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 $G_1, \ldots, G_m$ be given distribution functions and $p_1, \ldots, p_m$ 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 $G_1, \ldots, G_m$ with probabilities (or weights) $p_1, \ldots, p_m$.

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

1. Choose $k \in \{1, \ldots, m\}$ at random with probabilities (or weights) $p_1, \ldots, p_m$. 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 task, 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

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

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

2.1 Parametric Family Selection

Using the assigned formula $\ell := \left\lfloor \frac{\mathcal{I}_2 + 2.5}{2} \right\rfloor$, we find $\ell = 5$. Thus, we will use the parametric family $\mathcal{G}_5(\Theta)$ in this task.

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

It can be expressed as:

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

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

3 Task 1: Testing Goodness-of-Fit

3.1 Basic Distribution Function

The problem gives a specific basic parameter:

$$\boldsymbol{\theta}_0 = (-\mathcal{N}, \mathcal{S}+4) = (-9, 13).$$

Thus, the basic distribution function is:

$$G_0(u) = \mathcal{G}_5(u|\boldsymbol{\theta}_0) = U(-9, 13).$$

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

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

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

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

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

3.2 Finding Mixture Distributions

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

$$\mu_0 = \mu(\boldsymbol{\theta}_1), \quad \mathcal{N}v_0^2 = v^2(\boldsymbol{\theta}_1).$$

$$\mu_0 + 2v_0 = \mu(\boldsymbol{\theta}_2), \quad v_0^2 = \mathcal{S}v^2(\boldsymbol{\theta}_2).$$

3.2.1 Determining G_1

First we determine G_1. We have:

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

It is given that $\mu(\theta_1) = \mu_0 = 2$.
Plugging in $\mathcal{N} = 9$ and $v_0^2 = \frac{121}{3}$, we get:

$$v^2(\boldsymbol{\theta}_1) = \mathcal{N}v_0^2 = 9 \times \frac{121}{3} = 363.$$

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

$$\mu(\boldsymbol{\theta}_1) = \frac{a_1+b_1}{2}, \quad v^2(\boldsymbol{\theta}_1) = \frac{(b_1-a_1)^2}{12}.$$

Since $\mu(\boldsymbol{\theta}_1) = 2$:

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

Since $v^2(\boldsymbol{\theta}_1) = 363$:

$$\frac{(b_1-a_1)^2}{12} = 363 \implies b_1 - a_1 = \sqrt{(b_1-a_1)^2} = \sqrt{4356} = 66.$$

Solving the system:

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

Adding the two equations:

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

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

Thus:

$$\boldsymbol{\theta}_1 = (-31, 35) \implies G_1(u) = U(-31, 35).$$

3.2.2 Determining G_2

Repeating the process for G_2. We have:

$$\mu_0 + 2v_0 = \mu(\boldsymbol{\theta}_2), \quad v_0^2 = \mathcal{S}v^2(\boldsymbol{\theta}_2).$$

Given $\mu_0 = 2$ and $v_0^2 = 40.3333$, we have:

$$\mu(\boldsymbol{\theta}_2) = \mu_0 + 2v_0 = 2 + 2 \times \sqrt{40.3333} = 2 + 2 \times 6.3509 = 2 + 12.7018 = 14.7018.$$

Also:

$$v_0^2 = \mathcal{S}v^2(\boldsymbol{\theta}_2) \implies 40.3333 = 9v^2(\boldsymbol{\theta}_2) \implies v^2(\boldsymbol{\theta}_2) = \frac{40.3333}{9} \approx 4.4815.$$

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

$$\frac{a_2 + b_2}{2} = 14.7018 \implies a_2 + b_2 = 29.4036.$$

$$\frac{(b_2 - a_2)^2}{12} = 4.4815 \implies (b_2 - a_2)^2 = 4.4815 \times 12 = 53.7777.$$

