Monter Carlo test for the Kolmogorov-Smirnov test in R

Piotr Nowicki

2024-11-14

The Kolmogorv-Smirnov (KS) test, along with other goodness-of-fit tests, assumes that the parameters of a distribution are provided a priori. However, these parameters are often estimated in practice. When parameters are estimated, the calculated p-values from the KS test may be innacurate. To obtain correct p-values, we can conduct a Monte Carlo test. This involves drawing samples from an assumed distribution and obtaining the distribution of the KS statistic or p-value under parameter estimation. This work provides an example of using the Monte Carlo test to correct the KS test.

```
library(ggplot2)
library(gridExtra)
library(grid)

set.seed(141124)

# Number of simulations
N <- 100000
```

Prepare data

```
# Suppose data come from N(5, 2) when the null hypothesis is true
norm_sims <- matrix(rnorm(30*N, mean=5, sd=2), nrow=N, ncol=30)

# Suppose data come from a gamma distribution when the alternative hypothesis is true
gamma_sims <- matrix(rgamma(30*N, shape=9, scale=1/2), nrow=N, ncol=30)
```

P-value distributions when parameters are known

```r
# P-value distribution when the null hypothesis is true
pvals_true <- apply(norm_sims, 1, function(row) ks.test(row, "pnorm", 5, 2)$p.value)

# How many p-values are below the threshold?
err_true_rate <- mean(pvals_true < 0.05)

# P-value distribution when the null hypothesis is false
pvals_false <- apply(gamma_sims, 1, function(row) ks.test(row, "pnorm", 5, 2)$p.value)

# How many p-values are below the threshold?
err_false_rate <- mean(pvals_false < 0.05)
```

Uncorrected p-value distributions when parameters are unknown

```r
# P-value distribution when the family is true
pvals_true_uncorrected <- apply(norm_sims, 1,
                                function(row) ks.test(
                                  row, "pnorm", mean(row), sd(row))$p.value)

# How many p-values are below the threshold?
err_true_rate_uncorrected <- mean(pvals_true_uncorrected < 0.05)

# P-value distribution when the family is false
pvals_false_uncorrected <- apply(gamma_sims, 1,
                                 function(row) ks.test(
                                  row, "pnorm", mean(row), sd(row))$p.value)

# How many p-values are below the threshold?
err_false_rate_uncorrected <- mean(pvals_false_uncorrected < 0.05)
```

Corrected p-value distributions when parameters are unknown

```r
# Simulate data from a referenced distribution and calculate p-values
ref_sims <- matrix(rnorm(30*N, mean=0, sd=1), nrow=N, ncol=30)
pval_distr <- apply(ref_sims, 1,
                    function(row) ks.test(
                      row, "pnorm", mean(row), sd(row))$p.value)

# Calculate p-value for your data and compare to the obtained p-value distribution above
pvals_true_corrected <- sapply(pvals_true_uncorrected,
                               function(pval) mean(pval_distr < pval))
```

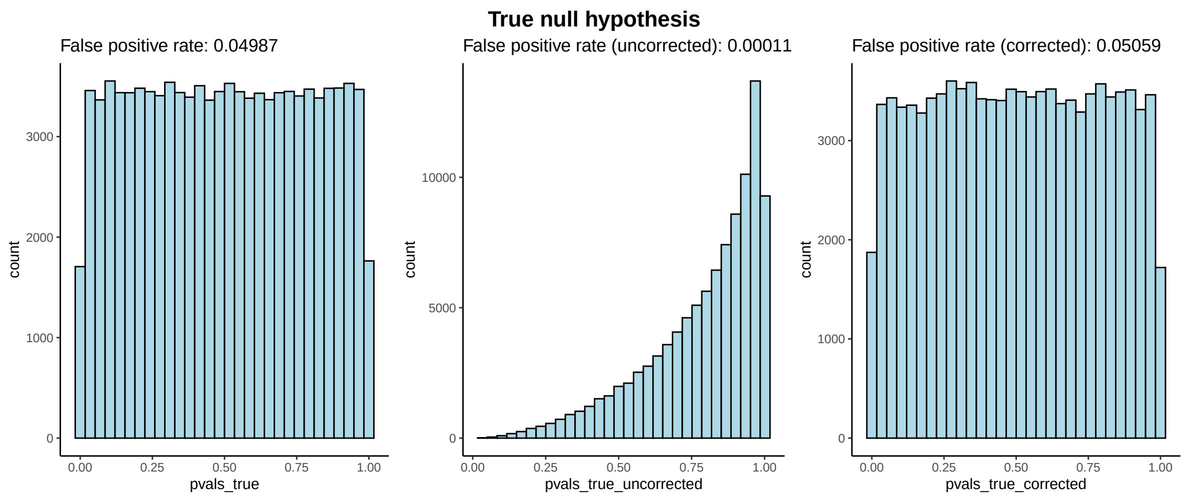
```
# How many p-values are below the threshold?
err_true_rate_corrected <- mean(pvals_true_corrected < 0.05)

# Repeat as above but suppose your data come from the gamma distribution
pvals_false_corrected <- sapply(pvals_false_uncorrected,
                                function(pval) mean(pval_distr < pval))

# How many p-values are below the threshold?
err_false_rate_corrected <- mean(pvals_false_corrected < 0.05)
```

