Testing Distributions: Kolmogorov-Smirnov

2024-05-03

We'll first use two-sample Kolmogorov-Smirnov tests to evaluate whether paired groups of EVI values are from the same distribution. The K-S test is a nonparametric test, which seems more appropriate in this case when we don't have normally distributed distributions. The null hypothesis of the K-S test is that the distributions are the same, and the D-statistic indicates the size of the difference between the two distributions, ranging from 0 (from the same distribution) to 1 (from very different distributions).

Another alternative test we can consider is the Mann-Whitney U Test, which is a nonparametric test to compare the medians of unpaired data, though median does not necessarily seem like the statistic we most care about for these EVI distributions.

Another approach would be to use permutation tests to compare the groups according to any statistic we specify - if there are any which would be most appropriate for vegetation indices, we can consider this.

```
plant_timing_data <- readRDS('plant_timing_data.Rds')


occlen_extremes_relig <- plant_timing_data |>
                      filter(!is.na(isis_occ_status)) |>
                      mutate(sunni_status=ifelse(sunni_dom==1,'Sunni Dominated',
                              ifelse(sunni_mix==1,'Sunni Mixed',
                                     ifelse(no_sunni==1,'No Sunni Pop',
                                            NA))),
                             occ_len_cat = ifelse(occ_length_mon<=12, "Short",
                                                  ifelse(occ_length_mon>=35,"Long",
                                                         "Mid"))) |>
                      mutate(occlen_and_relig = paste0(sunni_status," - ",occ_len_cat),
                             occlen_and_timing = paste0(occ_len_cat, " - " ,isis_occ_status))

occlen_extremes_relig_filt <- occlen_extremes_relig |>
                                filter(occ_len_cat %in% c('Long','Short')) |>
                                filter(isis_occ_status %in% c('Pre','Post'))


occ_vs_noocc_relig <- plant_timing_data |>
                       mutate(sunni_status=ifelse(sunni_dom==1,'Sunni Dominated',
                               ifelse(sunni_mix==1,'Sunni Mixed',
                                      ifelse(no_sunni==1,'No Sunni Pop',
                                             NA))),
                              occ_len_cat = ifelse(!is.na(isis_occ_status),
                                                   ifelse(occ_length_mon<=12, "Short",
                                                   ifelse(occ_length_mon>=35,"Long",
                                                          "Mid")),NA),
                              occ_status = ifelse(!is.na(isis_occ_status), "Occupied",
                                                    "Not Occupied"),
                              isis_presence = ifelse(is.na(isis_occ_status),
```

```
                                                    ifelse(year_num <= 2013, 'Pre',
                                                     ifelse(year_num >= 2018, 'Post',
                                                    ifelse((year_num > 2013 & year_num < 2018), 'During
                                                          ))),isis_occ_status)) |>
                          mutate(occ_and_relig = paste0(sunni_status," - ",occ_status),
                                 occ_and_timing = paste0(occ_status, " - " ,isis_presence))


occ_vs_noocc_relig_filt <- occ_vs_noocc_relig |>
                           filter(isis_presence %in% c('Pre','Post'))
```

```
seasonal_data <- readRDS('seasonal_data.Rds')

seasonal_data_occ_relig <- seasonal_data |>
                      filter(!is.na(isis_occ_status)) |>
                      mutate(sunni_status=ifelse(sunni_dom==1,'Sunni Dominated',
                              ifelse(sunni_mix==1,'Sunni Mixed',
                                     ifelse(no_sunni==1,'No Sunni Pop',
                                            NA)))) |>
                      mutate(occ_and_relig = paste0(sunni_status," - ",isis_occ_status))

seasonal_occ_vs_noocc_relig <- seasonal_data |>
                        mutate(sunni_status=ifelse(sunni_dom==1,'Sunni Dominated',
                                ifelse(sunni_mix==1,'Sunni Mixed',
                                       ifelse(no_sunni==1,'No Sunni Pop',
                                              NA))),
                               occ_status = ifelse(!is.na(isis_occ_status), "Occupied",
                                                   "Not Occupied"),
                               isis_presence = ifelse(is.na(isis_occ_status),
                                                    ifelse(year_num <= 2013, 'Pre',
                                                     ifelse(year_num >= 2018, 'Post',
                                                    ifelse((year_num > 2013 & year_num < 2018), 'During
                                                          ))),isis_occ_status)) |>
                      mutate(occ_and_relig = paste0(sunni_status," - ",occ_status),
                             occ_and_timing = paste0(occ_status, " - " ,isis_presence))


seasonal_occ_vs_noocc_relig_filt <- seasonal_occ_vs_noocc_relig |>
                           filter(isis_presence %in% c('Pre','Post'))
```

For data by plant seasons (Barley-Wheat growing seasons of December - April, Summer crop growing season of May - September), create vectors of the mean and max residual EVI values matching a variety of designations of interest:

```
# Occupied vs Non-Occupied Mean Residuals
occ_mean_resids <- filter(occ_vs_noocc_relig, occ_status=='Occupied')$mean_evi_resids
occ_mean_resids <- occ_mean_resids[!is.na(occ_mean_resids)]
noocc_mean_resids <- filter(occ_vs_noocc_relig, occ_status=='Not Occupied')$mean_evi_resids
noocc_mean_resids <- noocc_mean_resids[!is.na(noocc_mean_resids)]

# Occupied vs Non-Occupied Max Residuals
occ_max_resids <- filter(occ_vs_noocc_relig, occ_status=='Occupied')$max_evi_resids
occ_max_resids <- occ_max_resids[!is.na(occ_max_resids)]
```

```
noocc_max_resids <- filter(occ_vs_noocc_relig, occ_status=='Not Occupied')$max_evi_resids
noocc_max_resids <- noocc_max_resids[!is.na(noocc_max_resids)]


# Occupied vs Non-Occupied Pre-ISIS Mean Residuals
occ_pre_mean_resids <- filter(occ_vs_noocc_relig, (occ_status=='Occupied'&isis_presence=='Pre'))$mean_ev
occ_pre_mean_resids <- occ_pre_mean_resids[!is.na(occ_pre_mean_resids)]
noocc_pre_mean_resids <- filter(occ_vs_noocc_relig, (occ_status=='Not Occupied'&isis_presence=='Pre'))$m
noocc_pre_mean_resids <- noocc_pre_mean_resids[!is.na(noocc_pre_mean_resids)]

# Occupied vs Non-Occupied Post-ISIS Mean Residuals
occ_post_mean_resids <- filter(occ_vs_noocc_relig, (occ_status=='Occupied'&isis_presence=='Post'))$mean_
occ_post_mean_resids <- occ_post_mean_resids[!is.na(occ_post_mean_resids)]
noocc_post_mean_resids <- filter(occ_vs_noocc_relig, (occ_status=='Not Occupied'&isis_presence=='Post'))
noocc_post_mean_resids <- noocc_post_mean_resids[!is.na(noocc_post_mean_resids)]

# Occupied vs Non-Occupied Pre-ISIS Max Residuals
occ_pre_max_resids <- filter(occ_vs_noocc_relig, (occ_status=='Occupied'&isis_presence=='Pre'))$max_evi_
occ_pre_max_resids <- occ_pre_max_resids[!is.na(occ_pre_max_resids)]
noocc_pre_max_resids <- filter(occ_vs_noocc_relig, (occ_status=='Not Occupied'&isis_presence=='Pre'))$ma
noocc_pre_max_resids <- noocc_pre_max_resids[!is.na(noocc_pre_max_resids)]

# Occupied vs Non-Occupied Post-ISIS Max Residuals
occ_post_max_resids <- filter(occ_vs_noocc_relig, (occ_status=='Occupied'&isis_presence=='Post'))$max_ev
occ_post_max_resids <- occ_post_max_resids[!is.na(occ_post_max_resids)]
noocc_post_max_resids <- filter(occ_vs_noocc_relig, (occ_status=='Not Occupied'&isis_presence=='Post'))
noocc_post_max_resids <- noocc_post_max_resids[!is.na(noocc_post_max_resids)]


# Long vs Short Occupation Pre-ISIS Mean Residuals
long_occ_pre_mean_resids <- filter(occ_vs_noocc_relig, (occ_len_cat=='Long'&isis_presence=='Pre'))$mean_
long_occ_pre_mean_resids <- long_occ_pre_mean_resids[!is.na(long_occ_pre_mean_resids)]
short_occ_pre_mean_resids <- filter(occ_vs_noocc_relig, (occ_len_cat=='Short'&isis_presence=='Pre'))$mea
short_occ_pre_mean_resids <- short_occ_pre_mean_resids[!is.na(short_occ_pre_mean_resids)]

# Long vs Short Occupation Post-ISIS Mean Residuals
long_occ_post_mean_resids <- filter(occ_vs_noocc_relig, (occ_len_cat=='Long'&isis_presence=='Post'))$mea
long_occ_post_mean_resids <- long_occ_post_mean_resids[!is.na(long_occ_post_mean_resids)]
short_occ_post_mean_resids <- filter(occ_vs_noocc_relig, (occ_len_cat=='Short'&isis_presence=='Post'))$m
short_occ_post_mean_resids <- short_occ_post_mean_resids[!is.na(short_occ_post_mean_resids)]

# Long vs Short Occupation Pre-ISIS Max Residuals
long_occ_pre_max_resids <- filter(occ_vs_noocc_relig, (occ_len_cat=='Long'&isis_presence=='Pre'))$max_ev
long_occ_pre_max_resids <- long_occ_pre_max_resids[!is.na(long_occ_pre_max_resids)]
short_occ_pre_max_resids <- filter(occ_vs_noocc_relig, (occ_len_cat=='Short'&isis_presence=='Pre'))$max_
short_occ_pre_max_resids <- short_occ_pre_max_resids[!is.na(short_occ_pre_max_resids)]

# Long vs Short Occupation Post-ISIS Max Residuals
long_occ_post_max_resids <- filter(occ_vs_noocc_relig, (occ_len_cat=='Long'&isis_presence=='Post'))$max_
long_occ_post_max_resids <- long_occ_post_max_resids[!is.na(long_occ_post_max_resids)]
short_occ_post_max_resids <- filter(occ_vs_noocc_relig, (occ_len_cat=='Short'&isis_presence=='Post'))$ma
short_occ_post_max_resids <- short_occ_post_max_resids[!is.na(short_occ_post_max_resids)]
```

