Probability Theory and Mathematical Statistics

Homework 5, Vitaliy Pozdnyakov

Task 1. 535 **bombs fell on a southern London during World War II. The southern London had been divided on** 576 **districts with area of each district =** 0.25 **km. The table shows the number of bombs that fell on the districts.**

The number of bombs **k**	0	1	2	3	4	5
The number of districts $\mathbf{n_k}$	299	211	93	35	7	1

Using the goodness-of-fit test procedure based on the chi-square distribution and considering the $\alpha = 0.05$**, test:**

The form of the distribution of the number of bombs that fell on one district is Poisson.

$H_0 : X \sim Pois(\theta)$

The form of the distribution of the number of bombs that fell on one district is NOT Poisson.

$H_1 : X \nsim Pois(\theta)$

Test statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2(m - r - 1)$$

where m is the number of intervals, r is the number of parameters,

$$p_i = P(X = x_i = m) = \frac{\hat{\lambda}^m e^{-\hat{\lambda}}}{m!}$$

and let $\hat{\lambda} = \overline{X}$

Checking intervals:

In [1]:

```
import pandas as pd
import math
```

In [2]:

```
df = pd.DataFrame({
    'bombs': [0, 1, 2, 3, 4, 5 ],
    'districts': [229, 211, 93, 35, 7, 1]})
```

In [3]:

```
n = df['districts'].sum()
n
```

Out[3]:

```
576
```

$n = 576$

In [4]:

```
bar_x = 0
for _, val in df.iterrows():
    bar_x += val['bombs'] * val['districts']
bar_x = bar_x / n
round(bar_x, 2)
```

Out[4]:

```
0.93
```

$\overline{X} = 0.93$

In [5]:

```
# probability of X = m
def p_pois(m):
    if type(m) is not list:
        m = [m]
    res = 0
    for i in m:
        res += bar_x ** i * math.exp(-bar_x) / math.factorial(i)
    return res
```

In [6]:

```
for i, val in df.iterrows():
    p = p_pois(val['bombs'])
    if n * p < 10: print('critical interval:', i)
```

```
critical interval: 4
critical interval: 5
```

Interavals with indexes 4 and 5 have $np_i < 10$, so we need to union these intervals.

In [7]:

```
df = pd.DataFrame({
    'bombs': [0, 1, 2, 3, [4, 5]],
    'districts': [229, 211, 93, 35, 7+1]})
```

In [8]:

```
for i, val in df.iterrows():
    p = p_pois(val['bombs'])
    if n * p < 10: print('critical interval:', i)
```

```
critical interval: 4
```

The interval with index 4 has $np_i < 10$, so we need to union this interval with the neighbor (index 3).

In [9]:

```
df = pd.DataFrame({
    'bombs': [0, 1, 2, [3, 4, 5]],
    'districts': [229, 211, 93, 35+7+1]})
```

In [10]:

```
for i, val in df.iterrows():
    p = p_pois(val['bombs'])
    if n * p < 10: print('critical interval:', i)
```

Now there is no interval with $np_i < 10$

Critical area for the right-tail case $\overline{G} = [\chi^2_{cr}, +\infty]$

Critical values for the right-tail case

$$[1 - \chi^2(\chi^2_{cr}, m - r - 1) = 1 - \chi^2(\chi^2_{cr}, 4 - 1 - 1) = 1 - \chi^2(\chi^2_{\alpha}, 2) = 0.95]$$

(the number of intervals decreased by 2 after the union)

$$\Rightarrow \chi^2_{cr} = \chi^2_{0.05,3} = 7.82$$

Calculation:

In [11]:

```
chi2 = 0
for i, val in df.iterrows():
    p = p_pois(val['bombs'])
    chi2 += (val['districts'] - n * p) ** 2 / (n * p)
round(chi2, 2)
```

Out[11]:

```
0.75
```

$$\chi^2 = 0.75$$

$\chi^2 \notin \overline{G}$ so we assept H_0

Task 2. The results of men height measures are in tables

Height (cm)	143-146	146-149	149-152	152-155	155-158
The number of men	1	2	8	26	65
Height (cm)	**158-161**	**161-164**	**164-167**	**167-170**	**170-173**
The number of men	120	180	201	170	120
Height (cm)	**173-176**	**176-179**	**179-182**	**182-185**	**185-188**
The number of men	64	28	10	3	1

Using the goodness-of-fit test procedure based on the chi-square distribution and considering the $\alpha = 0.05$, test:

The form of the distribution of the men height is normal.

