Parametric & Nonparametric Statistics Project

Aleksandr Jan Smoliakov

2024–12–12

1 Introduction

2 Preliminaries

In the project below, we will use the following parameters:

- $\mathcal{N} = 9$ (first name: 'Aleksandr', 9 letters)
- $\mathcal{S} = 9$ (last name: 'Smoliakov', 9 letters)
- $\mathcal{I}_1 = 5$ (last digit of study book number)
- $\mathcal{I}_2 = 8$ (second last digit of study book number)

Let $G_1, \dots, G_m$ be given distribution functions and $p_1, \dots, p_m$ be probabilities that sum to 1. The distribution function G defined by

$$G(u) := p_1 G_1(u) + \cdots + p_m G_m(u) = \sum_{k=1}^{m} p_k G_k(u), \quad u \in \mathbb{R}$$

is called a mixture of distribution functions $G_1, \dots, G_m$ with probabilities (or weights) $p_1, \dots, p_m$.

G is the distribution function of the random variable Z generated in the following way:

1. Choose $k \in \{1, \dots, m\}$ at random with probabilities (or weights) $p_1, \dots, p_m$. The chosen number is denoted by k^*.

2. Generate a random variable $Z^*_{k^*}$ according to the distribution function G_{k^*} and assign $Z \leftarrow Z^*_{k^*}$.

In this project, we will have $m = 2$, so the algorithm for generating Z is as follows:

$$Z \leftarrow Z^*_{1+k^*} \quad k^* \sim \text{Binomial}(1, p_2), \quad Z^*_k \sim G_k \ (k = 1, 2).$$

Let

$$\mathcal{G}(\Theta) = \{G(\cdot|\theta) : \theta \in \Theta\}$$

be a given parametric family of absolutely continuous parametric functions $G(\cdot|\theta)$ with the respective distribution densities $g(\cdot|\theta)$ dependent on the unknown parameter $\theta \in \Theta$. It is assumed that θ is two-dimensional, i.e., $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$.

2.1 Parametric Family Selection

Using the assigned formula $l := \left\lfloor \frac{I_2+2.5}{2} \right\rfloor$, we find $l = 5$. Thus, we will use the parametric family $\mathcal{G}_5(\Theta)$ in this project.

$\mathcal{G}_5(\Theta)$ contains distribution functions of random variables uniformly distributed on $[\theta_1, \theta_2]$, where $\theta_1 < \theta_2$.

The family $G_5(\Theta)$ consists of uniform distributions:

$$G(u|\theta) = \begin{cases} 0 & u < \theta_1 \\ \frac{u-\theta_1}{\theta_2-\theta_1} & \theta_1 \le u \le \theta_2 \\ 1 & u > \theta_2 \end{cases}$$

with $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ and $\theta_1 < \theta_2$.

3 Task 1: Testing Goodness-of-Fit

3.1 Basic Distribution Function

The problem gives a specific basic parameter:

$$\theta_0 = (-\mathcal{N}, \mathcal{S} + 4) = (-9, 13).$$

with $\mathcal{N} = 9$ and $\mathcal{S} = 9$. Thus:

$$\theta_0 = (-9, 13).$$

Thus, the basic distribution function is:

$$G_0(u) = G(u|\theta_0) = U(-9, 13).$$

For a uniform distribution $U(a, b)$:

- Mean: $\mu = \frac{a+b}{2}$
- Variance: $v^2 = \frac{(b-a)^2}{12}$

For $G_0 = U(-9, 13)$:

$$\mu_0 = \frac{-9+13}{2} = \frac{4}{2} = 2$$

$$v_0^2 = \frac{22^2}{12} = \frac{484}{12} = \frac{121}{3} \approx 40.3333$$

So:

$$\mu_0 = 2, \quad v_0^2 \approx 40.3333.$$

3.2 Finding G_1 and G_2

We are given the following equations for the mixture distributions G_1 and G_2:

We have the equations:

$$\mu_0 = \mu(\theta_1), \quad \mathcal{N} v_0^2 = v^2(\theta_1).$$

$$\mu_0 + 2v_0 = \mu(\theta_2), \quad v_0^2 = \mathcal{S} v^2(\theta_2).$$

First we determine G_1, we have:

$$\mu_0 = \mu(\theta_1), \quad \mathcal{N} v_0^2 = v^2(\theta_1).$$

Since $\mathcal{N} = 9$ and $v_0^2 = \frac{121}{3}$, we have:

$$v^2(\theta_1) = \mathcal{N} v_0^2 = 9 \times \frac{121}{3} = 363.$$

Let $G_1(u) = U(a_1, b_1)$. For a uniform distribution:

$$\mu(\theta_1) = \frac{a_1 + b_1}{2}, \quad v^2(\theta_1) = \frac{(b_1 - a_1)^2}{12}.$$

From $\mu_0 = 2$:

$$\frac{a_1 + b_1}{2} = 2 \implies a_1 + b_1 = 4.$$

From $N v_0^2 = v^2(\theta_1)$:

$$\frac{(b_1 - a_1)^2}{12} = 363 \implies (b_1 - a_1)^2 = 4356.$$

$$b_1 - a_1 = 66 \quad \text{(taking the positive root since } b_1 > a_1\text{)}.$$