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

Solving the system:

$$a_2 + b_2 = 29.4036, \quad b_2 - a_2 = 7.3333.$$

Adding the two equations:

$$2b_2 = 36.7369 \implies b_2 = 18.3685.$$

$$a_2 = 29.4036 - 18.3685 = 11.0351.$$

Thus:

$$\boldsymbol{\theta}_2 = (-11.0351, 18.3685) \implies G_2(u) = U(11.0351, 18.3685).$$

3.3 Computing p_1 and p_2

Given:

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

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

$$p_1 = (\alpha_1)^{1-\tau}(\alpha_2)^{\tau} = (0.1)^{5/6}(0.01)^{1/6} \approx 0.06813.$$

Then:

$$p_2 = \frac{5p_1}{\sqrt{S}} = \frac{5 \times 0.06813}{\sqrt{9}} = \frac{0.3406}{3} \approx 0.1135.$$

3.4 Determining Mixture Distributions

We consider testing:

$$H_0 : F_Y = G_0 \textbf{ versus } H' : F_Y \neq G_0$$

We will compare the empirical distribution of samples generated from:

1. $F_Y = (1 - p_1)G_0 + p_1G_1$, i.e. a mixture of G_0 and G_1.
2. $F_Y = (1 - p_2)G_0 + p_2G_2$, i.e. a mixture of G_0 and G_2.

The tests are conducted for sample sizes:

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

$$N_2 = 100 \times (2 + \mathcal{N}) = 100 \times (2 + 9) = 1100.$$

3.5 Goodness-of-Fit Tests

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

The test statistic is:

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

where F_n is the empirical distribution function (EDF) based on the sample and F is the theoretical distribution function. In this case, $F = G_0$.

Since:

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

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 difference between F_Y and 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 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:

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

To find $\sup |G_1(u) - G_0(u)|$, we investigate ranges of u piecewise between the breakpoints of the two functions.

1. For $u < -31$: $G_0(u) = G_1(u) = 0$, so the difference is 0.
2. For $-31 \leq u < -9$: $G_0(u) = 0$, $G_1(u) = \frac{u+31}{66}$. The difference is $\frac{u+31}{66}$, which is increasing as u approaches -9, where it is $\frac{-9+22}{66} = \frac{1}{3}$.

3. For $-9 \leq u < 13$: $G_0(u) = \frac{u+9}{22}$, $G_1(u) = \frac{u+31}{66}$. The difference is $\frac{u+31}{66} - \frac{u+9}{22} = \frac{2u-4}{66}$, which is increasing from $-\frac{1}{3}$ at -9 to $\frac{1}{3}$ at 13.

4. For $13 \leq u < 35$: $G_0(u) = 1$, $G_1(u) = \frac{u+31}{66}$. The difference is $\frac{u+31}{66} - 1 = \frac{u-35}{66}$, which is increasing from $-\frac{1}{3}$ at 13 to 0 at 35.

5. For $u \geq 35$: $G_0(u) = G_1(u) = 1$, so the difference is 0.

The maximum absolute difference is $\frac{1}{3}$ at the endpoints of the range $[-9, 13]$.
Hence:

$$\sup_u |G_1(u) - G_0(u)| = 1/3 \approx 0.3333.$$

For the mixture, taking $p_1 \approx 0.06813$:

$$\sup_u |F_Y(u) - G_0(u)| = p_1 \times 0.3333 = 0.06813 \times 0.3333 \approx 0.02271.$$

3.5.2 G_2 vs. G_0

Repeating the process for G_2:
$G_0 = U(-9, 13)$, so:

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

$G_2 = U(11.0351, 18.3685)$, so:

$$G_2(u) = \begin{cases} 0 & u < 11.0351 \\ \frac{u-11.0351}{7.3333} & 11.0351 \leq u \leq 18.3685 \\ 1 & u > 18.3685 \end{cases}$$

To find $\sup |G_2(u) - G_0(u)|$, we investigate ranges of u piecewise between the breakpoints of the two functions.