$H_0 : X \sim N(\theta_1, \theta_2^2)$

The form of the distribution of the men height is nonnormal.

$H_1 : X \nsim N(\theta_1, \theta_2^2)$

Test statistic:

$\chi^2 = \sum_{i=1}^{n} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2(m - r - 1)$ where m is the number of intervals, r is the number of parameters,

$$p_i = P(x_i \leq X \leq x_{i+1}) = \int_{x_i}^{x_{i+1}} \varphi_{N(\hat{\mu}, \hat{\sigma^2})}(t)dt$$

and let $\hat{\mu} = \overline{X}$ and $\hat{\sigma^2} = \frac{1}{n}\sum(x_i - \hat{\mu})^2$

Checking intervals:

In [12]:

```
from scipy.stats import norm
```

In [13]:

```
df = pd.DataFrame({
    'mens': [1, 2, 8, 26, 65, 120, 180, 201, 170, 120, 64, 28, 10, 3, 1],
    'height': [[143, 146],
               [146, 149],
               [149, 152],
               [152, 155],
               [155, 158],
               [158, 161],
               [161, 164],
               [164, 167],
               [167, 170],
               [170, 173],
               [173, 176],
               [176, 179],
               [179, 182],
               [182, 185],
               [185, 188]]})
```

In [14]:

```
n = df['mens'].sum()
n
```

Out[14]:

```
999
```

$n = 999$

In [15]:

```
def mean(interval):
    return (interval[0] + interval[1]) / 2
```

In [16]:

```
bar_x = 0
for _, val in df.iterrows():
    bar_x += val['mens'] * mean(val['height'])
bar_x = bar_x / n
round(bar_x, 2)
```

Out[16]:

```
165.53
```

$\overline{X} = 165.53$

In [17]:

```
hat_sigma = 0
for _, val in df.iterrows():
    hat_sigma += val['mens'] * (mean(val['height']) - bar_x) ** 2
hat_sigma = (hat_sigma / n) ** (1/2)
round(hat_sigma, 2)
```

Out[17]:

```
6.05
```

$\hat{\sigma} = 6.05$

In [18]:

```
# probability of a <= X <= b
def p_norm(interval):
    l_bound = (interval[0] - bar_x) / hat_sigma
    r_bound = (interval[1] - bar_x) / hat_sigma
    return norm.cdf(r_bound) - norm.cdf(l_bound)
```

In [19]:

```
for i, val in df.iterrows():
    p = p_norm(val['height'])
    if n * p < 10: print('critical interval:', i)
```

```
critical interval: 0
critical interval: 1
critical interval: 2
critical interval: 12
critical interval: 13
critical interval: 14
```

There are intervals with $np_i < 10$, so we need to union these.

In [20]:

```
df = pd.DataFrame({
    'mens': [1+2+8, 26, 65, 120, 180, 201, 170, 120, 64, 28, 10+3+1],
    'height': [[143, 152],
               [152, 155],
               [155, 158],
               [158, 161],
               [161, 164],
               [164, 167],
               [167, 170],
               [170, 173],
               [173, 176],
               [176, 179],
               [179, 188]]})
```

In [21]:

```
hat_sigma = 0
for _, val in df.iterrows():
    hat_sigma += val['mens'] * (mean(val['height']) - bar_x) ** 2
hat_sigma = (hat_sigma / n) ** (1/2)
round(hat_sigma, 2)
```