We can also ensure that the vector lengths match the count breakdowns for each type of category combination:

```
occ_df <- occ_vs_noocc_relig |>
          filter(!is.na(mean_evi_resids)) |>
          select(c('occ_len_cat','isis_presence','occ_status')) |>
          filter(occ_status=='Occupied')
table(occ_df)
```

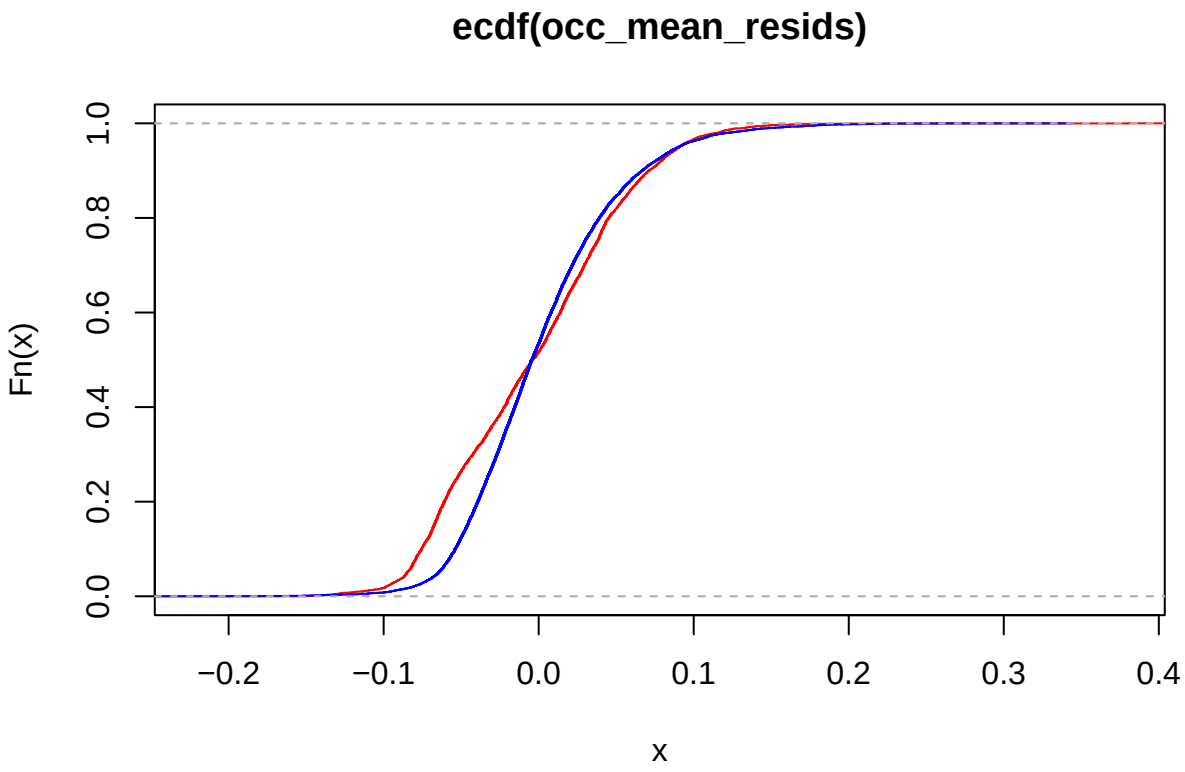
```
## , , occ_status = Occupied
##
##            isis_presence
## occ_len_cat During Post  Pre
##       Long     178  186  572
##       Mid      219  464 1191
##       Short     67  499  964
```

```
noocc_df <- occ_vs_noocc_relig |>
          filter(!is.na(mean_evi_resids)) |>
          select(c('occ_status','isis_presence'))
table(noocc_df)
```

```
##               isis_presence
## occ_status     During Post  Pre
##   Not Occupied   2172 2231 6905
##   Occupied        464 1149 2727
```

K-S test to compare mean residual EVI for occupied and non-occupied areas over the entire time period. Relatively small difference identified, but statistically significant.

```
plot(ecdf(occ_mean_resids),
     xlim=range(c(occ_mean_resids, noocc_mean_resids)),
     col='red')
plot(ecdf(noocc_mean_resids),
     add=TRUE,
     col='blue')
```

