eCDF: expectation

Suppose that we observe a sample $x_1, ..., x_n$ which are realizations of i.i.d. r.v. $X_1, ..., X_n$, with continuous distribution L_x. The empirical Cummulative Distribution Function (eCDF) is defined as

$$F_{x,n}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq t}$$

Taking the expectation, we get

$$\begin{aligned} E[F_{x,n}(t)] &= E\left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq t}\right] = \frac{1}{n} \sum_{i=1}^{n} E\left[\mathbb{1}_{x_i \leq t}\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} P(x_i \leq t) \\ &= \frac{1}{n} n F_x(t) = F_x(t) \end{aligned}$$

where $F_x(t)$ is the true CDF of the data.

eCDF: variance

We can also prove, since $\mathbb{1}_{x_i \leq t} \sim B(F_x(t))$, that

$$\begin{aligned} var(F_{x,n}(t)) &= var\left(\frac{1}{n}\sum_{i=1}^{n} E\mathbb{1}_{x_i \leq t}\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} var\left(\mathbb{1}_{x_i \leq t}\right) \\ &= \frac{1}{n^2} n\left((F_x(t))(1 - F_x(t))\right) \\ &= \frac{1}{n}\left((F_x(t))(1 - F_x(t))\right) \end{aligned}$$

Besides, from the Law of Large Numbers (LLN), we have that

$$F_{x,n}(t) \xrightarrow{a.s.} F_x(t) \qquad \text{as } n \text{ goes to } \infty.$$

eCDF: distribution and pointwise CI

And from the Central Limit Theorem (CLT), we have that

$$F_{x,n}(t) \xrightarrow{L} N\left(F_x(t), \frac{1}{n}F_x(t)\big(1 - F_x(t)\big)\right) \qquad \text{as } n \text{ goes to } \infty.$$

As a consequence, a pointwise $(1-\alpha)$ Confidence Interval for $F_x(t)$ is given by

$$\Bigg[F_{x,n}(t) - q_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}F_{x,n}(t)\big(1 - F_{x,n}(t)\big)},$$
$$F_{x,n}(t) + q_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}F_{x,n}(t)\big(1 - F_{x,n}(t)\big)}\Bigg]$$

KS test: test statistic and its distribution

One-sample test: We observe a sample $x_1, ..., x_n$ which are realizations of i.i.d. r.v. $X_1, ..., X_n$, and want to test the null hypothesis $H_0 : F_x = F$, i.e H_0 : 'the true CDF of the data is equal to a given specific CDF (for example the normal CDF)', versus $H_1 : F_x \neq F$. The test statistic is

$$D_{x,n} = \sup_{t \in \mathbb{R}} \Big| F_{x,n}(t) - F(t) \Big|$$

It can be shown that, under H_0, when n is large, $\sqrt{n} D_{x,n}$ has the distribution of a r.v. K having CDF

$$F_k = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} exp(-2k^2 t^2)$$

Decision rule

The null hypothesis H_0 is rejected at level α if $\sqrt{n}D_{x,n} \geq d_\alpha$, where d_α is the $1-\alpha$ quantile of the distribution of K. For $\alpha = 0.05$, $d_\alpha \approx 1.36$. As a consequence, a confidence band can be build for the unknown true CDF $F_x(t)$ and is given by

$$\left[F_{x,n}(t) - \frac{d_\alpha}{\sqrt{n}},\ F_{x,n}(t) + \frac{d_\alpha}{\sqrt{n}}\right]$$

Remark: The pointwise $(1-\alpha)$ Confidence Interval for $F_x(t)$ is usually tighter for $\alpha = 0.05$ or lower.

Two-sample test

Two-sample test: We observe a sample $x_1, ..., x_n$ which are realizations of i.i.d. r.v. $X_1, ..., X_n$ and $y_1, ..., y_n$ which are realizations of i.i.d. r.v. $Y_1, ..., Y_n$, with respective continuous distributions L_x and L_y. We want to test $H_0 : F_x = G_y$ versus $H_1 : F_x \neq G_y$. Let $F_{x,n}$ be the eCDF of the sample $x_1, ..., x_n$ and $G_{y,m}$ be the eCDF of the sample $y_1, ..., y_m$. The test relies on the statistic

$$D_{n,m} = \sup_{t \in \mathbb{R}} \left| F_{x,n}(t) - G_{y,m}(t) \right|$$

Under H_0, if n and m are large, the distribution of $\frac{D_{n,m}}{\sqrt{\frac{1}{n}+\frac{1}{m}}}$ is F_k.

So H_0 is rejected at level α if $\frac{D_{n,m}}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \geq d_\alpha$.

References

Bagdonavičius V., Kruopis J., Nikulin M. S., Non-parametric Tests for Complete Data (2011), Wiley, ISBN 978-1-84821-269-5 (hardback)

The R Project for Statistical Computing: https://www.r-project.org/

course notes