```
import pandas as pd
from scipy.stats import kendalltau


years_ca = [1991, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001,
2002, 2003, 2004,
            2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013,
2014, 2015, 2016,
            2017, 2018, 2019, 2020, 2021, 2022, 2023]
male_counts_ca = [2, 1, 1, 2, 1, 8, 8, 21, 18, 21, 44, 76, 92, 131,
122, 166, 186, 253, 237, 289,
                  281, 346, 373, 466, 421, 498, 444, 596, 690, 772,
650]
female_counts_ca = [0, 0, 1, 0, 0, 2, 3, 1, 7, 11, 11, 17, 28, 33, 58,
52, 76, 85, 107, 110,
                    148, 166, 165, 206, 223, 263, 280, 316, 381, 412,
440]
gender_gap_ca = [f - m for m, f in zip(male_counts_ca,
female_counts_ca)]

# FA Dataset
years_fa = [1991, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002,
2003, 2004,
            2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013,
2014, 2015,
            2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023]
male_counts_fa = [1, 2, 2, 2, 1, 6, 10, 9, 9, 25, 32, 43, 90, 104,
125, 161, 215, 216, 242,
                  245, 301, 298, 322, 297, 326, 308, 402, 461, 483,
453]
female_counts_fa = [1, 0, 1, 2, 0, 4, 2, 2, 9, 7, 17, 25, 57, 74, 98,
99, 132, 134, 172, 188,
                    210, 234, 281, 270, 339, 318, 369, 438, 436, 460]
gender_gap_fa = [m-f for m, f in zip(male_counts_fa,
female_counts_fa)]

# Mann-Kendall Test
mk_test_result_ca = kendalltau(years_ca, gender_gap_ca)
mk_test_result_fa = kendalltau(years_fa, gender_gap_fa)

# Results
print("CA Dataset Mann-Kendall Test Result:", mk_test_result_ca)
print("FA Dataset Mann-Kendall Test Result:", mk_test_result_fa)
```

```
CA Dataset Mann-Kendall Test Result: SignificanceResult(statistic=-
0.8716334749099672, pvalue=6.4413867693831084e-12)
FA Dataset Mann-Kendall Test Result:
SignificanceResult(statistic=0.313238548254737,
pvalue=0.015871017738004995)
```

```
from scipy.stats import chi2_contingency

# Observed frequencies
data = {
    "China": [1747, 839],
    "USA": [1663, 715],
    "Germany": [506, 203],
    "Japan": [442, 73],
    "England": [256, 124]
}

# Perform chi-squared test for each country
results = {}
for country, frequencies in data.items():
    chi2, p_value, _, _ = chi2_contingency([frequencies,
[sum(frequencies) - f for f in frequencies]])
    results[country] = {'Chi-squared': chi2, 'p-value': p_value}

results
'''This test helps us determine if there is a significant difference
between the number of males and females in each country'''
```

```
{'China': {'Chi-squared': 636.2327919566899,
  'p-value': 2.204229036293869e-140},
 'USA': {'Chi-squared': 754.2548359966358, 'p-value':
4.7670587255318835e-166},
 'Germany': {'Chi-squared': 257.2750352609309,
  'p-value': 6.737599742675509e-58},
 'Japan': {'Chi-squared': 525.9184466019417,
  'p-value': 2.1823244606041034e-116},
 'England': {'Chi-squared': 90.32105263157895,
  'p-value': 2.0248655429940715e-21}}
```

```
import pandas as pd

# Load the Excel file
file_path = 'Publications_Forecast.xlsx'
data = pd.read_excel(file_path)

# Display the first few rows of the dataframe to understand its
structure
data.head()
# Load the Excel file and list the sheet names
sheets = pd.ExcelFile(file_path)
sheet_names = sheets.sheet_names

# Read and preview the first few rows of each sheet
sheets_data = {}
for sheet in sheet_names:
    sheets_data[sheet] = pd.read_excel(file_path,
```

```
sheet_name=sheet).head()

sheets_data, sheet_names
from scipy.stats import ttest_rel

# Read complete data from each sheet
overall_data = pd.read_excel(file_path, sheet_name='Sheet1')
ca_data = pd.read_excel(file_path, sheet_name='CA')
fa_data = pd.read_excel(file_path, sheet_name='FA')

# Calculate t-test for each dataset
def perform_ttest(data, actual_col, forecast_col):
    actual = data[actual_col]
    forecasted = data[forecast_col]
    t_stat, p_value = ttest_rel(actual, forecasted)
    return t_stat, p_value

# Overall data
overall_t_stat, overall_p_value = perform_ttest(overall_data,
'Actual', 'Adjusted Forecast')

# Corresponding author data
ca_t_stat, ca_p_value = perform_ttest(ca_data, 'Actual CA
Publications', 'Forecasted CA Publications')

# First author data
fa_t_stat, fa_p_value = perform_ttest(fa_data, 'Actual FA
Publications', 'Forecasted FA Publications')

overall_t_stat, overall_p_value, ca_t_stat, ca_p_value, fa_t_stat,
fa_p_value
```

```
(-0.21525973209006324,
 0.8310215675387781,
 0.4000006666683333,
 0.6919900307504462,
 -0.004371903516951845,
 0.9965416732635004)
```

```
import pandas as pd

# Load the Excel files
data_ca = pd.read_excel("HCGenderGapCA.xlsx")
data_fa = pd.read_excel("HCGenderGapFA.xlsx")

# Display the first few rows of each dataset to understand their
structure
data_ca.head(), data_fa.head()
```

```
(   Publication Year  female  male  Gender Gap
 0              1991       0     2           2
```

```
 1                1994       0     1           1
 2                1995       1     1           0
 3                1996       0     2           2
 4                1997       0     1           1,
    Publication Year  female  male  gender gap
 0                1991       1     1           0
 1                1995       0     2           2
 2                1996       1     2           1
 3                1997       1     2           1
 4                1998       0     1           1)
```

```
from scipy.stats import kendalltau

# Extract columns for Mann-Kendall Test
years_ca = data_ca['Publication Year']
gender_gap_ca = data_ca['Gender Gap']

years_fa = data_fa['Publication Year']
gender_gap_fa = data_fa['gender gap']  # Note different column name
casing

# Perform the Mann-Kendall Test
mk_test_result_ca = kendalltau(years_ca, gender_gap_ca)
mk_test_result_fa = kendalltau(years_fa, gender_gap_fa)

mk_test_result_ca, mk_test_result_fa
```

```
(SignificanceResult(statistic=0.8728468546741887,
pvalue=5.7569446585837285e-12),
 SignificanceResult(statistic=0.2519121366275708,
pvalue=0.04832109670847903))
```