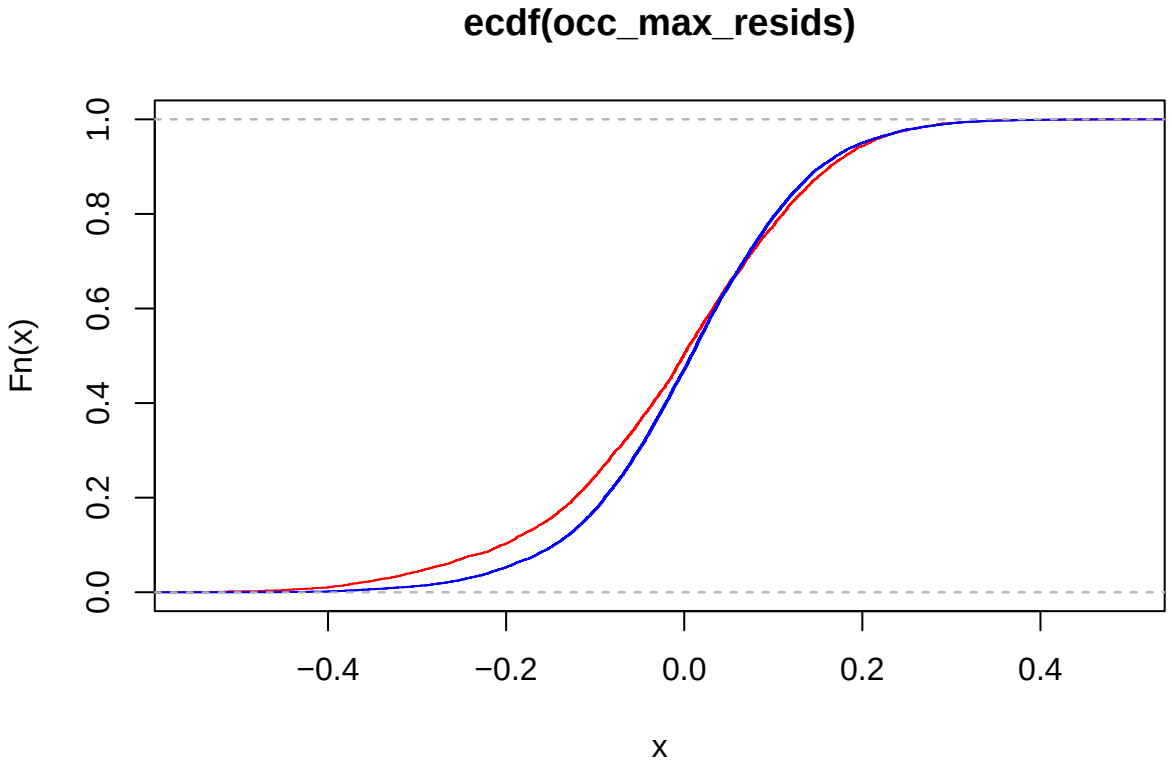



```
ks.test(occ_mean_resids, noocc_mean_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  occ_mean_resids and noocc_mean_resids
## D = 0.14538, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

K-S test to compare max residual EVI for occupied and non-occupied areas over the entire time period. Relatively small difference identified, but statistically significant.

```
plot(ecdf(occ_max_resids),
     xlim=range(c(occ_max_resids, noocc_max_resids)),
     col='red')
plot(ecdf(noocc_max_resids),
     add=TRUE,
     col='blue')
```

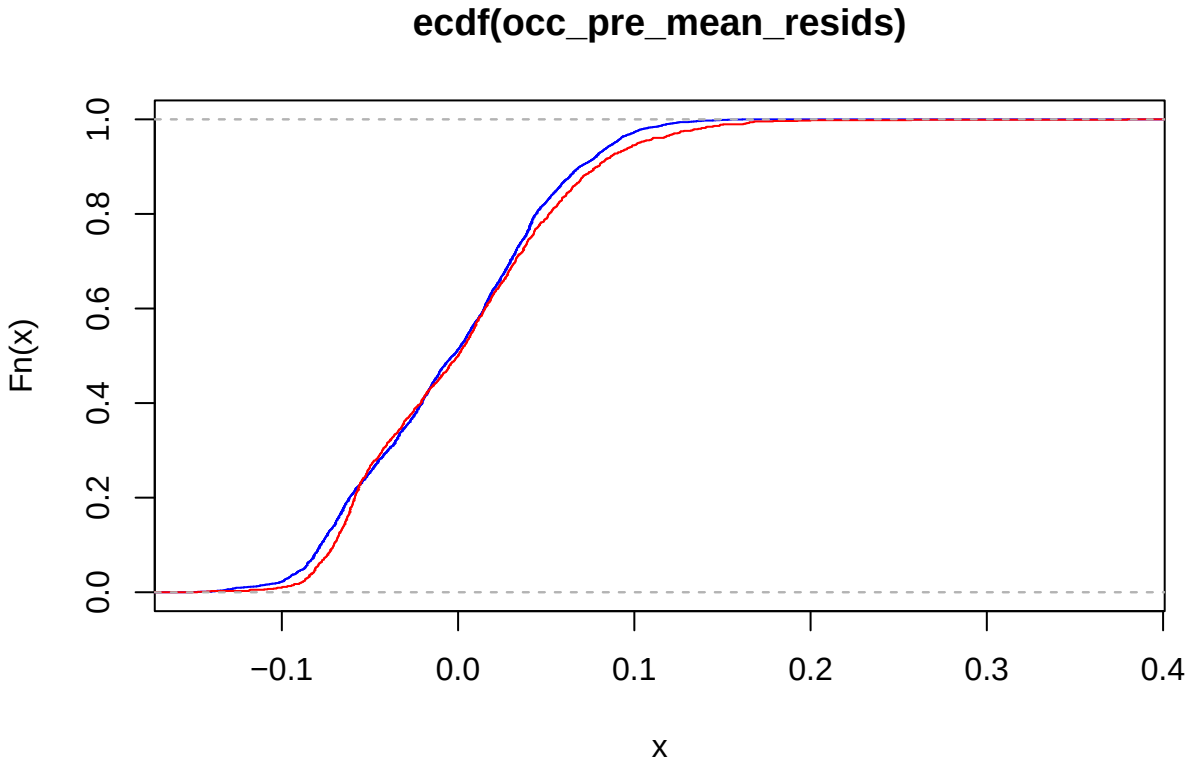



```
ks.test(occ_max_resids, noocc_max_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  occ_max_resids and noocc_max_resids
## D = 0.071713, p-value = 1.954e-14
## alternative hypothesis: two-sided
```

K-S test to compare mean residual EVI for occupied areas pre- and post-ISIS occupation. Very small difference identified, statistically significant at the 0.1 level but not the 0.05 level.

```
plot(ecdf(occ_pre_mean_resids),
     xlim=range(c(occ_pre_mean_resids, occ_post_mean_resids)),
     col='blue')
plot(ecdf(occ_post_mean_resids),
     add=TRUE,
     col='red')
```

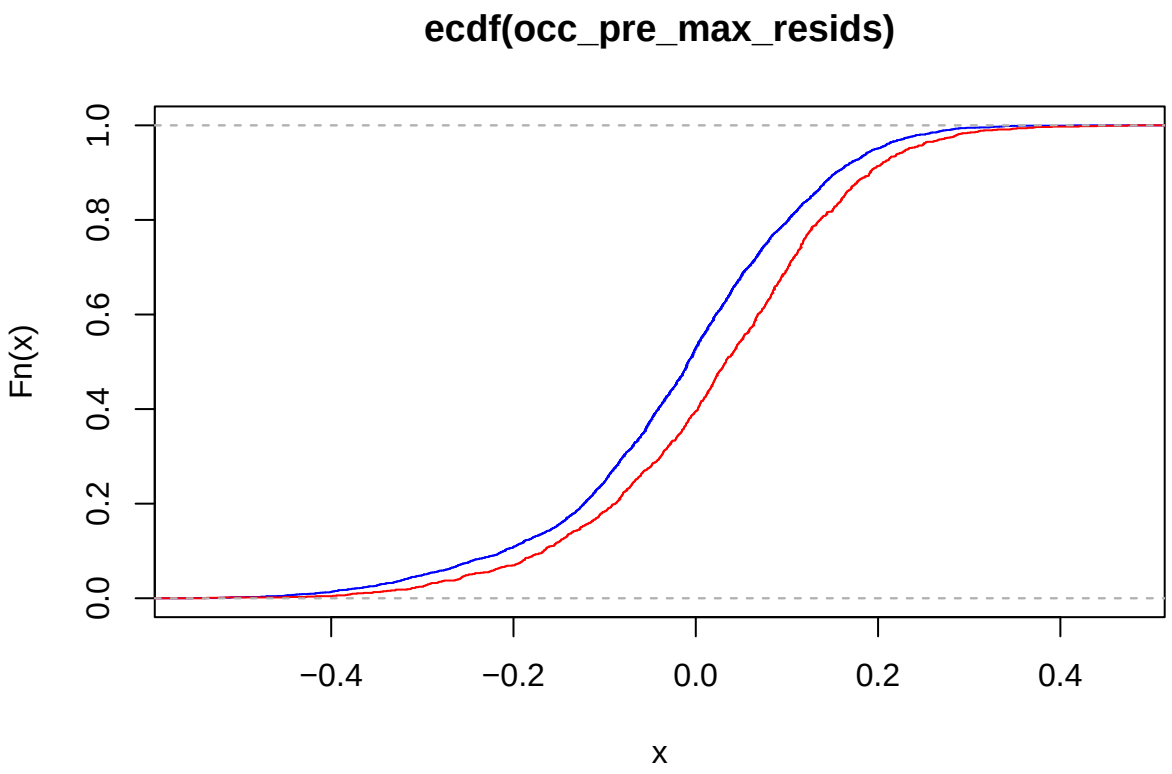



