Probability Theory and Mathematical Statistics

Homework 4, Vitaliy Pozdnyakov

Task 1. The table shows the historical data on temperature of January for two cities:

City \ Year	1891	1892	1893	1894	1895	1896	1897
City X	-19.2	-14.8	-19.6	-11.1	-9.4	-16.9	-13.7
City Y	-21.8	-15.4	-20.8	-11.3	-11.6	-19.2	-13.0

City \ Year	1898	1899	1900	1901	1902	1903	
City X	-4.9	-13.9	-9.4	-8.3	-7.9	-5.3	
City Y	-7.4	-15.1	-14.4	-11.1	-10.5	-7.2	

Considering significance level 0.05, test that the mathematical expectation of on temperature of January in City X is -13.75 ($H_0 : \mu_X = -13.75$), with alternative hypothesis $H_1 : \mu_X \neq -13.75$. $X \sim N(\mu_X, \sigma^2)$, where σ^2 is unknown.

$\alpha = 0.05$

Let $\mu_0 = -13.75$

We use the t-score for a mean where σ^2 is unknown:

$t = \frac{|\hat{\mu}_X - \mu_0|\sqrt{n-1}}{\sqrt{\hat{\sigma^2_X}}} \sim S(n-1)$ where $\hat{\sigma^2}$ is the biased sample variance $\hat{\sigma^2} = \frac{\sum(x_i - \hat{\mu})^2}{n}$

Critical area $\overline{G} = [-\infty, -t_{cr}] \cup [t_{cr}, +\infty]$

Critical values (via the table for a central confidence level)
$[S(t_{cr}, n-1) = S(t_{1-\alpha}, 12) = 0.95] \Rightarrow t_{cr} = t_{0.95} = 2.18$

Calculating of the t-score:

In [1]:

```
import pandas as pd
```

In [2]:

```
df = pd.DataFrame({
    'cityx': [-19.2, -14.8, -19.6, -11.1, -9.4, -16.9, -13.7, -4.9, -13.9, -9.4, -8.
    'cityy': [-21.8, -15.4, -20.8, -11.3, -11.6, -19.2, -13.0, -7.4, -15.1, -14.4, -
```

In [3]:

```
mean = df['cityx'].mean()
round(mean, 2)
```

Out[3]:

```
-11.88
```

In [4]:

```
round(df['cityx'].var(ddof=0), 2)
```

Out[4]:

```
22.14
```

$t = \frac{|-11.88-(-13.75)|\sqrt{12}}{\sqrt{22.14}} = 1.38$

$t \notin \overline{G}$ thereby accept H_0

Task 2. Considering significance level 0.05, test that the variance of on temperature of January in City X is 20 ($H_0 : \sigma^2 = 20$), with alternative hypothesis $H_1 : \sigma^2 \neq 20$. $X \sim N(\mu_X, \sigma^2)$, where μ_X is unknown.

$\alpha = 0.05$

Let $\sigma_0^2 = 20$

$\chi^2 = \frac{n\hat{\sigma}_X^2}{\sigma_0^2} \sim \chi^2(n-1)$

Critical area $\overline{G} = [-\infty, \chi_1^2] \cup [\chi_2^2, +\infty]$

Critical values $[\chi^2(\chi_1^2, n-1) - \chi^2(\chi_2^2, n-1) = \chi^2(\chi_{1-\alpha/2}^2, n-1) - \chi^2(\chi_{\alpha/2}^2, n-1) = 0.95]$

$\Rightarrow \chi_1^2 = \chi_{0.975}^2 = 4.4, \chi_2^2 = \chi_{0.025}^2 = 23.34$

$\chi^2 = \frac{13 \cdot 22.14}{20} = 14.39$

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

Task 3. Considering significance level 0.05, test that the $\sigma_X^2 = \sigma_Y^2$

$\alpha = 0.05$

$H_0 : \sigma_X^2 = \sigma_Y^2$

$H_1 : \sigma_X^2 \neq \sigma_Y^2$

$F = \frac{\hat{s}_X^2}{\hat{s}_Y^2} \sim F(k_1, k_2)$ where $k_1 = n_1 - 1$ and $k_2 = n_2 - 1$ and $\hat{s}^2$ is the unbiased sample variance

$\hat{s}^2 = \frac{1}{n-1} \sum (x_i - \hat{\mu})^2$

Critical area $\overline{G} = [-\infty, F_1] \cup [F_2, +\infty]$

Critical values $[F(F_1, k_1, k_2) - F(F_2, k_1, k_2) = F(F_{1-\alpha/2}, 12, 12) - F(F_{\alpha/2}, 12, 12) = 0.95]$

