Statistics for Applications

Chapter 5: Testing goodness of fit

Cdf and empirical cdf (1)

Let $X_1, \ldots, X_n$ be i.i.d. real random variables. Recall the cdf of X_1 is defined as:

$$F(t) = \mathbb{P}[X_1 \leq t], \quad \forall t \in \mathbb{R}.$$

It completely characterizes the distribution of X_1.

Definition

The *empirical cdf* of the sample $X_1, \ldots, X_n$ is defined as:

$$\begin{aligned} F_n(t) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t} \\ &= \frac{\#\{i = 1, \ldots, n : X_i \leq t\}}{n}, \quad \forall t \in \mathbb{R}. \end{aligned}$$

Cdf and empirical cdf (2)

- Let Z be a r.v. chosen at random among $X_1, \ldots, X_n$.
- Conditionally on the sample $X_1, \ldots, X_n$, if the X_i's are pairwise distinct,

$$Z \sim \mathcal{U}\left(\{X_1, \ldots, X_n\}\right).$$

- F_n is the conditional cdf of Z given the sample.
- F is the (unconditional) cdf of Z.

Cdf and empirical cdf (3)

By the LLN, for all $t \in \mathbb{R}$,

$$F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t).$$

Glivenko-Cantelli Theorem (*Fundamental theorem of statistics*)

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0.$$

Hence, $F_n(t)$ brings a lot of information about the distribution of the X_i's when n is large.

Cdf and empirical cdf (4)

By the CLT, for all $t \in \mathbb{R}$,

$$\sqrt{n}\,\left(F_n(t) - F(t)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\big(0, F(t)\left(1 - F(t)\right)\big).$$

Donsker's Theorem

If F is continuous, then

$$\sqrt{n} \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{(d)} \sup_{0\le t\le 1} |\mathbb{B}(t)|,$$

where $\mathbb{B}$ is a Brownian bridge on $[0, 1]$.

Kolmogorov-Smirnov test (1)

- Let $X_1, \ldots, X_n$ be i.i.d. real random variables with unknown cdf F and let F^0 be a **continuous** cdf.
- Consider the two hypotheses:

$$H_0 : "F = F^0" \quad \text{v.s.} \quad H_1 : "F \neq F^0".$$

- Let F_n be the empirical cdf of the sample $X_1, \ldots, X_n$.
- If $F = F^0$, then $F_n(t) \approx F^0(t)$, uniformly in $t \in [0, 1]$.

Kolmogorov-Smirnov test (2)

- Let $T_n = \sup_{t \in \mathbb{R}} \sqrt{n} \left| F_n(t) - F^0(t) \right|$.

- By Donsker's theorem, if H_0 is true, then $T_n \xrightarrow[n \to \infty]{(d)} Z$, where Z has a known distribution (supremum of a Brownian bridge).

- **KS test with asymptotic level α:**

$$\delta_\alpha^{KS} = \mathbb{1}_{T_n > q_{1-\alpha}},$$

  where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of Z (obtained in tables).

- p-value of KS test: $\mathbb{P}[Z > T_n | T_n]$.

Kolmogorov-Smirnov test (3)

Remarks:

- In practice, how to compute T_n ?
- F^0 is non decreasing, F_n is piecewise constant, with jumps at $t_i = X_i, i = 1, \ldots, n$.
- Let $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ be the reordered sample.
- The expression for T_n reduces to the following practical formula:

$$T_n = \sqrt{n} \max_{i=1,\ldots,n} \left\{ \max \left(\left| \frac{i-1}{n} - F^0(X_{(i)}) \right|, \left| \frac{i}{n} - F^0(X_{(i)}) \right| \right) \right\}.$$

Kolmogorov-Smirnov test (4)

- T_n is called a *pivotal statistic*: If H_0 is true, the distribution of T_n does not depend on the distribution of the X_i's and it is easy to reproduce it in simulations.

- Indeed, let $U_i = F^0(X_i), i = 1, \ldots, n$ and let G_n be the empirical cdf of $U_1, \ldots, U_n$.

- If H_0 is true, then $U_1, \ldots, U_n \overset{i.i.d.}{\sim} \mathcal{U}([0.1])$ and $T_n = \sup_{0 \leq x \leq 1} \sqrt{n}\,|G_n(x) - x|$.