Visualisation of the p-value distributions

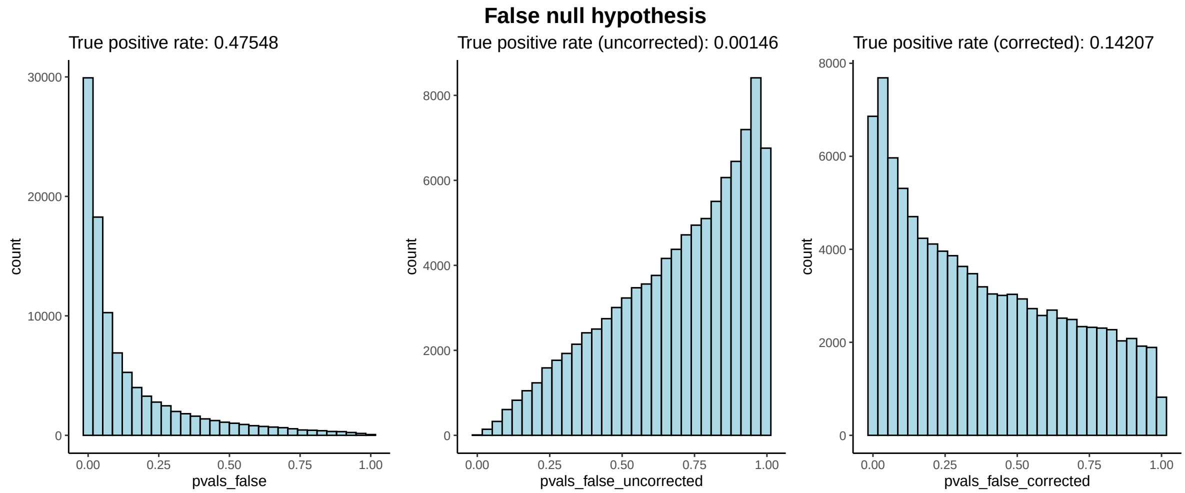
```
# Create histograms using ggplot
p1 <- ggplot(data.frame(pvals_true), aes(x = pvals_true)) +
  geom_histogram(bins = 30, color = "black", fill = "lightblue") +
  ggtitle(paste("False positive rate:", err_true_rate)) +
  theme_classic()

p2 <- ggplot(data.frame(pvals_true_uncorrected), aes(x = pvals_true_uncorrected)) +
  geom_histogram(bins = 30, color = "black", fill = "lightblue") +
  ggtitle(paste("False positive rate (uncorrected):", err_true_rate_uncorrected)) +
  theme_classic()

p3 <- ggplot(data.frame(pvals_true_corrected), aes(x = pvals_true_corrected)) +
  geom_histogram(bins = 30, color = "black", fill = "lightblue") +
  ggtitle(paste("False positive rate (corrected):", err_true_rate_corrected)) +
  theme_classic()

p4 <- ggplot(data.frame(pvals_false), aes(x = pvals_false)) +
  geom_histogram(bins = 30, color = "black", fill = "lightblue") +
  ggtitle(paste("True positive rate:", err_false_rate)) +
  theme_classic()

p5 <- ggplot(data.frame(pvals_false_uncorrected), aes(x = pvals_false_uncorrected)) +
  geom_histogram(bins = 30, color = "black", fill = "lightblue") +
  ggtitle(paste("True positive rate (uncorrected):", err_false_rate_uncorrected)) +
  theme_classic()

p6 <- ggplot(data.frame(pvals_false_corrected), aes(x = pvals_false_corrected)) +
  geom_histogram(bins = 30, color = "black", fill = "lightblue") +
  ggtitle(paste("True positive rate (corrected):", err_false_rate_corrected)) +
```

```
  theme_classic()
```

```
grid.arrange(
  p1, p2, p3,
  ncol = 3,
  nrow = 1,
  top = textGrob("True null hypothesis",
                 gp = gpar(fontsize = 16, fontface = "bold", col = "black"))
)
```




```
grid.arrange(
  p4, p5, p6,
  ncol = 3,
  nrow = 1,
  top = textGrob("False null hypothesis",
                 gp = gpar(fontsize = 16, fontface = "bold", col = "black"))
)
```


False null hypothesis
True positive rate: 0.47548
True positive rate (uncorrected): 0.00146
True positive rate (corrected): 0.14207
count
pvals_false
pvals_false_uncorrected
pvals_false_corrected