KSgeneral: Computing P-Values of the KS Test for (Dis)Continuous Null Distribution

Senren Tan (Joint work with Dimitrina S. Dimitrova, Vladimir K. Kaishev)

Faculty of Actuarial Science and Insurance, Cass Business School

senren.tan@cass.city.ac.uk

July 16, 2018

One-sample two-sided Kolmogorov-Smirnov (KS) statistic

Goodness-of-fit test statistic measuring how well the distribution of a sample $\{x_1, ..., x_n\}$ from n i.i.d. random variables $\{X_1, ..., X_n\}$ agrees with some unknown distribution $F_X(x)$

The test statistic, D_n, Kolmogorov (1933)

$$D_n = \sup_x |F_n(x) - F(x)|, \quad (1)$$

- n: sample size
- $F_n(x)$: empirical distribution function of $\{x_1, ..., x_n\}$, $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(X_i \le x)}$
- $F(x)$: a pre-specified cumulative distribution function (cdf) under the null hypothesis H_0 that the sample $\{X_1, ..., X_n\}$ follows, i.e., $F_X(x) = F(x)$.
- Distribution-free when $F(x)$ is continuous!

When $F(x)$ is not continuous

- Distribution of D_n in (1) depends on the cdf under H_0, i.e., distribution-free property fails!
- Many real-life applications in insurance and finance require goodness-of-fit tests on discrete or mixed distribution to data
- Examples:
  - claim amount modelled by mixed distributions in multi-layer excess-of-loss reinsurance
  - claim numbers modelled by discrete distributions in general insurance, e.g., car insurance
  - bank loan recovery rate modelled by mixed distribution (clustering of 0's and 1's, and continuous segment in between)
- KS test is conservative on mixed or discrete data, when $F(x)$ is assumed to be continuous (Noether (1963))
- KS test could have greater power than chi-squared test, when $F(x)$ is not continuous (Petitt and Stephens (1977))

R package **KSgeneral**

- To the best of our knowledge, no statistical packages exist for computing exact p-values of KS test, when $F(x)$ is mixed
- When $F(x)$ is purely discrete, `ks.test` function in R package **dgof** (Arnold and Emerson (2011)) calculates exact p-values of the KS test, (but only exact for sample size $n \leq 30$)
- We provide an R package **KSgeneral** which efficiently computes $P(D_n < q)$ when $F(x)$ is continuous, mixed or purely discrete, and thus obtain exact p-values of the KS test for any (small or large) sample size n, and any $q \in [0, 1]$, available from `https://CRAN.R-project.org/package=KSgeneral`
- The algorithm is based on expressing $P(D_n < q)$ as an appropriate double-boundary non-crossing probability and computing the latter probability using fast Fourier transform (FFT) technique, paper recently accepted by *Journal of Statistical Software*, available from `http://openaccess.city.ac.uk/18541`

Computing $P(D_n < q)$, when $F(x)$ is mixed

KSgeneral function:

- *mixed_ks_c_cdf(q, n, jump_points, Mixed_dist, ..., tol)*

An example of a mixed $F(x)$

- $$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ x + 0.2 & \text{if } 0 \leq x < 0.6, \\ 1 & \text{if } x \geq 0.6, \end{cases}$$
- Sample size n: 5
- q: $\frac{1}{5000}$, $\frac{2}{5000}$, ..., 1.

```
> n <- 5
> q <- 1:5000/5000
```

Computing $P(D_n < q)$, when $F(x)$ is mixed

```
> Mixed_cdf_example <- function(x){
+   result <- 0
+   if (x < 0){
+     result <- 0
+   }
+   else if (x == 0){
+     result <- 0.2
+   }
+   else if (x < 0.6){
+     result <- 0.2 + x
+   }
+   else{
+     result <- 1
+   }
+   return (result)
+ }
```

Computing $P(D_n < q)$, when $F(x)$ is mixed

```
> plot(q, sapply(q, function(x) 1-KSgeneral::mixed_ks_c_cdf(x,
+      n, c(0, 0.6), Mixed_cdf_example)), type='l',
+      ylab = "P(D_{n} < q)")
> lines(q, sapply(q, function(x) 1-KSgeneral::cont_ks_c_cdf(x,
+      n)), type='l', col = 'red')
```

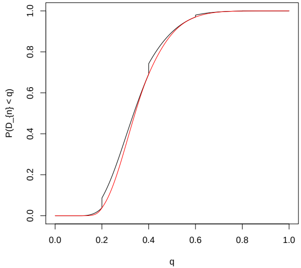

Figure: Plot of $P(D_n \geq q)$, $q \in [0, 1]$, when $F(x)$ is mixed

Computing p-values of KS test, when $F(x)$ is discrete

KSgeneral function:

```
disc_ks_test(x, y, ..., exact = NULL, tol = 1e-08,
             sim.size = 1e+06, num.sim = 10)
```

dgof function:

```
ks.test(x, y, ..., alternative = "two.sided",
   exact = NULL, tol=1e-8, simulate.p.value=FALSE, B=2000)
```

An example of a discrete $F(x)$

- $F(x)$ is the cdf of discrete uniform $[1, 10]$ random variable
- Sample sizes n: 25 or 500
- Sample simulated from discrete uniform $[1, 10]$ random variable

Computing p-values of KS test, when $F(x)$ is discrete

When the sample size is 25

```
> x3 <- sample(1:10, 25, replace = TRUE)
> KSgeneral::disc_ks_test(x3, ecdf(1:10), exact = TRUE)

        One-sample Kolmogorov-Smirnov test


data:  x3
D = 0.08, p-value = 0.9353949771749
alternative hypothesis: two-sided

> dgof::ks.test(x3, ecdf(1:10), exact = TRUE)

        One-sample Kolmogorov-Smirnov test


data:  x3
D = 0.08, p-value = 0.9353949771749
alternative hypothesis: two-sided
```

Computing p-values of KS test, when $F(x)$ is discrete

When the sample size is 500

```
> x4 <- sample(1:10, 500, replace = TRUE)
> KSgeneral::disc_ks_test(x4, ecdf(1:10), exact = TRUE)

        One-sample Kolmogorov-Smirnov test

data:  x4
D = 0.032, p-value = 0.4241393907967
alternative hypothesis: two-sided

> dgof::ks.test(x4, ecdf(1:10), exact = TRUE)

        One-sample Kolmogorov-Smirnov test

data:  x4
D = 0.032, p-value = 1
alternative hypothesis: two-sided
```

Conclusion

- R package **KSgeneral** efficiently computes $P(D_n < q)$ for small or large n, and any $q \in [0, 1]$, when $F(x)$ is continuous, mixed or discrete
- **KSgeneral** also efficiently computes p-values of KS test, when $F(x)$ is continuous, mixed or discrete
- Detailed numerical analysis can be found in Dimitrova et al. (2018)
- Our algorithm is also applicable for the weighted KS-type statistics (will update the package soon)

References

Arnold, T. B., and Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2), 34-39.

Dimitrova, D.S., Kaishev, V.K and Tan, S. (2018). Computing the Kolmogorov-Smirnov distribution when the underlying cdf is purely discrete, mixed or continuous. *Journal of Statistical Software*, forthcoming.

Kolmogorov, A. (1933). Sulla Determinazione Empirica di una Legge di Distributione. *Giornale dell'Istituto Italiano degli Attuari*, 4, 1-11.

Noether, G. E. (1963). Note on the Kolmogorov statistic in the discrete case. *Metrika*, 7(1), 115-116.

Petitt, A. N., and Stephens, M. A. (1977). The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics*, 19(2), 205-210.