Categorical variable correlation tests

September 15, 2022

[1]:
```
catalog
```

[1]:
```
<kedro.io.data_catalog.DataCatalog at 0x1dda8476f70>
```

[2]:
```
catalog.list()
```

[2]:
```
['Research data Mongolian',
 'Research data English',
 'Patient',
 'Diagnosis',
 'External beam radiotherapy',
 'Brachytherapy',
 'Chemotherapy',
 'Acute toxicity',
 'Response status',
 'Late morbidity',
 'Disease status',
 'parameters']
```

[3]:
```
df = catalog.load('Research data English')
```

```
2022-09-15 15:44:39,826 - kedro.io.data_catalog - INFO - Loading data from
`Research data English` (CSVDataSet)…
```

[4]:
```
df.describe()
```

[4]:

	age	perfor_status	N	M	growth_type	\
count	120.000000	120.000000	120.000000	120.0	120.000000	
mean	50.841667	0.475000	0.716667	0.0	2.058333	
std	10.637650	0.501468	0.452506	0.0	0.490169	
min	25.000000	0.000000	0.000000	0.0	1.000000	
25%	44.000000	0.000000	0.000000	0.0	2.000000	
50%	51.000000	0.000000	1.000000	0.0	2.000000	
75%	59.000000	1.000000	1.000000	0.0	2.000000	
max	78.000000	1.000000	1.000000	0.0	3.000000	

	treatment_total_days	pelvic_total_dose	Pelv_fr	Dose_per_fr	\
count	120.000000	120.000000	120.000000	120.0	

```
mean                1.491667          48.850000   24.616667          2.0
std                 0.502027           3.513838    2.908069          0.0
min                 1.000000          40.000000   20.000000          2.0
25%                 1.000000          50.000000   25.000000          2.0
50%                 1.000000          50.000000   25.000000          2.0
75%                 2.000000          50.000000   25.000000          2.0
max                 2.000000          56.000000   50.000000          2.0

       midline_block_dose  …  hrctv_volume  eqd2_bladder  eqd2_rectum  \
count          120.000000  …    120.000000    120.000000   120.000000
mean             1.600000  …     29.185000     74.213333    60.719167
std              3.626235  …      8.920947      8.892192     6.299971
min              0.000000  …     13.100000     51.200000    47.200000
25%              0.000000  …     23.475000     69.175000    56.300000
50%              0.000000  …     28.100000     76.000000    60.300000
75%              0.000000  …     34.000000     80.850000    64.725000
max             10.000000  …     65.900000     89.100000    79.600000

       eqd2_sigmoid  eqd2_hrbrachy_dose  eqd2_total_dose  \
count    119.000000          119.000000       119.000000
mean      64.419328           36.589076        85.315126
std        6.956771            2.133382         2.484141
min       48.600000           29.400000        76.600000
25%       60.450000           35.800000        83.750000
50%       64.300000           36.100000        85.800000
75%       68.950000           38.100000        86.700000
max       82.100000           40.700000        91.000000

       before_brachy_categor  Last_follup_timing  rect_sig_grade          id
count             120.000000          120.000000       47.000000  120.000000
mean                1.350000           16.616667        0.446809  160.500000
std                 0.478969            8.680301        0.618853   34.785054
min                 1.000000            1.000000        0.000000  101.000000
25%                 1.000000            9.000000        0.000000  130.750000
50%                 1.000000           16.000000        0.000000  160.500000
75%                 2.000000           24.000000        1.000000  190.250000
max                 2.000000           38.000000        2.000000  220.000000

[8 rows x 27 columns]
```

[5]:
```
df.columns
```

[5]:
```
Index(['age', 'perfor_status', 'has_concom', 'tumour_stage', 'N', 'M', 'figo',
       'pathological_type', 'diagnos_tumor_size', 'growth_type',
       'vaginal_invasion', 'PaR', 'PaL', 'uterine_invasion', 'ct_abdomen',
       'ct_pelvic', 'mri_diagnostic', 'mri_before_brachy', 'pelvic_node_mts',
       'Paraaort_node_mts', 'treatment_total_days', 'pelvic_total_dose',
```

```
       'Pelv_fr', 'Dose_per_fr', 'midline_block_dose', 'midline_block_frac',
       'paramet_boost_dose', 'paramet_boost_fr', 'card_no', 'last_brachy_date',
       'applicator_name', 'total_apoint_left', 'total_apoint_right',
       'icru_rectum', 'icru_bladder', 'hrctv_volume', 'eqd2_bladder',
       'eqd2_rectum', 'eqd2_sigmoid', 'eqd2_hrbrachy_dose', 'eqd2_total_dose',
       'chemo_dose', 'chemo_numcycles', 'before_brachy_tumor_size',
       'before_brachy_categor', 'post_treatment_response',
       'post_treatment_response_date', 'last_response_date',
       'last_response_status', 'Last_follup_timing', 'rect_sig_grade',
       'Last_rect_eval_date', 'id', 'age_group'],
      dtype='object')
```