```
ks.test(occ_pre_mean_resids, occ_post_mean_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  occ_pre_mean_resids and occ_post_mean_resids
## D = 0.045988, p-value = 0.06546
## alternative hypothesis: two-sided
```

K-S test to compare max residual EVI for occupied areas pre- and post-ISIS occupation. Relatively small difference identified, statistically significant.

```
plot(ecdf(occ_pre_max_resids),
     xlim=range(c(occ_pre_max_resids, occ_post_max_resids)),
     col='blue')
plot(ecdf(occ_post_max_resids),
     add=TRUE,
     col='red')
```

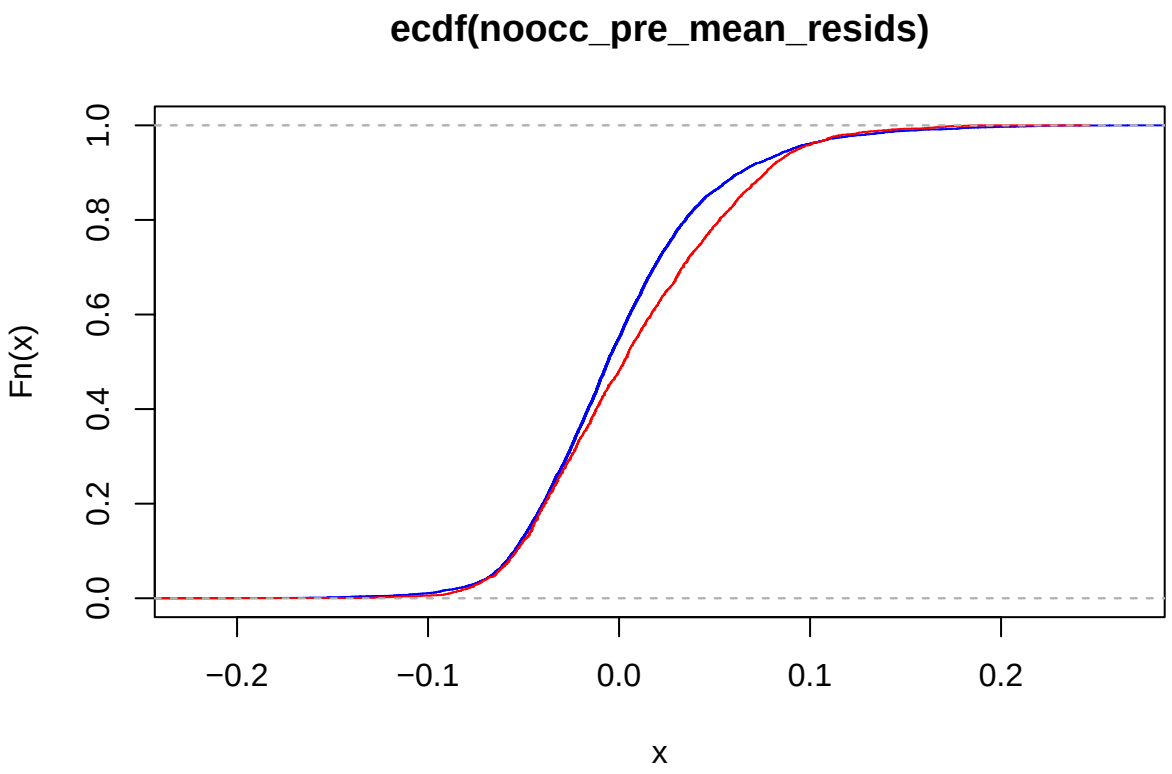



```
ks.test(occ_pre_max_resids, occ_post_max_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  occ_pre_max_resids and occ_post_max_resids
## D = 0.13991, p-value = 3.608e-14
## alternative hypothesis: two-sided
```

K-S test to compare mean residual EVI for non-occupied areas pre- and post-ISIS presence. Relatively small difference identified, statistically significant.

```
plot(ecdf(noocc_pre_mean_resids),
     xlim=range(c(noocc_pre_mean_resids, noocc_post_mean_resids)),
     col='blue')
plot(ecdf(noocc_post_mean_resids),
     add=TRUE,
     col='red')
```

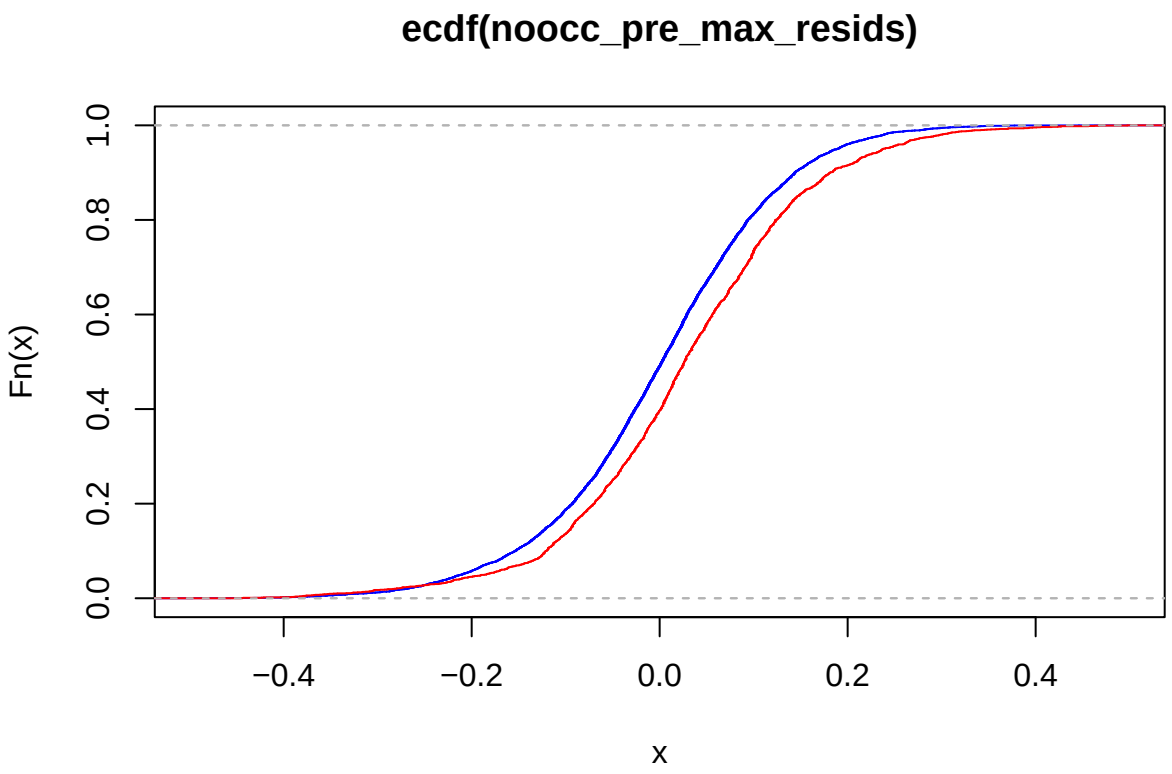



