Towards multivariate p-values and FDR control based on a class of Kolmogorov-Smirnov two sample tests

Soma S. Dhavala
** working paper **

Abstract:

We consider a class of two sample tests which is a generalization of the widely used Kolomogorov-Smirnov tests. The "two-sample Kolmogorov-Smirnov test" tests the alternate hypothesis that the two samples under consideration come from different populations. The test statistic is the supremum of the absolute deviations between the two empirical distribution functions. We propose to use order statistics of these absolute deviations as test statistics. We empirically observed that the distribution of these test statistics is also independent of true distribution under null hypothesis. We conjecture that these ideas can be used to generate multivariate p-values which offer better control over false discovery rate (FDR) or in choosing one candidate test among the class which offers the desired FDR.

Some Key Words: Distribution-free tests, Nonparametric tests, Two sample tests, Kolomogorov-Smirnov, Robust tests

Short title: Generalized Kolmogorov-Smirnov tests,

Kolmogorov-Smirnov test

We want to the test the hypothesis:

$$H_0 : F(x) = G(x) \text{ Vs } H_a : F(x) \neq G(x)$$

where $F(x)$ and $G(x)$ are the distribution functions of two populations under consideration. A widely used distribution-free test is the Kolmogorov-Smirnov test. It is based on the following statistic:

$$D = \sup |F_n(x) - G_m(x)|$$

where $F_n(x)$ is the empirical distribution function of the samples $x_1, x_2, \ldots, x_n$ and $G_m(x)$ is the empirical distribution function of the samples $y_1, y_2, \ldots, y_m$. The test is consistent because the empirical distributions are consistent estimators of the distributions.

Difficulty

As can be seen, the test statistic is based on the largest discrepancy between the empirical distributions. Consequently, the statistic is very sensitive to noisy or rough empirical distributions. For example, in Figure 1, we plotted a pdf and its corrupted version in the top-panel. As can be seen, the noisy pdf resembles the shape of the true pdf but wiggly. The K-S test is sensitive to these shapes. This effect is dramatic as the sample size increases.

Proposal

Instead of considering the largest deviation, we can consider the median absolute deviation. More, generally, we can consider any quantile of this random-variable as the statisitc. And all of them are consistent (due to Corollory 1.1).

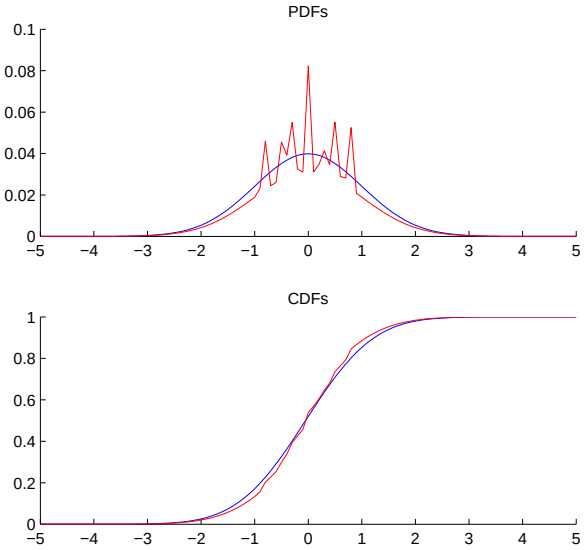

Figure 1: Noisy PDF

Proposal Statistic

$$D = \text{median}|F_n(x) - G_m(x)|$$

The above statistic is also distribution free. Analytical derivation of the distribution of this statistic is quite difficult. Instead, we propose to generate the critical values based on simulations.

Monte-Carlo simulation

- for b=1:B (B number of MC paths)
  - Generate n of $u_1 \sim U(0,1)$ and $u_{1s} = \text{sort}(u)$ (in ascending order)
  - Generate m of $u_1 \sim U(0,1)$ and $u_{2s} = \text{sort}(u)$ (in ascending order)
  - Compute $D_b = \text{median}|u_{1s} - u_{2s}|$
- Compute the desired critical values

Below, we show the histogram of the statisitc with n=m=1000 based on 20000 iterations. The algorithm is extremely simple and one can easily generate the critical values on the fly.