Kolmogorov-Smirnov test (5)

- For some large integer M:
  - Simulate M i.i.d. copies $T_n^1, \ldots, T_n^M$ of T_n;
  - Estimate the $(1-\alpha)$-quantile $q_{1-\alpha}^{(n)}$ of T_n by taking the sample $(1-\alpha)$-quantile $\hat{q}_{1-\alpha}^{(n,M)}$ of $T_n^1, \ldots, T_n^M$.
- Test with approximate level α:

$$\delta_\alpha = \mathbb{1}_{T_n > \hat{q}_{1-\alpha}^{(n,M)}}.$$

- Approximate p-value of this test:

$$\text{p-value} \approx \frac{\#\{j = 1, \ldots, M : T_n^j > T_n\}}{M}.$$

- See Optional Exercises 3, Ex. 4.

Test of independence (1)

- Consider a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of i.i.d. vectors on a finite space $\{a_1, \ldots, a_K\} \times \{b_1, \ldots, b_L\}$.

- Consider the two hypotheses:

  $H_0 : "X_1 \perp\!\!\!\perp Y_1"$ v.s. $H_1 : "X_1$ and Y_1 are not independent".

- For $(k, l) \in \{1, \ldots, K\} \times \{1, \ldots, L\}$, set:
  - $p_{k,l} = \mathbb{P}[X_1 = a_k, Y_1 = b_l]$,
  - $p_{k,\cdot} = \mathbb{P}[X_1 = a_k]$,
  - $p_{\cdot,l} = \mathbb{P}[Y_1 = b_l]$.

- H_0 reduces to:

$$p_{k,l} = p_{k,\cdot} \times p_{\cdot,l}, \quad \forall k, l.$$

Test of independence (2)

- Set $\hat{p}_{k,l} = \dfrac{N_{k,l}}{n}$, $\hat{p}_{k,\cdot} = \dfrac{N_{k,\cdot}}{n}$ and $\hat{p}_{\cdot,l} = \dfrac{N_{\cdot,l}}{n}$, where
  - $N_{k,l} = \#\{i = 1 \ldots, n : X_i = a_k, Y_i = b_l\}$,
  - $N_{k,\cdot} = \#\{i : X_i = a_k\}$,
  - $N_{\cdot,l} = \#\{i : Y_i = b_l\}$.
- Remark: $\hat{p}_{k,l}$ is the MLE of $p_{k,l}$.
- If H_0 is true, then $\hat{p}_{k,l} \approx \hat{p}_{k,\cdot}\hat{p}_{\cdot,l}$, $\quad \forall k, l$.
- Let $T_n = n \sum_{k=1}^{K} \sum_{l=1}^{L} \dfrac{(\hat{p}_{k,l} - \hat{p}_{k,\cdot}\hat{p}_{\cdot,l})^2}{\hat{p}_{k,\cdot}\hat{p}_{\cdot,l}}$.

Test of independence (3)

- If H_0 is true, then

$$T_n \xrightarrow[n\to\infty]{(d)} \chi^2_{(K-1)(L-1)}.$$

- Test of independence with asymptotic level α:

$$\delta_\alpha = \mathbb{1}_{T_n > q_{1-\alpha}},$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\chi^2_{(K-1)(L-1)}$.

- p-value: $\mathbb{P}[Z > T_n | T_n]$, where $Z \sim \chi^2_{(K-1)(L-1)}$ and $Z \perp\!\!\!\perp T_n$.

χ^2 goodness-of-fit test, finite case (1)

- Let $X_1, \ldots, X_n$ be i.i.d. random variables on some finite space $E = \{a_1, \ldots, a_K\}$, with some probability measure $\mathbb{P}$.
- Let $(\mathbb{P}_\theta)_{\theta \in \Theta}$ be a parametric family of probability distributions on E.
- Example: On $E = \{1, \ldots, K\}$, consider the family of binomial distributions $(\mathcal{B}(K, p))_{p \in (0,1)}$.
- For $j = 1, \ldots, K$ and $\theta \in \Theta$, set

$$p_j(\theta) = \mathbb{P}_\theta[Y = a_j], \quad \text{where } Y \sim \mathbb{P}_\theta$$

and

$$p_j = \mathbb{P}[X_1 = a_j].$$

χ^2 goodness-of-fit test, finite case (2)

- Consider the two hypotheses:

$$H_0 : "\mathbb{P} \in (\mathbb{P}_\theta)_{\theta \in \Theta}" \quad \text{v.s.} \quad H_1 : "\mathbb{P} \notin (\mathbb{P}_\theta)_{\theta \in \Theta}".$$

- Testing H_0 means testing whether the statistical model $\left(E, \mathcal{P}(E), (\mathbb{P}_\theta)_{\theta \in \Theta}\right)$ fits the data (e.g., whether the data are indeed from a binomial distribution).