[6]:
```
diagnos_tumor_size = df['diagnos_tumor_size']
tumour_stage = df['tumour_stage']
before_brachy_tumor_size = df['before_brachy_tumor_size']
before_brachy_categor = df['before_brachy_categor']
```

[7]:
```
from scipy.stats import chi2_contingency
import pandas as pd
import scipy.stats as stats
```

[8]:
```
def chi2_contingency_test (test_data):

    stat, p, dof, ex = chi2_contingency(test_data)
    print("Expected cell frequencies: " + str(ex))
    print("Degree of freedom: " + str(dof))

    significance_level = 0.05
    print("p-value: " + str(format(p, '.10f')))
    if p <= significance_level:
        print('Reject NULL HYPOTHESIS. Thus, the variables are associated with␣
↪each other and happen to have a correlation between the variables.')
    else:
        print('ACCEPT NULL HYPOTHESIS. Thus, the grouping variables have no␣
↪association or correlation amongst them.')
```

[9]:
```
def fishers_exact_test (test_data):

    oddsratio, p = stats.fisher_exact(test_data)

    significance_level = 0.05
    print("p-value: " + str(format(p, '.10f')))
    if p <= significance_level:
        print('Reject NULL HYPOTHESIS. Thus, the variables are associated with␣
↪each other and happen to have a correlation between the variables.')
    else:
```

```
        print('ACCEPT NULL HYPOTHESIS. Thus, the grouping variables have no␣
↪association or correlation amongst them.')
```

[10]:
```
def label_before_brachy_tumor_size (row):
    if row['before_brachy_tumor_size'] == '>2 ':
        return 2
    if row['before_brachy_tumor_size'] == 'CR':
        return 0
    if row['before_brachy_tumor_size'] == ' 2 ':
        return 1
```

[11]:
```
def label_diagnos_tumor_size (row):
    if row['diagnos_tumor_size'] == '>4 ':
        return 1
    if row['diagnos_tumor_size'] == ' 4 ':
        return 0
```

[12]:
```
def label_tumour_stage (row):
    if row['tumour_stage'] == '2B':
        return 0
    if row['tumour_stage'] == '3B':
        return 1
    if row['tumour_stage'] == '4A':
        return 2
```

[13]:
```
# Creating ordinal attributes
df['before_brachy_tumor_size_ordinal'] = df.apply (lambda row:␣
↪label_before_brachy_tumor_size(row), axis=1)
df['diagnos_tumor_size_ordinal'] = df.apply (lambda row:␣
↪label_diagnos_tumor_size(row), axis=1)
df['tumour_stage_ordinal'] = df.apply (lambda row: label_tumour_stage(row),␣
↪axis=1)
```

[14]:
```
def spearman_corr_test(test_data):
    value = test_data.corr(method="spearman")
    print(value)
```

Diagnostic tumor size and and tumour stage

[15]:
```
# Create a contingency table
test_data = pd.crosstab(diagnos_tumor_size, tumour_stage)
print(test_data)
```

```
tumour_stage        2B  3B  4A
diagnos_tumor_size
>4                  20  40   3
 4                  51   6   0
```

Chi-Square Test of Independence

[16]:
```
chi2_contingency_test (test_data)
```

```
Expected cell frequencies: [[37.275 24.15   1.575]
 [33.725 21.85   1.425]]
Degree of freedom: 2
p-value: 0.0000000010
Reject NULL HYPOTHESIS. Thus, the variables are associated with each other and
happen to have a correlation between the variables.
```

In this test, there are expected test frequencies that are <5. Therefore, we use the Fisher's Exact test instead.