Out[21]:

```
6.18
```

$\hat{\sigma} = 6.18$

In [22]:

```
for i, val in df.iterrows():
    p = p_norm(val['height'])
    if n * p < 10: print('critical interval:', i)
```

Now there is no interval with $np_i < 10$

Critical area for the right-tail case $\overline{G} = [\chi^2_{cr}, +\infty]$

Critical values for the right-tail case

$$[1 - \chi^2(\chi^2_{cr}, m - r - 1) = 1 - \chi^2(\chi^2_{cr}, 11 - 2 - 1) = 1 - \chi^2(\chi^2_{\alpha}, 8) = 0.95]$$

(the number of intervals decreased by 4 after the union)

$$\Rightarrow \chi^2_{cr} = \chi^2_{0.05,8} = 15.51$$

Calculation:

In [23]:

```
chi2 = 0
for i, val in df.iterrows():
    p = p_norm(val['height'])
    chi2 += (val['mens'] - n * p) ** 2 / (n * p)
round(chi2, 2)
```

Out[23]:

```
2.73
```

$\chi^2 = 2.73$

$\chi^2 \notin \overline{G}$ so we assept H_0

Task 3. According to the census of Sweden in 1936, a sample of $25,263$ couples who married during 1931-1936 was obtained from the aggregate of all married couples. The table shows the distribution of the annual income and the number of children of the married couples in this sample.

Income (thousands) / Children (persons)	0-1	1-2	2-3	>3	Sum
0	2,161	3,577	2,184	1,636	9,558
1	2,755	5,081	2,222	1,052	11,110
2	936	1,753	640	306	3,635
3	225	419	96	38	778
≥ 4	39	98	31	14	182
Sum	6,116	10,928	5,173	3,016	25,263

X – the number of children in a married couple

Y – the annual income in a married couple

Using the contingency table tests procedure based on the chi-square distribution and considering the $\alpha = 0.05$, test:

H_0 : X is independent of Y

H_1 : X depends on Y

Test statistic:

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{(v_{ij} - (v_i v_j)/n)^2}{(v_i v_j)/n} \right) \sim \chi^2 \left((n_1 - 1) \cdot (n_2 - 1) \right)$$

Checking intervals:

In [24]:

```
import numpy as np
```

In [25]:

```
data = np.matrix([
    [2161, 3577, 2184, 1636],
    [2755, 5081, 2222, 1052],
    [936, 1753, 640, 306],
    [225, 419, 96, 38],
    [39, 98, 31, 14]])
```

In [26]:

```
n = data.sum()
n
```

Out[26]:

```
25263
```

$n = 25,264$

In [27]:

```
for i in range(data.shape[0]):
    for j in range(data.shape[1]):
        if data[i].sum() * data[:, j].sum() / n < 10:
            print('critical interval:', i, j)
```

There is no interval $(v_i v_j)/n < 10$, so we do not need to do the union.

Critical area for the right-tail case $\overline{G} = [\chi^2_{cr}, +\infty]$

Critical values for the right-tail case

$$[1 - \chi^2(\chi^2_{cr}, (n_1 - 1) \cdot (n_2 - 1)) = 1 - \chi^2(\chi^2_{cr}, (4 - 1) \cdot (5 - 1)) = 1 - \chi^2(\chi^2_{\alpha}, 12) = 0.95]$$

$$\Rightarrow \chi^2_{cr} = \chi^2_{0.05,12} = 21.03$$

Calculating:

In [28]:

```
chi2 = 0
for i in range(data.shape[0]):
    for j in range(data.shape[1]):
        chi2 += (
            (data[i, j] - data[i].sum() * data[:, j].sum() / n) ** 2
            /
            (data[i].sum() * data[:, j].sum() / n)
        )
round(chi2, 2)
```

Out[28]:

```
568.57
```

$$\chi^2 = 568.57$$

$\chi^2 \in \overline{G}$ so we reject H_0

Task 4. The table shows the income distribution of all industrial workers and employees of Sweden in 1930 for the age groups of 40-50 years and 50-60 years