Solve the system:

$$a_1 + b_1 = 4, \quad b_1 - a_1 = 66.$$

Add the two equations:

$$2b_1 = 70 \implies b_1 = 35.$$

$$a_1 = 4 - 35 = -31.$$

Thus:

$$\theta_1 = (-31, 35) \implies G_1(u) = U(-31, 35).$$

—

For G_2:

First, compute $v_0 = \sqrt{40.3333} \approx 6.349$.

$$\mu_0 + 2v_0 = 2 + 2 \times 6.349 = 2 + 12.698 = 14.698.$$

Also:

$$v_0^2 = Sv^2(\theta_2) \implies 40.3333 = 9v^2(\theta_2) \implies v^2(\theta_2) = \frac{40.3333}{9} \approx 4.48148.$$

For $G_2(u) = U(a_2, b_2)$:

$$\frac{a_2 + b_2}{2} = 14.698 \implies a_2 + b_2 = 29.396.$$

$$\frac{(b_2 - a_2)^2}{12} = 4.48148 \implies (b_2 - a_2)^2 = 53.7777.$$

$$b_2 - a_2 = \sqrt{53.7777} \approx 7.3333.$$

Solve:

$$a_2 + b_2 = 29.396, \quad b_2 - a_2 = 7.3333.$$

Add the two:

$$2b_2 = 36.7293 \implies b_2 = 18.36465.$$

$$a_2 = 29.396 - 18.36465 = 11.03135.$$

Thus:

$$\theta_2 = (11.03135, 18.36465) \implies G_2(u) = U(11.03135, 18.36465).$$

3.3 Computing p_1 and p_2

Given:

$$\tau = \frac{1}{1 + I_1}, \quad I_1 = 5 \implies \tau = \frac{1}{6}.$$

$$\alpha_1 = 0.1, \quad \alpha_2 = 0.01.$$

$$p_1 = (\alpha_1)^{1-\tau}(\alpha_2)^{\tau} = (0.1)^{5/6}(0.01)^{1/6}.$$

Compute approximately: - $\alpha_1^{5/6} = 0.1^{0.8333...} = e^{0.8333 \ln(0.1)} \approx 0.146$. - $\alpha_2^{1/6} = 0.01^{1/6} = e^{(1/6)\ln(0.01)} \approx 0.464$.

Thus:

$$p_1 \approx 0.146 \times 0.464 = 0.0677.$$

Then:

$$p_2 = \frac{5p_1}{\sqrt{S}} = \frac{5 \times 0.0677}{\sqrt{9}} = \frac{0.3385}{3} \approx 0.11283.$$

3.4 The Mixture Distributions for Testing

We consider testing:

$$H_0 : F_Y = G_0 . H' : F_Y \neq G_0 .$$

We will compare the empirical distribution of samples generated from:
1. $F_Y = (1 - p_1)G_0 + p_1 G_1$, i.e. a mixture of G_0 and G_1. 2. $F_Y = (1 - p_2)G_0 + p_2 G_2$, i.e. a mixture of G_0 and G_2.

The tests are conducted for sample sizes:

$$N_1 = 10 \times (2 + \mathcal{N}) = 10 \times (2 + 9) = 10 \times 11 = 110,$$

$$n_2 = 100 \times (2 + \mathcal{N}) = 100 \times 11 = 1100.$$

3.5 Goodness-of-Fit Tests

We use the Kolmogorov-Smirnov (KS) test for the given samples $(Y_t)_{t=1}^{n}$:

The test statistic is:

$$D_n = \sup_u |F_n(u) - G_0(u)|,$$

where F_n is the empirical distribution function (EDF) based on the sample.

Since:

$$F_Y(u) = (1 - p_k)G_0(u) + p_k G_k(u),$$

we have:

$$F_Y(u) - G_0(u) = p_k[G_k(u) - G_0(u)],$$

for $k = 1$ or $k = 2$.

Thus, the maximum deviation from G_0 is:

$$\sup_u |F_Y(u) - G_0(u)| = p_k \sup_u |G_k(u) - G_0(u)|.$$

We need $\sup_u |G_1(u) - G_0(u)|$ and $\sup_u |G_2(u) - G_0(u)|$.

3.5.1 Case 1: G_1 vs. G_0

$G_0 = U(-9, 13)$, so:

$$G_0(u) = \begin{cases} 0 & u < -9 \\ \frac{u+9}{22} & -9 \leq u \leq 13 \\ 1 & u > 13 \end{cases}$$

$G_1 = U(-31, 35)$, so:

$$G_1(u) = \begin{cases} 0 & u < -31 \\ \frac{u+31}{66} & -31 \leq u \leq 35 \\ 1 & u > 35 \end{cases}$$

3.5.2 Case 2: G_2 vs. G_0

3.6 KS Test Critical Values and Detection Probability

3.7 Approximate p-values

3.8 Conclusion