1. For $u < -9$: $G_0(u) = G_2(u) = 0$, so the difference is 0.

2. For $-9 \leq u < 11.0351$: $G_0(u) = \frac{u+9}{22}$, $G_2(u) = 0$. The difference is $\frac{u+9}{22}$, which is increasing as u approaches 11.0351, where it is $\frac{11.0351+9}{22} \approx 0.9107$.

3. For $11.0351 \leq u < 13$: $G_0(u) = \frac{u+9}{22}$, $G_2(u) = \frac{u-11.0351}{7.3333}$. The difference is $\frac{u-11.0351}{7.3333} - \frac{u+9}{22} = \frac{3u-33.1053}{22} - \frac{u+9}{22} = \frac{2u-42.1053}{22}$, which is increasing from -0.9107 at 11.0351 to $\frac{13-42.1053}{22} \approx -0.7321$ at 13.

4. For $13 \leq u < 18.3685$: $G_0(u) = 1$, $G_2(u) = \frac{u-11.0351}{7.3333}$. The difference is $\frac{u-11.0351}{7.3333} - 1 = \frac{u-18.3685}{7.3333}$, which is increasing from $-\frac{18.3685-13}{7.3333} = -\frac{5.3685}{7.3333} \approx -0.7321$ at 13 to 0 at 18.3685.

5. For $u \geq 18.3685$: $G_0(u) = G_2(u) = 1$, so the difference is 0.

The maximum absolute difference is ≈ 0.9107 at 11.0351.
Hence:

$$\sup_u |G_2(u) - G_0(u)| \approx 0.9107.$$

For the mixture, taking $p_2 \approx 0.1135$:

$$\sup_u |F_Y(u) - G_0(u)| = p_2 \times 0.9107 = 0.1135 \times 0.9107 \approx 0.1034.$$

3.6 Critical Values and Detection Probability

Under H_0, the Kolmogorov-Smirnov test critical values at significance $\alpha_1 = 0.1$ and $\alpha_2 = 0.01$ are approximately:

$$D_{N,\alpha_1} \approx \frac{1.22}{\sqrt{N}} \quad \text{for} \quad \alpha_1 = 0.1,$$

$$D_{N,\alpha_2} \approx \frac{1.63}{\sqrt{N}} \quad \text{for} \quad \alpha_2 = 0.01.$$

For $N = 110$:

$$D_{110,0.1} \approx \frac{1.22}{\sqrt{110}} \approx 0.1163 \quad \text{and} \quad D_{110,0.01} \approx \frac{1.63}{\sqrt{110}} \approx 0.1554.$$

- For G_1: $\sup |F_Y - G_0| \approx 0.02271 < 0.1163 < 0.1554$. Thus, at $N = 110$, it's unlikely we reject H_0. $p > 0.1$
- For G_2: $\sup |F_Y - G_0| \approx 0.1034 < 0.1163 < 0.1554$. There is some chance to reject at a higher α level. $p > 0.1$ (but it's closer to the borderline)

For $N = 1100$:

$$D_{1100,0.1} \approx \frac{1.22}{\sqrt{1100}} \approx 0.03678 \quad \text{and} \quad D_{1100,0.01} \approx \frac{1.63}{\sqrt{1100}} \approx 0.04915.$$

- For G_1: $\sup |F_Y - G_0| \approx 0.02271 < 0.03678 < 0.04915$. Even with 1100 samples, we will likely not reject H_0 at $\alpha = 0.1$. $p > 0.1$
- For G_2: $\sup |F_Y - G_0| \approx 0.03678 < 0.04915 < 0.1034$. We will almost certainly reject H_0 at $\alpha = 0.01$. $p < 0.01$

Thus, the results of the Kolmogorov-Smirnov test are as follows:

Mixture	**Sample Size**	**p-value**	**Result**
G_1	110	> 0.1	No rejection
G_1	1100	> 0.1	No rejection
G_2	110	> 0.1	No rejection
G_2	1100	< 0.01	Rejection at $\alpha = 0.01$

3.7 Conclusions

By analytically comparing the theoretical distributions, we have:

- **Mixture with G_1:**
  $\sup |F_Y - G_0| \approx 0.02271$.
  Even at $N = 1100$, we will likely not reject H_0 at $\alpha = 0.1$. The p-value is higher than 0.1.
- **Mixture with G_2:**
  $\sup |F_Y - G_0| \approx 0.1034$.
  We will almost certainly reject H_0 at $\alpha = 0.01$ even with 1100 samples. The p-value is < 0.01. We will likely not reject with 110 samples at $\alpha = 0.1$, but the p-value may be close.

It is evident that for Kolmogorov-Smirnov tests, the magnitude of the deviation from G_0 and the sample size play the decisive role.

As $N \to \infty$, if $F_Y \neq G_0$, the empirical distribution F_N converges to F_Y, and thus D_N converges to $\sup_u |F_Y(u) - G_0(u)|$.

4 Task 2: Applications of Bootstrap Technique

In this section we will

- Test Complex Goodness of Fit Hypothesis,
- Check bootstrap consistency,
- Compare bootstrap confidence interval construction methods.

4.1 Testing Goodness-of-Fit by Bootstrap

In Task 1, we considered a simple hypothesis test for the goodness-of-fit of the data to the distribution G_0. In this section, we will test the complex Goodness-of-Fit hypothesis that the unknown distribution function F_Y belongs to the parametric family $\mathcal{G}(\Theta)$:

$$H_0 : F_Y \in \mathcal{G}(\Theta) \quad \text{versus} \quad H' : F_Y \notin \mathcal{G}(\Theta).$$

We will use the same parametric family $\mathcal{G}_5(\Theta)$ and the same distributions G_0, G_1, G_2 as in Task 1:

$$G_0(u) = U(-9, 13), \quad G_1(u) = U(-31, 35), \quad G_2(u) = U(11.0351, 18.3685).$$

We will make use of the parametric bootstrap technique to test the hypothesis. The test statistic is the Kolmogorov-Smirnov test statistic and the significance level is $\alpha = 0.1$.

The parametric bootstrap algorithm for testing this hypothesis is as follows:

1. **Generate the sample:** With sample sizes $N_1 = 110$ and $N_2 = 1100$, generate data $(Y_t)_1^N$ from the mixture distributions F_Y from Task 1:

$$(1 - p_1)\, G_0 + p_1\, G_1 \quad \text{or} \quad (1 - p_2)\, G_0 + p_2\, G_2,$$

2. **Estimate $\hat{\boldsymbol{\theta}}_N$:** For the sample Y^N, assume that $F_Y \in \mathcal{G}(\Theta)$. We estimate the parameter $\boldsymbol{\theta}$ by the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_N$. For the uniform distribution, the MLE amounts to $\hat{a} = \min(Y^N)$, $\hat{b} = \max(Y^N)$. We denote the fitted distribution by $\hat{G}_N(u) := G(u|\hat{\theta}_N)$.

3. **Calculate the test statistic:** Calculate the Kolmogorov-Smirnov test statistic T comparing the EDF of the data Y^N and the fitted distribution $\hat{G}_N$. We denote the test statistic by $\hat{T}_N$.

4. **Generate the bootstrap samples:** Generate $B = 100 \times N$ bootstrap samples $Y_b^{N*}, b \in \{1, \ldots, B\}$ by resampling with replacement from the original sample Y^N. For each bootstrap sample Y_b^{N*}, estimate the parameter θ by the same method as in step 2, obtaining $\hat{\theta}_b^{N*}$.