Fisher's Exact Test

[17]:
```
# Create a contingency table
test_data = pd.crosstab(diagnos_tumor_size, df.loc[df['tumour_stage'].
 ↪isin(['2B','3B'])]['tumour_stage'])
print(test_data)
```

```
tumour_stage        2B  3B
diagnos_tumor_size
>4                  20  40
 4                  51   6
```

[18]:
```
fishers_exact_test (test_data)
```

```
p-value: 0.0000000003
Reject NULL HYPOTHESIS. Thus, the variables are associated with each other and
happen to have a correlation between the variables.
```

Testing correlation of ordinal variables using Spearman's Rank Correlation Coefficient

[19]:
```
spearman_corr_test(df[['diagnos_tumor_size_ordinal','tumour_stage_ordinal']])
```

```
                           diagnos_tumor_size_ordinal  tumour_stage_ordinal
diagnos_tumor_size_ordinal                    1.000000              0.586854
tumour_stage_ordinal                          0.586854              1.000000
```

0.0.1 Diagnostic tumor size and before brachy tumor size

[20]:
```
# Create a contingency table
test_data = pd.crosstab(diagnos_tumor_size, before_brachy_tumor_size)
print(test_data)
```

```
before_brachy_tumor_size  >2   CR   2
diagnos_tumor_size
>4                        43   6    14
 4                         9  15    33
```

Chi-Square Test of Independence

[21]:
```
chi2_contingency_test (test_data)
```

```
Expected cell frequencies: [[27.3   11.025 24.675]
 [24.7    9.975 22.325]]
Degree of freedom: 2
p-value: 0.0000000518
Reject NULL HYPOTHESIS. Thus, the variables are associated with each other and
happen to have a correlation between the variables.
```

Testing correlation of ordinal variables using Spearman's Rank Correlation Coefficient

[22]:
```
df['before_brachy_tumor_size_ordinal'].value_counts()
```

[22]:
```
2    52
1    47
0    21
Name: before_brachy_tumor_size_ordinal, dtype: int64
```

[23]:
```
df['before_brachy_tumor_size'].value_counts()
```

[23]:
```
>2     52
 2     47
CR      21
Name: before_brachy_tumor_size, dtype: int64
```

[24]:
```
spearman_corr_test(df[['diagnos_tumor_size_ordinal','before_brachy_tumor_size_ordinal']])
```

```
                                diagnos_tumor_size_ordinal  \
diagnos_tumor_size_ordinal                        1.000000
before_brachy_tumor_size_ordinal                  0.494411

                                before_brachy_tumor_size_ordinal
diagnos_tumor_size_ordinal                              0.494411
before_brachy_tumor_size_ordinal                        1.000000
```

0.0.2 Before brachy tumor size and before brachy category

[25]:
```
# Create a contingency table
test_data = pd.crosstab(before_brachy_categor, before_brachy_tumor_size)
print(test_data)
```

```
before_brachy_tumor_size  >2   CR   2
before_brachy_categor
1                         15   21    42
2                         37    0     5
```

Chi-Square Test of Independence

[26]:
```
chi2_contingency_test (test_data)
```

```
Expected cell frequencies: [[33.8  13.65 30.55]
 [18.2   7.35 16.45]]
Degree of freedom: 2
p-value: 0.0000000000
Reject NULL HYPOTHESIS. Thus, the variables are associated with each other and
happen to have a correlation between the variables.
```

Testing correlation of ordinal variables using Spearman's Rank Correlation Coefficient

[27]:
```
spearman_corr_test(df[['before_brachy_categor','before_brachy_tumor_size_ordinal']])
```

```
                                before_brachy_categor  \
before_brachy_categor                        1.000000
before_brachy_tumor_size_ordinal             0.644586

                                before_brachy_tumor_size_ordinal
before_brachy_categor                                   0.644586
before_brachy_tumor_size_ordinal                        1.000000
```

0.1 References:

1. McHugh ML. The chi-square test of independence. Biochem Med (Zagreb). 2013;23(2):143-9. doi: 10.11613/bm.2013.018. PMID: 23894860; PMCID: PMC3900058. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/#:~:text=The%20Chi%2Dsquare%20test%20of%
2. Schober, Patrick MD, PhD, MMedStat; Boer, Christa PhD, MSc; Schwarte, Lothar A. MD, PhD, MBA. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia: May 2018 - Volume 126 - Issue 5 - p 1763-1768 doi: 10.1213/ANE.0000000000002864 https://journals.lww.com/anesthesia-analgesia/fulltext/2018/05000/correlation_coefficients___appropriate_use_and.50.aspx

[ ]:
```

```