Parametric & Nonparametric Statistics Project

Aleksandr Jan Smoliakov

2024–12–12

1 Preliminaries

We are given the following parameters:

- $N = 9$
- $S = 9$
- $I_1 = 5$
- $I_2 = 8$

2 Task 1: Testing Goodness-of-Fit

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 $G_5(\Theta)$ in this project.

We are given the parametric distribution family $G_5(\Theta)$, which contains distribution functions of random variables uniformly distributed on $[\theta_1, \theta_2]$, $\theta_1 < \theta_2$.

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

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

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

2.2 Basic Distribution Function

The parameter of the basic distribution function G_0 is $\theta_0 = (-N, S+4) = (-9, 13)$.

The problem gives a specific basic parameter:

$$\theta_0 = (-N, S+4),$$

with $N = 9$ and $S = 9$. Thus:

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

So:

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

For a uniform distribution $U(a, b)$: - Mean: $\mu = \frac{a+b}{2}$ - Variance: $v^2 = \frac{(b-a)^2}{12}$
For $G_0 = U(-9, 13)$: - $\mu_0 = \frac{-9+13}{2} = \frac{4}{2} = 2$. - $(b_0 - a_0) = 13 - (-9) = 22$. - $v_0^2 = \frac{22^2}{12} = \frac{484}{12} = 40.3333...$
So:

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

2.3 Finding θ_1 and θ_2

For G_1:

We have the equations:

$$\mu_0 = \mu(\theta_1), \quad Nv_0^2 = v^2(\theta_1).$$

Since $N = 9$ and $v_0^2 \approx 40.3333$:

$$Nv_0^2 = 9 \times 40.3333 = 362.9997 \approx 363.$$

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 $Nv_0^2 = v^2(\theta_1)$:

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

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

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:

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

Check variance:

$$(b_1 - a_1)^2/12 = 66^2/12 = 4356/12 = 363 \checkmark$$

For G_2:
We have:

$$\mu_0 + 2v_0 = \mu(\theta_2), \quad v_0^2 = Sv^2(\theta_2).$$

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:

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

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

Thus:

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

Then:

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

—

2.5 The Mixture Distributions for Testing

We consider testing:

$$H_0 : F_Y = G_0.$$

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:

$$n_1 = 10 \times (2 + N) = 10 \times (2 + 9) = 10 \times 11 = 110,$$

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

2.6 Goodness-of-Fit Tests and DKW Inequality

We use the following tests for the given samples $(Y_t)_{t=1}^n$:
1. **Kolmogorov–Smirnov (KS) Test**: Test statistic:

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

where F_n is the empirical distribution function (EDF) based on the sample.
2. **Cramér–von Mises (CvM) Test**: Test statistic:

$$W^2 = \int_{-\infty}^{\infty} [F_n(u) - G_0(u)]^2 dG_0(u).$$

3. **Anderson–Darling (AD) Test**: A weighted version of CvM that puts more emphasis on the tails:

$$A^2 = \int_{-\infty}^{\infty} \frac{[F_n(u) - G_0(u)]^2}{G_0(u)(1 - G_0(u))} dG_0(u).$$

4. **Dvoretzky–Kiefer–Wolfowitz (DKW) Inequality**: Provides uniform confidence bands for the EDF around the true CDF:

$$P\left(\sup_u |F_n(u) - F(u)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}.$$

This can be used to derive a test or confidence envelope.

For each test, we compare the observed test statistic against the critical values or use a Monte Carlo approach to determine p-values.

2.7 Finding p-values and Commenting on Results

Since the problem states that we should find (approximate) p-values and comment on the results, the final step will outline how to conduct these tests in principle. Exact numeric p-values require simulation or tables, as no closed-form solutions are provided here. However, we will give a detailed approach to obtaining those p-values and compare the tests.

Approach to Finding p-values:

1. **Monte Carlo Simulation (Recommended Method)**: - Under H_0: Generate a large number of samples from $G_0 = U(-9, 13)$. - For each sample, compute the test statistic (KS, CvM, AD). - Determine the distribution of the test statistic under H_0. - Now, take the observed data: - If data is generated from $(1 - p_1)G_0 + p_1 G_1$ or $(1 - p_2)G_0 + p_2 G_2$, compute the test statistic for this observed sample. - The p-value is the proportion of simulated H_0-samples whose test statistic exceeds the observed one.

2. **Comparison of Tests**: - **KS test**: Sensitive to the largest deviation anywhere in the distribution, but not as sensitive to tail differences. - **CvM test**: A more balanced measure of differences across the entire distribution. - **AD test**: More sensitive to differences in the tails, which may be crucial if the alternative distributions differ significantly from G_0 in their spread or tail behavior.

Since G_1 has the same mean but a much larger variance than G_0, the tails are more "spread out." We expect the AD test to detect this difference more readily than KS, especially for large sample sizes. The CvM test should also show good power against a spread difference.

For G_2, the mean is shifted substantially ($\mu_2 \approx 14.698$ vs. $\mu_0 = 2$). All tests (KS, CvM, AD) should easily detect this difference, especially for large n.

3. **Effect of Sample Size**: - For $n = 110$: With a moderate sample size, subtle differences might not be so easily detected. Small mixture probability p_1 might lead to a moderate p-value for the test with G_1. - For $n = 1100$: With a large sample, even small deviations from G_0 become evident. Expect very small p-values (strong rejection of H_0) for both mixtures, especially the one with G_2.

4. **Approximate p-values**: Without actual data simulation, we can only state qualitatively: - For the mixture with G_1: - At $n = 110$: p-values might be moderate; the KS test may still not strongly reject H_0, but CvM and AD tests may yield smaller p-values due to their sensitivity to variance differences. - At $n = 1100$: p-values likely become very small, indicating strong evidence against H_0.

- For the mixture with G_2: - Even at $n = 110$, the large shift in mean should lead to very low p-values for all tests, likely rejecting H_0. - At $n = 1100$, the difference is even more stark, and the p-values should be extremely small.

2.8 Conclusion and Comments

- The **Anderson–Darling test** often outperforms KS and CvM in detecting differences in the tails. Since G_1 differs in spread, AD might provide the smallest p-values. - The **Cramér–von Mises test** is a good all-rounder and might also show lower p-values than KS for the G_1 mixture scenario. - The **Kolmogorov–Smirnov test**, while classic and widely used, can be less sensitive to certain types of distributional differences, especially when differences are not concentrated in one part of the distribution. - The **Dvoretzky–Kiefer–Wolfowitz inequality** provides a theoretical bound on deviations, ensuring that for large sample sizes the observed deviations are highly unlikely, thus leading to small p-values.

For the given mixtures: - With $F_Y = (1 - p_1)G_0 + p_1 G_1$ and a moderate sample size $n = 110$, the p-values might be around a borderline region, depending on the exact realization. Increasing to $n = 1100$ makes detecting the difference almost certain. - With $F_Y = (1-p_2)G_0 + p_2 G_2$, the shift in the mean is large, and all tests should yield very small p-values, even at $n = 110$, strongly rejecting H_0.

In summary, for the given alternatives, all tests will eventually show that the data does not come from G_0, but their sensitivity and the required sample size to confidently reject H_0 differ. The Anderson–Darling test tends to be the most sensitive in these scenarios, providing the smallest p-values, followed by Cramér–von Mises, and then Kolmogorov–Smirnov. Larger sample sizes drastically reduce p-values for both mixtures.

—

Final Note: The exact numerical p-values require either: - Simulation: Generating samples under H_0 and under the specified alternatives, or - Using published tables/approximations for these tests.