```
ks.test(noocc_pre_mean_resids, noocc_post_mean_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  noocc_pre_mean_resids and noocc_post_mean_resids
## D = 0.10171, p-value = 1.443e-15
## alternative hypothesis: two-sided
```

K-S test to compare max residual EVI for non-occupied areas pre- and post-ISIS presence. Relatively small difference identified, statistically significant.

```
plot(ecdf(noocc_pre_max_resids),
     xlim=range(c(noocc_pre_max_resids, noocc_post_max_resids)),
     col='blue')
plot(ecdf(noocc_post_max_resids),
     add=TRUE,
     col='red')
```

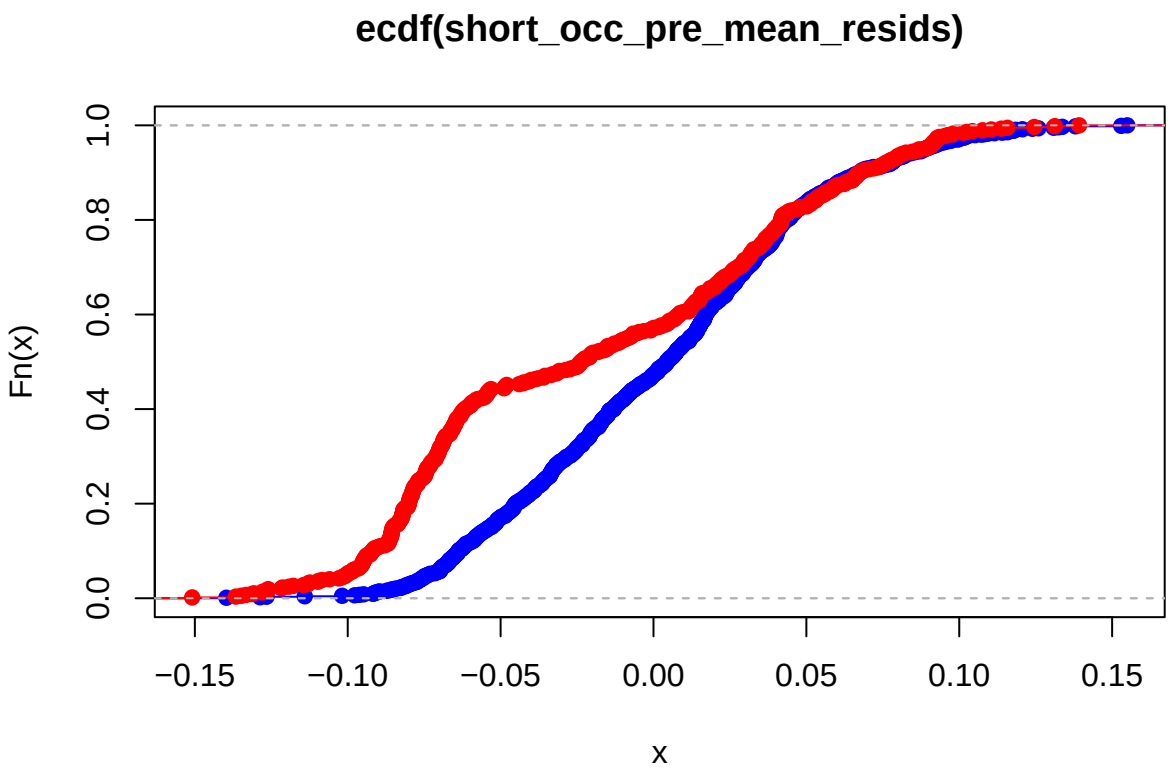



```
ks.test(noocc_pre_max_resids, noocc_post_max_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  noocc_pre_max_resids and noocc_post_max_resids
## D = 0.097907, p-value = 1.832e-14
## alternative hypothesis: two-sided
```

K-S test to compare mean residual EVI for ISIS-occupied areas with long vs short occupation periods (short being a year or less, long being 35 months or more) in the pre-ISIS period. Moderate difference identified, statistically significant.

```
plot(ecdf(short_occ_pre_mean_resids),
     xlim=range(c(short_occ_pre_mean_resids, long_occ_pre_mean_resids)),
     col='blue')
plot(ecdf(long_occ_pre_mean_resids),
     add=TRUE,
     col='red')
```

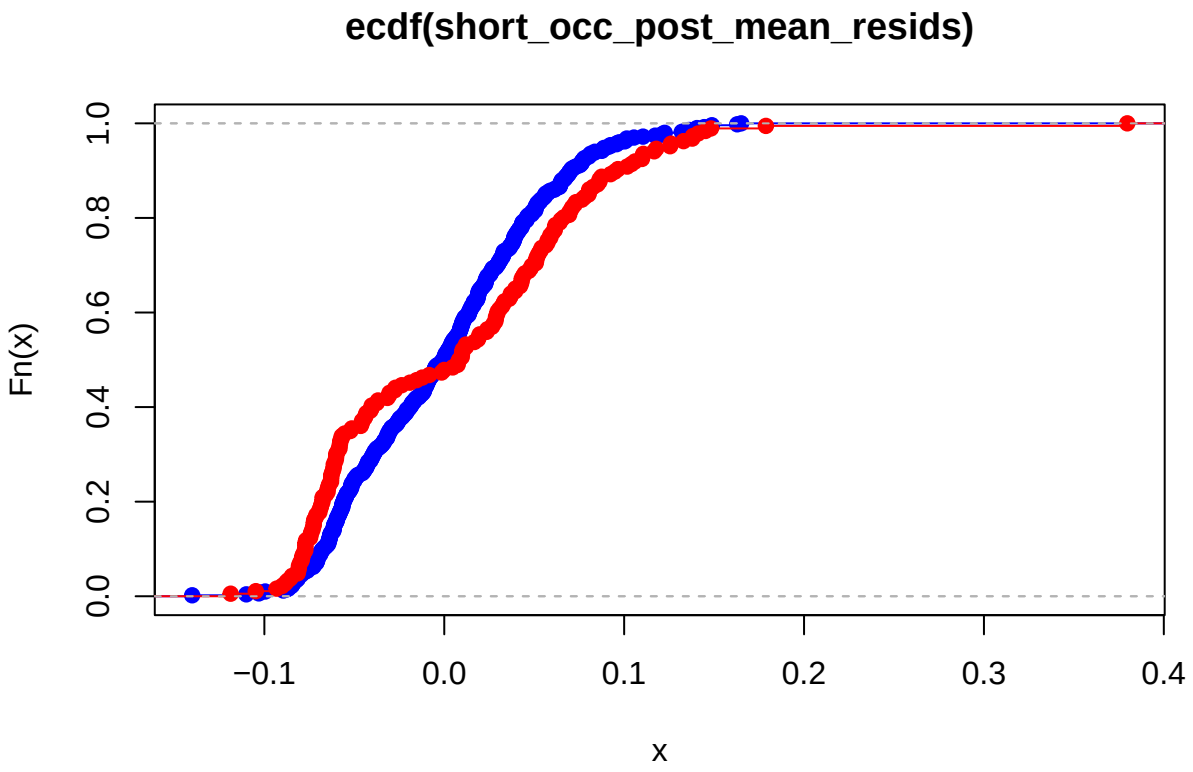