- H_0 is equivalent to:

$$"p_j = p_j(\theta), \quad \forall j = 1, \ldots, K, \text{ for some } \theta \in \Theta."$$

χ^2 goodness-of-fit test, finite case (3)

- Let $\hat{\theta}$ be the MLE of θ when assuming H_0 is true.
- Let $\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i = a_j} = \frac{\#\{i : X_i = a_j\}}{n}, \quad j = 1, \ldots, K.$
- **Idea:** If H_0 is true, then $p_j = p_j(\theta)$ so both $\hat{p}_j$ and $p_j(\hat{\theta})$ are *good* estimators or p_j. Hence, $\hat{p}_j \approx p_j(\hat{\theta}), \quad \forall j = 1, \ldots, K.$
- Define the test statistic: $$T_n = n \sum_{j=1}^{K} \frac{\left(\hat{p}_j - p_j(\hat{\theta})\right)^2}{p_j(\hat{\theta})}.$$

χ^2 goodness-of-fit test, finite case (4)

- Under some technical assumptions, if H_0 is true, then

$$T_n \xrightarrow[n\to\infty]{(d)} \chi^2_{K-d-1},$$

where d is the size of the parameter θ ($\Theta \subseteq \mathbb{R}^d$ and $d < K-1$).

- Test with asymptotic level $\alpha \in (0,1)$:

$$\delta_\alpha = \mathbb{1}_{T_n > q_{1-\alpha}},$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of χ^2_{K-d-1}.

- p-value: $\mathbb{P}[Z > T_n | T_n]$, where $Z \sim \chi^2_{K-d-1}$ and $Z \perp\!\!\!\perp T_n$.

χ^2 goodness-of-fit test, infinite case (1)

- If E is infinite (e.g. $E = \mathbb{N}, E = \mathbb{R}$, ...):
- Partition E into K disjoint bins:

$$E = A_1 \cup \ldots \cup A_K.$$

- Define, for $\theta \in \Theta$ and $j = 1, \ldots, K$:
  - $p_j(\theta) = \mathbb{P}_\theta[Y \in A_j]$, for $Y \sim \mathbb{P}_\theta$,
  - $p_j = \mathbb{P}[X_1 \in A_j]$,
  - $\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \in A_j} = \frac{\#\{i : X_i \in A_j\}}{n}$,
  - $\hat{\theta}$: same as in the previous case.

χ^2 goodness-of-fit test, infinite case (2)

- As previously, let $T_n = n \sum_{j=1}^{K} \frac{\left(\hat{p}_j - p_j(\hat{\theta})\right)^2}{p_j(\hat{\theta})}$.

- Under some technical assumptions, if H_0 is true, then

$$T_n \xrightarrow[n\to\infty]{(d)} \chi^2_{K-d-1},$$

  where d is the size of the parameter θ ($\Theta \subseteq \mathbb{R}^d$ and $d < K-1$).

- Test with asymptotic level $\alpha \in (0,1)$:

$$\delta_\alpha = \mathbb{1}_{T_n > q_{1-\alpha}},$$

  where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of χ^2_{K-d-1}.

χ^2 goodness-of-fit test, infinite case (3)

- Practical issues:
  - Choice of K ?
  - Choice of the bins $A_1, \ldots, A_K$?
  - Computation of $p_j(\theta)$?
- Example 1: Let $E = \mathbb{N}$ and H_0 : "$\mathbb{P} \in (\mathcal{P}oiss(\lambda))_{\lambda>0}$".
- If one expects λ to be no larger than some λ_{max}, one can choose $A_1 = \{0\}, A_2 = \{1\}, \ldots, A_{K-1} = \{K-2\}, A_K = \{K-1, K, K+1, \ldots\}$, with K large enough such that $p_K(\lambda_{max}) \approx 0$.