2, F_n(c_2)), \dots, (c_k, 1)$$

is called an ogive.

- Pearson's χ^2 Test Statistic

For a vector of i.i.d. observations, $\mathbf{X} = X_1, X_2, \dots, X_n$, consider the null hypothesis,

$$H_0: F(x) = F_0(x).$$

To test this hypothesis, first partition the sample space of X_i into k mutually exclusive and collectively exhaustive subsets, indexed by j .

Pearson's χ^2 statistic is then given by

$$\chi_p^2 = \sum_{j=1}^k \frac{(f_j - np_j)^2}{np_j},$$

where f_j denotes the observed frequency with which the X_i fall into subset j , and p_j denotes the corresponding theoretical frequency under the null hypothesis.

Under H_0 , $\chi_p^2 \sim$ approximately χ_{k-1}^2 . Therefore, for a given level of significance α , the critical region is simply $(c, +\infty)$, where $c = \chi_{k-1, \alpha}^2$.

- A related test statistic is the Minimum χ^2 statistic, X_M^2 , which is calculated by minimizing Pearson's χ^2 over all $F_0 \in F$, for some family of distribution functions F .

This test statistic may be used to test the null hypothesis,

$$H_0: F(x) = F_0(x).$$

Under H_0 , $X_M^2 \sim \chi_{k-1-p}^2$ approximately, where p is the number of parameters needed to characterize the members of F .

- The Kolmogorov-Smirnov Test Statistic

For a vector of i.i.d. observations, $\mathbf{X} = X_1, X_2, \dots, X_n$, consider again the null hypothesis,

$$H_0: F(x) = F_0(x).$$

The Kolmogorov-Smirnov test statistic is given by

$$D_n = \max_x |F_n(x) - F_0(x)|.$$

If X is continuous, then it is necessary to check the values of $|F_n(x) - F_0(x)|$ at and "just before" each X_i ; i.e., it is necessary to check both $|F_n(X_i) - F_0(X_i)|$ and $|F_n(X_i -) - F_0(X_i -)|$ for all i .

If X is discrete, then it is necessary to check $|F_n(x) - F_0(x)|$ for all $x \in X$.

Under H_0 , the distribution of D_n is not well-known, but it has been tabulated for various values of n . For a given level of significance α , the critical region is $(c, +\infty)$, where $\Pr\{D_n > c | H_0\} = \alpha$.

- Note that $\Pr\{D_n > c | H_0\} = \alpha \implies \Pr\left\{\max_x |F_n(x) - F_0(x)| \leq c | H_0\right\} = 1 - \alpha$
 $\Pr\{F_n(x) - c \leq F_0(x) \leq F_n(x) + c\} = 1 - \alpha$ for any H_0 .

Thus, $(\max\{F_n(x) - c, 0\}, \min\{F_n(x) + c, 1\})$ constitutes a confidence interval for $F_0(x)$ with confidence level of at least $1 - \alpha$.

- A related test statistic is the Cramér-von Mises statistic,

$$T = \frac{1}{n} \sum_{i=1}^n [F_n(X_i) - F_0(X_i)]^2 .$$

This test statistic may be used to estimate the distribution function $F(x)$ by finding the $F_0(x)$ that minimizes T .

This type of approach is called minimum-distance estimation.