```
ks.test(short_occ_pre_mean_resids, long_occ_pre_mean_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  short_occ_pre_mean_resids and long_occ_pre_mean_resids
## D = 0.294, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

K-S test to compare mean residual EVI for ISIS-occupied areas with long vs short occupation periods (short being a year or less, long being 35 months or more) in the post-ISIS period. Relatively small difference identified, statistically significant.

```
plot(ecdf(short_occ_post_mean_resids),
     xlim=range(c(short_occ_post_mean_resids, long_occ_post_mean_resids)),
     col='blue')
plot(ecdf(long_occ_post_mean_resids),
     add=TRUE,
     col='red')
```

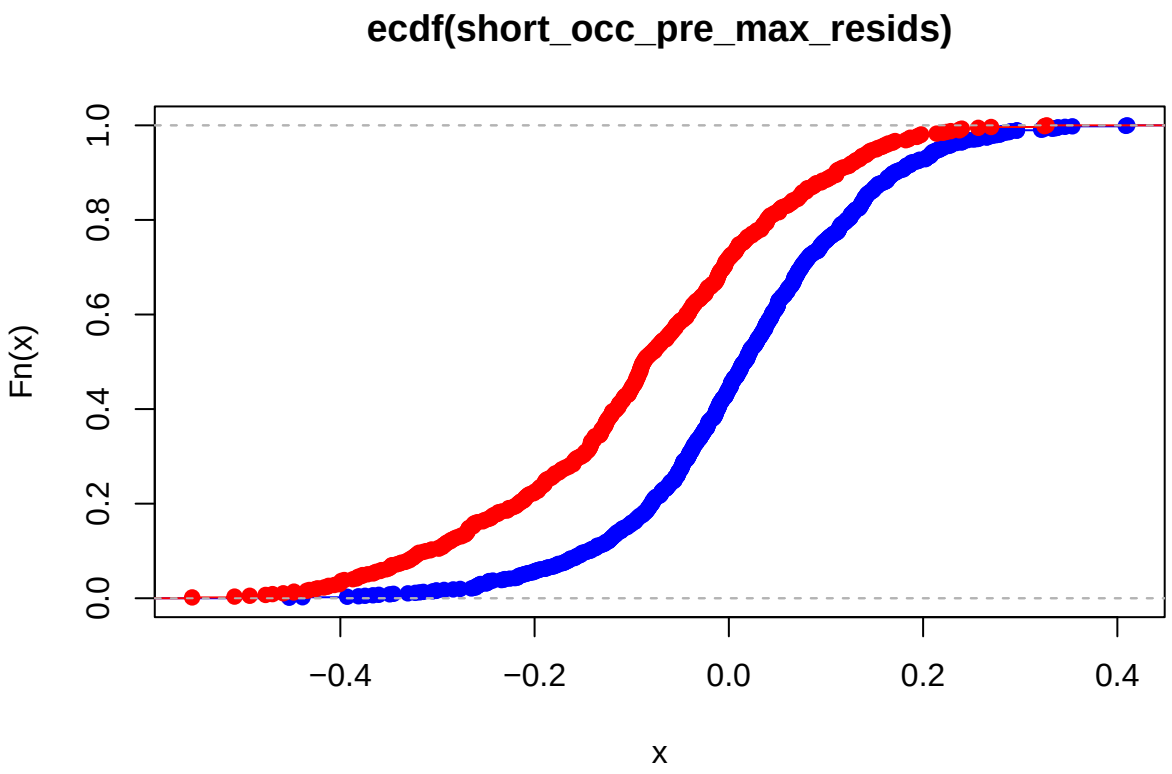



```
ks.test(short_occ_post_mean_resids, long_occ_post_mean_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  short_occ_post_mean_resids and long_occ_post_mean_resids
## D = 0.15234, p-value = 0.003714
## alternative hypothesis: two-sided
```

K-S test to compare max residual EVI for ISIS-occupied areas with long vs short occupation periods (short being a year or less, long being 35 months or more) in the pre-ISIS period. Moderate difference identified, statistically significant.

```
plot(ecdf(short_occ_pre_max_resids),
     xlim=range(c(short_occ_pre_max_resids, long_occ_pre_max_resids)),
     col='blue')
plot(ecdf(long_occ_pre_max_resids),
     add=TRUE,
     col='red')
```

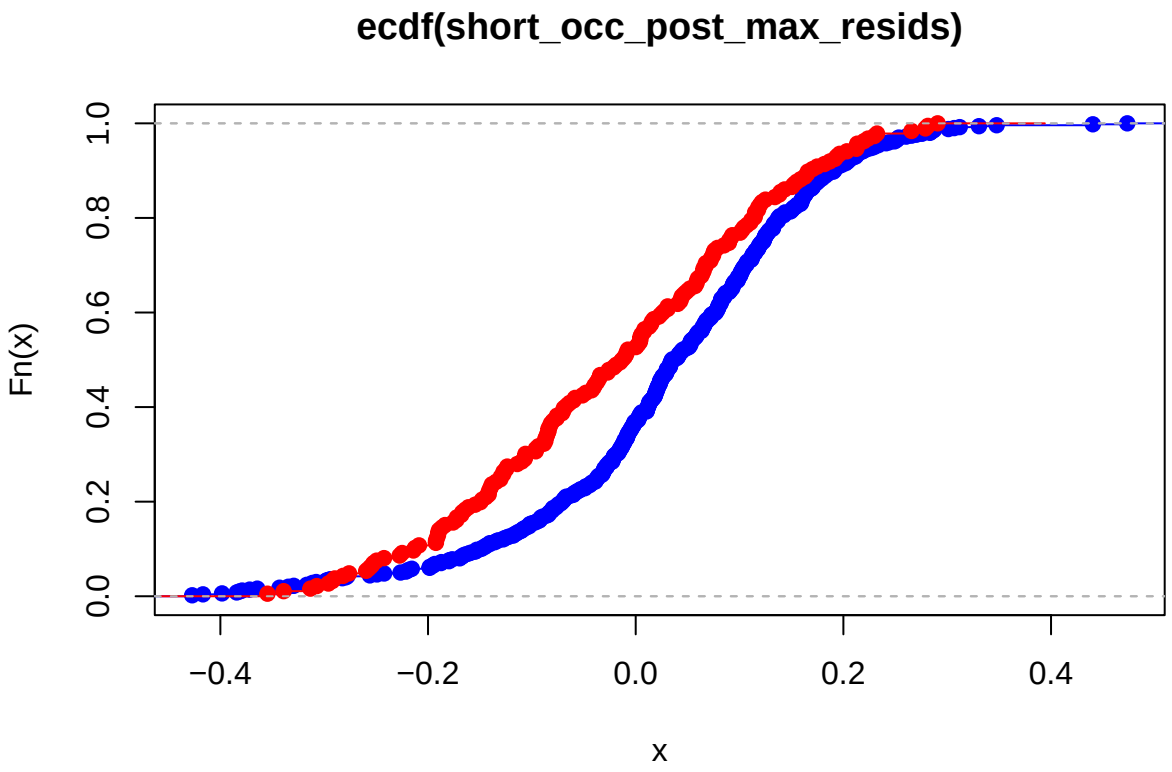



```
ks.test(short_occ_pre_max_resids, long_occ_pre_max_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  short_occ_pre_max_resids and long_occ_pre_max_resids
## D = 0.32546, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

K-S test to compare max residual EVI for ISIS-occupied areas with long vs short occupation periods (short being a year or less, long being 35 months or more) in the post-ISIS period. Moderate difference identified, statistically significant.