Age (years) / Income (thousands)	40-50 years	50-60 years
1-2	7,831	7,558
2-3	26,740	20,685
3-4	35,572	24,186
4-5	20,009	12,280
5-6	11,527	6,776
≥ 6	6,919	4,222
Sum	108,598	75,707

X **– the income of the industrial workers and employees 40-50 years**

Y – the income of the industrial workers and employees 50-60 years

Using the contingency table tests procedure based on the chi-square distribution and considering the $\alpha = 0.05$, test:

H_0 : X is independent of Y

H_1 : X depends on Y

I do not know how to check the independence between X and Y, so i change the hypotheses

H_0 : X and Y are homogeneous

H_1 : X and Y are not homogeneous

Test statistic:

$$\chi^2 = n_1 \cdot n_2 \sum_{i=l}^{n} \frac{1}{v_{1i} + v_{2i}} \left(\frac{v_{1i}}{n_1} - \frac{v_{2i}}{n_2} \right)^2 \sim \chi^2 \left((s-1) \cdot (l-1) \right)$$

where s — the number of groups (2), l — the number of intervals, v_{ij} — the number of observations in the i-th group and the j-th interval, n_i - the number of observations in the i-th group

Critical area for the right-tail case $\overline{G} = [\chi^2_{cr}, +\infty]$

Critical values for the right-tail case
$[1 - \chi^2(\chi^2_{cr}, (s-1) \cdot (l-1)) = 1 - \chi^2(\chi^2_{cr}, (2-1) \cdot (6-1)) = 1 - \chi^2(\chi^2_{\alpha}, 5) = 0.95]$

$\Rightarrow \chi^2_{cr} = \chi^2_{0.05,5} = 11.07$

Calculating:

In [32]:

```
df = pd.DataFrame({
    'group_40': [7831, 26740, 35572, 20009, 11527, 6919],
    'group_50': [7558, 20685, 24186, 12280, 6776, 4222]})
```

In [41]:

```
chi2 = 0
group_40sum = df['group_40'].sum()
group_50sum = df['group_50'].sum()
for i, val in df.iterrows():
    chi2 += (
        (val['group_40'] / group_40sum - val['group_50'] / group_50sum) ** 2
        /
        (val['group_40'] + val['group_50'])
    )
chi2 *= group_40sum * group_50sum
round(chi2, 2)
```

Out[41]:

```
840.62
```

$\chi^2 = 840.62$

$\chi^2 \in \overline{G}$ so we reject H_0

Task 5. Suppose that 10 sets of hypotheses of the form

$H_0 : \mu_X = \mu_0$ (const)

$H_1 : \mu_X \neq \mu_0$ (const)

have been tested and that the P-values for these tests are 0.12, 0.08. 0.93, 0.02, 0.01, 0.05, 0.88, 0.15, 0.13, and 0.06. Use Fisher's procedure to combine all of these P-values. What conclusions can you draw about these hypotheses?

$\chi^2 = -2 \sum_{i=1}^{k} \ln p_i \sim \chi^2(2k)$

Critical area for the right-tail case $\overline{G} = [\chi^2_{cr}, +\infty]$

Critical values for the right-tail case $[1 - \chi^2(\chi^2_{cr}, 2k) = 1 - \chi^2(\chi^2_{cr}, 2 \cdot 10) = 1 - \chi^2(\chi^2_{\alpha}, 20) = 0.95]$

$\Rightarrow \chi^2_{cr} = \chi^2_{0.05,20} = 18.31$

In [1]:

```
p_values = [0.12, 0.08, 0.93, 0.02, 0.01, 0.05, 0.88, 0.15, 0.06]
```

In [4]:

```
chi2 = 0
for i in p_values:
    chi2 += math.log(i)
chi2 *= -2
round(chi2, 2)
```

Out[4]:

```
42.14
```

$\chi^2 = 42.14$

$\chi^2 \in \overline{G}$ so we reject H_0