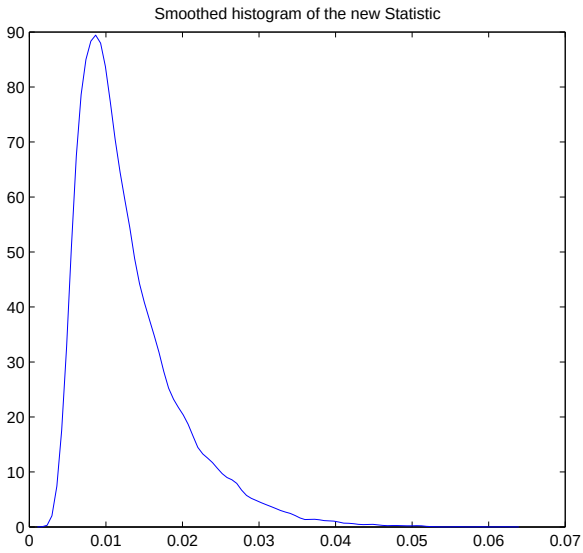

Figure 2: PDF of new K-S statistic at n=1000

Applications

- Testing for correlation in time-series data in the spectral domain
- Multiple hypotheis testing in Microarrays. It may be possible to develop an FDR controlling scheme by using a different quantile level in each of the individual hypothesis tests
- Diagnostic tests in MCMC, where the above tests can be used to check if the distributions in any two chains are the same or not.

Theorem 1 (Glivenko-Canteli)**.** *:*

$$\lim_{n\to\infty} \sup_{x\in\Re} |F_n(x) - F(x)| \xrightarrow{a.s} 0$$

Proof. Refer Csorgo (1983), pp.4. □

Corollary 1.1.

$$\lim_{n\to\infty} Q_p(|F_n(x) - F(x)|) \xrightarrow{a.s} 0$$

where Q_p is the pth quantile ($p \in [0, 1]$).

Proof. Let $z = |F_n(x) - F(x)|$.

Then,

$$\begin{aligned} \sup_{x\in\Re} |F_n(x) - F(x)| &= \sup z \\ &= \inf\{z \text{ s.t } F_Z^{-1}(z) \geq 1\} \\ &= Q_1 \end{aligned}$$

where F_Z is the distribution function of z. Which implies that $\lim_{n\to\infty} Q_1 \xrightarrow{a.s} 0$. Therefore, $Q_p \forall p \in [0, 1] \xrightarrow{a.s} 0$ as $n \to \infty$ □

Acknowledgements

Prof. Jeff Hart for listening to this weird idea that defies statistical thinking ☺

References

Gibbons, J. K and Chakraborti, S. (2003). Nonparametric Statistical Inference. *Marcel Dekker Inc.*, Newyork.

Durbin, J. (1973). Distribution Theory for Tests Based on the Sample Distribution Functions *SIAM*, Philadelphia, PN.

Csorgo, M. (1983). Quantile Process with Statistical Applications *SIAM*, Philadelphia, PN.

Klebanov, L., Gordon, A., Xiao, Y., Land, H. and Yakovlev, Y. (2006). A permutation test motivated by microarray data analysis. *Computational Statistics and Data Analysis*, **50**, 3618-3628

Xiao, Y., Gordon, A., and Yakovlev, Y. (2006). A L_1 version of the Cramer-von Mises Tet for Two-Sample Comparison in Microarray Data Analysis. *EURASIP Journal on Bioinformatics and Systems Biology*, **2006**, 1-9

Anderson, T. W. (1962). On the distribution of two-sample Cramer-von Mises criterion. *The Annals of the Mathematical Statistics*. **33**, 1148-1159.

Feltz, C. J. (2002). Customizing generalizations of the Kolmogorov-Smirnov Goodness-of-Fit test. *Journal of Statistical Computing an Simulation*. **72**, 179-186.

Reschenhofer, E. (1997). Generalizations of the Kolmogorov-Smirnov test. *Computational Statistics & Data Analysis*. **24**, 433-441.

Hodges, J. L. Jr (1958). The significance probability of the Smirnov two-sample test. *Arkiv For Matematik, Astronomi och Fysik*. **3**, 469-486.

Gail, M. H. andGreen, B. S. (1976). A Generalization of the one-sided two-sample Kolmogorov-Smirnov statistic for Evaluating Diagnostic tests. *Biometrics*. **32**, 561-570.

Li, Q., Maasoumi, E. and Racine, J. S. (2009). A nonparametric test for equality of distributions with mixed categorical and continous data. *Journal of Econometrics*. **148**, 186-200.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "'goodness-of-fit"' criterion based on stochastic process. *The Annals of the Mathematical Statistics*. **23**, 193-212.