```
plot(ecdf(short_occ_post_max_resids),
     xlim=range(c(short_occ_post_max_resids, long_occ_post_max_resids)),
     col='blue')
plot(ecdf(long_occ_post_max_resids),
     add=TRUE,
     col='red')
```

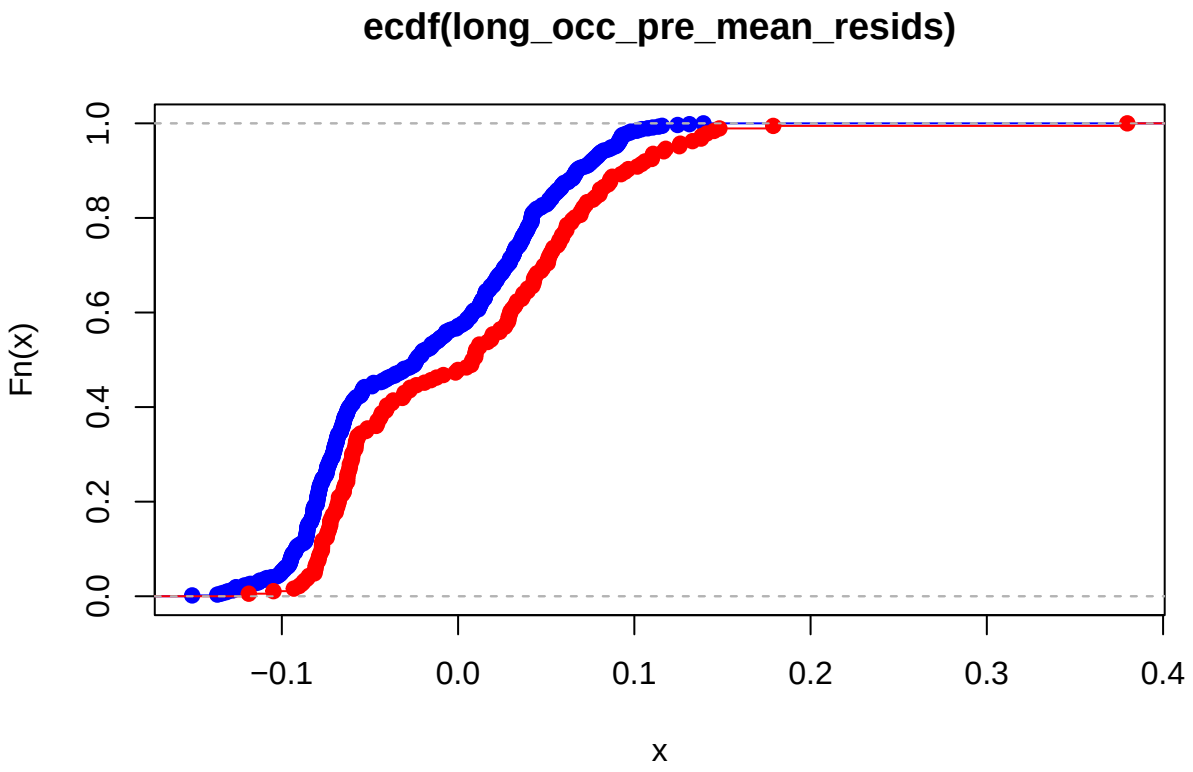



```
ks.test(short_occ_post_max_resids, long_occ_post_max_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  short_occ_post_max_resids and long_occ_post_max_resids
## D = 0.21323, p-value = 8.91e-06
## alternative hypothesis: two-sided
```

K-S test to compare mean residual EVI for long-ISIS-occupation areas (long being 35 months or more) pre- and post-ISIS occupation. Relatively small difference identified, statistically significant.

```
plot(ecdf(long_occ_pre_mean_resids),
     xlim=range(c(long_occ_pre_mean_resids, long_occ_post_mean_resids)),
     col='blue')
plot(ecdf(long_occ_post_mean_resids),
     add=TRUE,
     col='red')
```

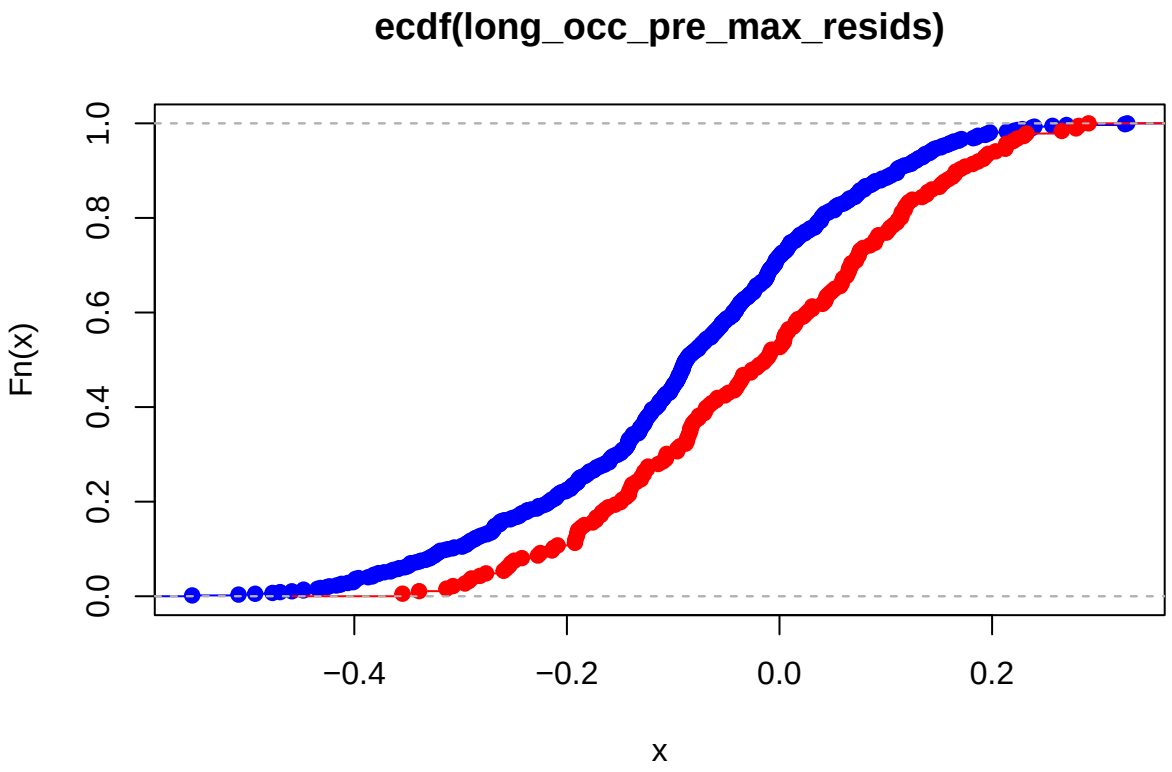



```
ks.test(long_occ_pre_mean_resids, long_occ_post_mean_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  long_occ_pre_mean_resids and long_occ_post_mean_resids
## D = 0.15715, p-value = 0.00195
## alternative hypothesis: two-sided
```

K-S test to compare max residual EVI for long-ISIS-occupation areas (long being 35 months or more) pre- and post-ISIS occupation. Moderate difference identified, statistically significant.

