Statistics for Applications

Chapter 6: Testing goodness of fit

Goodness of fit tests

Let X be a r.v. Given i.i.d copies of X we want to answer the following types of questions:

- Does X have distribution $\mathcal{N}(0,1)$? (Cf. Student's T distribution)
- Does X have distribution $\mathcal{U}([0,1])$? (Cf p-value under H_0)
- Does X have PMF $p_1 = 0.3,\ p_2 = 0.5,\ p_3 = 0.2$

These are all *goodness of fit* tests: we want to know if the hypothesized distribution is a good fit for the data.

Key characteristic of GoF tests: no parametric modeling.

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:

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

It completely characterizes the distribution of X_1.

Definition

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

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

Cdf and empirical cdf (2)

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*)

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

Cdf and empirical cdf (3)

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

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

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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$.

- If $F = F^0$, then $F_n(t) \approx F^0(t)$, for all $t \in [0, 1]$.

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:**

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

where q_α 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 $t_i = X_i, i = 1, \dots, n$.
- Let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ be the reordered sample.
- The expression for T_n reduces to the following practical formula:

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

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 $T_n = \sup_{0 \le x \le 1} \sqrt{n} \, |G_n(x) - x|$.

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 $q_\alpha^{(n)}$ of T_n by taking the sample $(1-\alpha)$-quantile $\hat{q}_\alpha^{(n,M)}$ of $T_n^1, \ldots, T_n^M$.
- Test with approximate level α:

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

- Approximate p-value of this test:

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

Kolmogorov-Smirnov test (6)

These quantiles are often precomputed in a table.

Other goodness of fit tests

We want to measure the distance between two functions: $F_n(t)$ and $F(t)$. There are other ways, leading to other tests:

- Kolmogorov-Smirnov:

$$d(F_n, F) = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$$

- Cramér-Von Mises:

$$d^2(F_n, F) = \int_{\mathbb{R}} [F_n(t) - F(t)]^2 \, dt$$

- Anderson-Darling:

$$d^2(F_n, F) = \int_{\mathbb{R}} \frac{[F_n(t) - F(t)]^2}{F(t)(1 - F(t))} dt$$

Composite goodness of fit tests

What if I want to test: "Does X have Gaussian distribution?" but I don't know the parameters?
Simple idea: plug-in

$$\sup_{t \in \mathbb{R}} \quad F_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t)$$

where

$$\hat{\mu} = \bar{X}_n, \qquad \hat{\sigma}^2 = S_n^2$$

and $\Phi_{\hat{\mu},\hat{\sigma}^2}(t)$ is the cdf of $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$.

In this case Donsker's theorem is *no longer valid*. This is a common and serious mistake!

Kolmogorov-Lilliefors test (1)

Instead, we compute the quantiles for the test statistic:

$$\sup_{t \in \mathbb{R}} \quad F_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t)$$

They do not depend on unknown parameters!

This is the Kolmogorov-Lilliefors test.

Kolmogorov-Lilliefors test (2)

These quantiles are often precomputed in a table.

Quantile-Quantile (QQ) plots (1)

- Provide a visual way to perform GoF tests
- Not formal test but quick and easy check to see if a distribution is plausible.
- Main idea: we want to check visually if the plot of F_n is close to that of F or equivalently if the plot of F_n^{-1} is close to that of F^{-1}.
- More convenient to check if the points

$$\big(F^{-1}(\frac{1}{n}), F_n^{-1}(\frac{1}{n})\big), \big(F^{-1}(\frac{2}{n}), F_n^{-1}(\frac{2}{n})\big), \ldots, \big(F^{-1}(\frac{n-1}{n}), F_n^{-1}(\frac{n-1}{n})\big)$$

are near the line $y = x$.

- F_n is not technically invertible but we define

$$F_n^{-1}(i/n) = X_{(i)},$$

the ith largest observation.

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 $(\mathsf{Bin}(K, p))_{p \in (0,1)}$.

- 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 $\big(E, (\mathbb{P}_\theta)_{\theta \in \Theta}\big)$ 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 $\alpha \in (0,1)$:

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

where q_α 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 $\alpha \in (0, 1)$:

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

where q_α 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 $H_0 : \mathbb{P} \in (\text{Poiss}(\lambda))_{\lambda>0}$.
- If one expects λ to be no larger than some $\lambda_{\max}$, 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 $p_K(\lambda_{\max}) \approx 0$.