5. **Calculate the bootstrap test statistics:** For each bootstrap sample Y_b^{N*}, calculate the Kolmogorov-Smirnov test statistic $\hat{T}_{N,b}^*$ comparing the EDF of the data Y_b^{N*} and the fitted distribution $G(u|\hat{\theta}_{N,b}^*)$.

6. **Calculate the approximate p-value:** Calculate the p-value as the proportion of bootstrap test statistics $\hat{T}_{N,b}^*$ that are greater than $\hat{T}_N$. If the p-value is less than the significance level α, reject the null hypothesis. Formally:

$$\hat{p}^* = \hat{p}^*(T, \mathcal{G}, Y^N, B) = \frac{\#\{b : \hat{T}_{N,b}^* > \hat{T}_N\}}{B}.$$

7. **Decision:** If $\hat{p}^* < \alpha$, reject the null hypothesis H_0. Otherwise, do not reject H_0.

4.1.1 Simulation Results

We will now simulate the parametric bootstrap procedure for the two sample sizes $N_1 = 110$ and $N_2 = 1100$. We will generate $B = 100 \times N$ bootstrap samples and calculate the p-values for each sample.

Mixture of Distributions	Sample Size	Test Statistic	p-value	Decision
G_0, G_1	110	0.3419	$< 10^{-5}$	Rejection at $\alpha = 0.1$
G_0, G_1	1100	0.3066	$< 10^{-5}$	Rejection at $\alpha = 0.1$
G_0, G_2	110	0.0944	0.266	No rejection
G_0, G_2	1100	0.1203	$< 10^{-5}$	Rejection at $\alpha = 0.1$

4.1.2 Comparison with Simple Hypothesis Tests and Discussion

In Task 1(b), we tested a simple hypothesis

$$H_0 : F_Y = G_0 \textbf{ versus } H' : F_Y \neq G_0$$

where G_0 was a fixed, known distribution. Here, by contrast, we do not know the parameter θ in advance. The parametric bootstrap procedure accounts for the uncertainty in θ by re-fitting the parameter for each bootstrap sample.

The results of the tests are interesting and not consistent with the Task 1 results. In Task 1, we found that the mixture with G_1 was not rejected at $\alpha = 0.1$ for both sample sizes. However, the parametric bootstrap test rejected the hypothesis for both sample sizes. The mixture with G_2 was rejected at $\alpha = 0.01$ for $N = 1100$ in Task 1, and the bootstrap test also rejected the hypothesis at $\alpha = 0.1$.

In this case, the discrepancy related to G_1 can be explained by the following reasons:

- We are using Kolmogorov-Smirnov tests. This statistic is sensitive to the maximum deviation between the empirical distribution function and the fitted distribution.
- Since p_1 and p_2 are small, the mixture distributions are close to the basic distribution G_0, the estimated Kolmogorov-Smirnov test statistics are low and the associated p-values are high.
- In Task 2, we do not know the parameter a priori. The maximum likelihood estimator is used to estimate θ from the data, and the estimated values have a significantly wider range than the basic distribution. This leads to higher test statistics and lower p-values in the G_1 mixture.

Tests on the mixture with G_2 are consistent between Task 1 and Task 2. The range of the parameter θ is narrower for G_2 than for G_1, and the estimated distribution is closer to the basic distribution G_0. This leads to lower test statistics and higher p-values in the G_2 mixture. More samples are needed to detect the deviation a simple uniform distribution.

If the parametric family $\mathcal{G}(\Theta)$ is known and correct, the parametric bootstrap can be used to test whether the data fits the parametric family. Nonparametric bootstrap can be used when the parametric family cannot be assumed. The nonparametric bootstrap procedure is similar to the parametric bootstrap, but the parameter θ is not estimated. While the parametric bootstrap is more powerful when the parametric family is correct, the nonparametric bootstrap is more robust when the parametric family is incorrect.

4.2 Checking Bootstrap Consistency

4.3 Bootstrap Confidence Intervals