```
plot(ecdf(long_occ_pre_max_resids),
     xlim=range(c(long_occ_pre_max_resids, long_occ_post_max_resids)),
     col='blue')
plot(ecdf(long_occ_post_max_resids),
     add=TRUE,
     col='red')
```

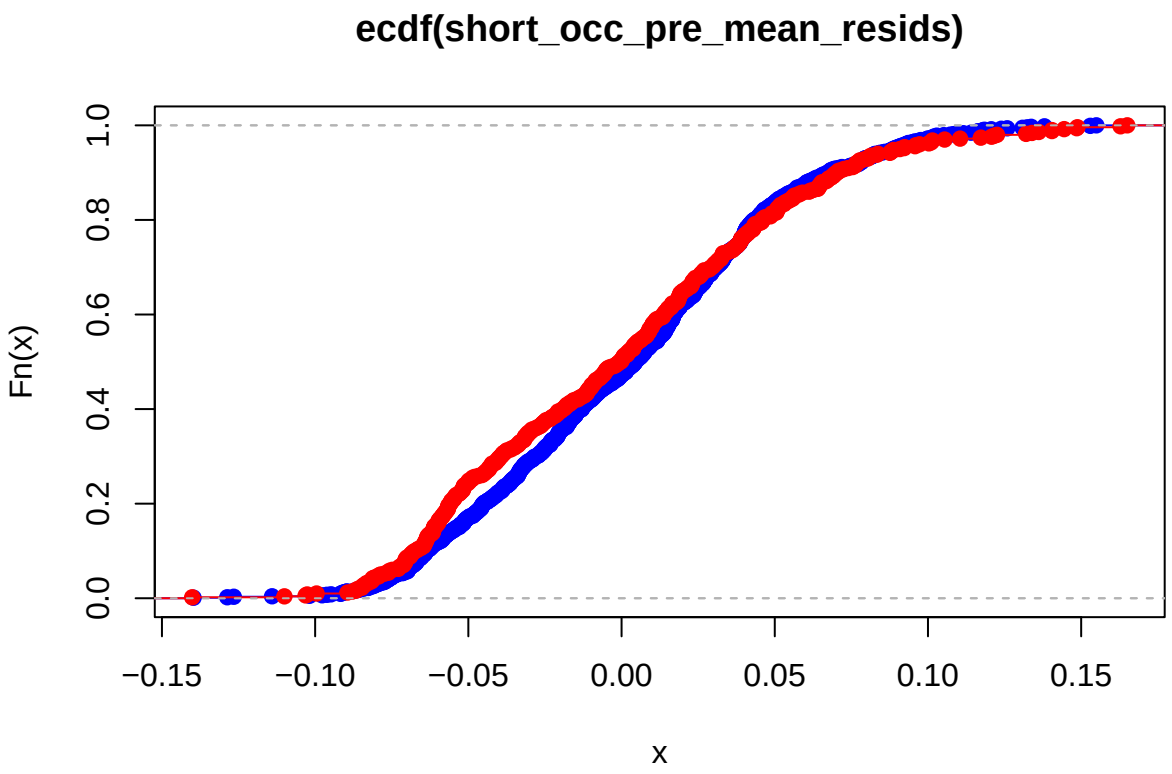



```
ks.test(long_occ_pre_max_resids, long_occ_post_max_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  long_occ_pre_max_resids and long_occ_post_max_resids
## D = 0.19877, p-value = 3.048e-05
## alternative hypothesis: two-sided
```

K-S test to compare mean residual EVI for short-ISIS-occupation areas (short being a year or less) pre- and post-ISIS occupation. Small difference identified, statistically significant at the 0.05 level.

```
plot(ecdf(short_occ_pre_mean_resids),
     xlim=range(c(short_occ_pre_mean_resids, short_occ_post_mean_resids)),
     col='blue')
plot(ecdf(short_occ_post_mean_resids),
     add=TRUE,
     col='red')
```

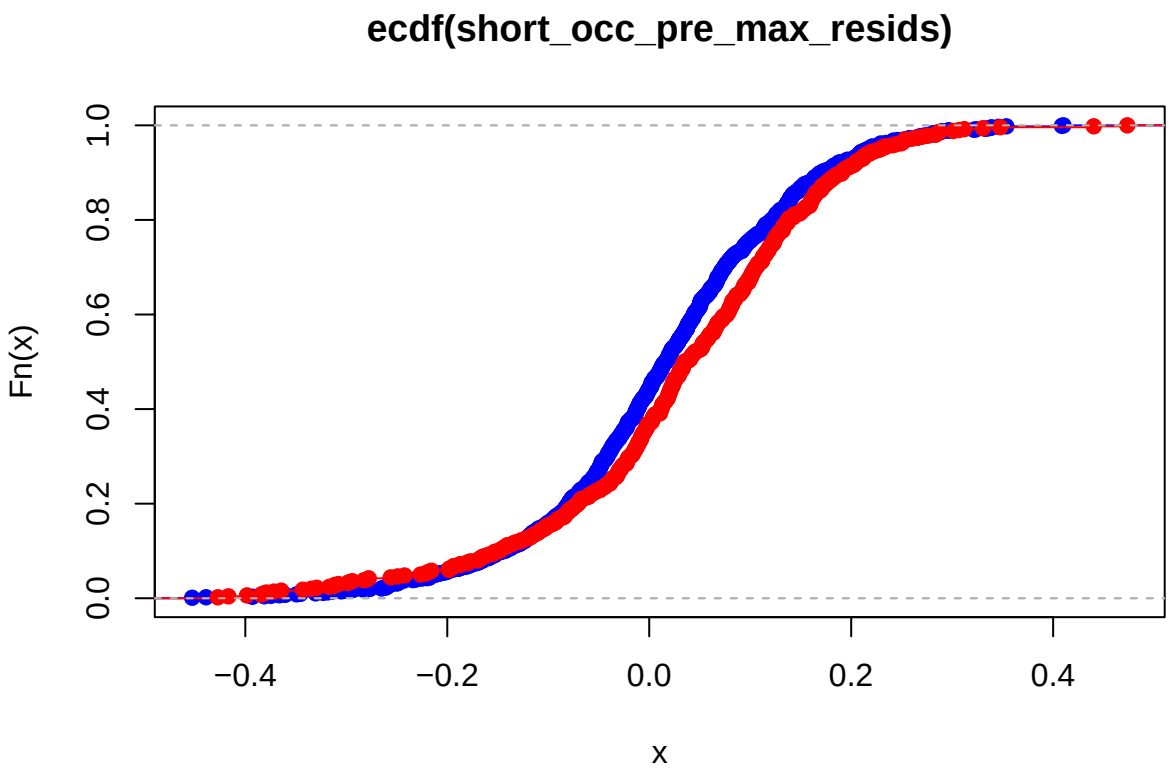



```
ks.test(short_occ_pre_mean_resids, short_occ_post_mean_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  short_occ_pre_mean_resids and short_occ_post_mean_resids
## D = 0.081273, p-value = 0.02598
## alternative hypothesis: two-sided
```

K-S test to compare max residual EVI for short-ISIS-occupation areas (short being a year or less) pre- and post-ISIS occupation. Relatively small difference identified, statistically significant.

```
plot(ecdf(short_occ_pre_max_resids),
     xlim=range(c(short_occ_pre_max_resids, short_occ_post_max_resids)),
     col='blue')
plot(ecdf(short_occ_post_max_resids),
     add=TRUE,
     col='red')
```




```
ks.test(short_occ_pre_max_resids, short_occ_post_max_resids)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  short_occ_pre_max_resids and short_occ_post_max_resids
## D = 0.11027, p-value = 0.0006731
## alternative hypothesis: two-sided
```