$\Rightarrow F_1 = F_{0.975} = 0.31, F_2 = F_{0.025} = 3.28$

In [5]:

```
round(df['cityx'].var(), 2)
```

Out[5]:

```
23.99
```

In [6]:

```
round(df['cityy'].var(), 2)
```

Out[6]:

```
21.76
```

$F = \frac{23.99}{21.76} = 1.1$

$F \notin \overline{G}$ thereby assept H_0

Task 4. Considering significance level 0.05, test that the $\mu_X = \mu_Y$

$\alpha = 0.05$

$H_0 : \mu_X = \mu_Y$

$H_1 : \mu_X \neq \mu_Y$

$$t = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{\frac{\left(\sum (x_i - \hat{\mu}_X)^2 + \sum (y_i - \hat{\mu}_Y)^2\right)(n_1 + n_2)}{n_1 n_2 (n_1 + n_2 - 2)}}} \sim S(n_1 + n_2 - 2)$$

Critical area $\overline{G} = [-\infty, -t_{cr}] \cup [t_{cr}, +\infty]$

Critical values (via the table for a central confidence level)

$[S(t_{cr}, n_1 + n_2 - 2) = S(t_{1-\alpha}, 24) = 0.95] \Rightarrow t_{cr} = t_{0.95} = 2.06$

In [7]:

```
meanx = df['cityx'].mean()
round(meanx, 2)
```

Out[7]:

```
-11.88
```

In [8]:

```
meany = df['cityy'].mean()
round(meany, 2)
```

Out[8]:

```
-13.75
```

In [9]:

```
sumx = 0
for i in list(df['cityx']):
    sumx += (i - meanx) ** 2
round(sumx, 2)
```

Out[9]:

```
287.88
```

In [10]:

```
sumy = 0
for i in list(df['cityy']):
    sumy += (i - meany) ** 2
round(sumy, 2)
```

Out[10]:

```
261.17
```

$$t = \frac{-11.88-(-13.75)}{\sqrt{\frac{(287.88+261.17)(13+13)}{13\cdot 13(13+13-2)}}} = 1$$

$t \notin \overline{G}$ thereby accept H_0

Task 5. In order to detect the effect of using a special seeder, 10 plots of land were seeded with an ordinary seeder, and another 10 plots of land were seeded with a special seeder. There are only two seeders (ordinary and special). The 20 total plots of land were divided into pairs (ordinary,special) with adjacent sections included in each pair. The question of which of the two adjacent sections should be processed by a special machine was decided by a coin toss. The table shows the differences in yields (harvest) from each pair (ordinary,special) of plots of land. Does using a special seeder give an increase in yield (harvest)?

The number of pair	The differences in yields (harvest)
1	2.4
2	1.0
3	0.7
4	0.0
5	1.1
6	1.6
7	-0.4
8	1.1
9	0.1
10	0.7

$\alpha = 0.05$

Let $X = \frac{\sum_{i=1}^{10} x_i}{10}$

$H_0 : X \sim (\mu_0, \sigma^2)$ where $\mu_0 = 0$

$H_1 : X \sim (\mu_0, \sigma^2)$ where $\mu_0 > 0$

$t = \frac{(\overline{X} - \mu_0)\sqrt{n-1}}{\sqrt{\hat{\sigma}_X^2}} \sim S(n-1)$

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

Critical value for the right-tail case (via the table for a central confidence level)
$[S(t_{cr}, n-1) = S(t_{1-2\alpha}, 9) = 0.9] \Rightarrow t_{cr} = t_{0.9} = 1.83$

In [11]:

```
df = pd.DataFrame({'diff': [2.4, 1, 0.7, 0, 1.1, 1.6, -0.4, 1.1, 0.1, 0.7]})
```

In [12]:

```
round(df['diff'].mean() * (9) ** (1/2) / (df['diff'].var(ddof=0)) ** (1/2), 2)
```

Out[12]:

```
3.21
```

$t = \frac{(0.83-0)\sqrt{10-1}}{\sqrt{0.6}} = 3.21$

$t \in \overline{G}$ thereby reject H_0

Answer: using a special seeder gives an statistically significant increase in yield (harvest).

Task 6. The researcher has two tools that measures an important scientific event:

Tool X: 0.8; 1.9; 3.0; 3.5; 3.8; 2.5; 1.7; 0.9; 1.0; 2.3; 3.3; 3.4

Tool Y: 1.4; 2.1; 3.1; 3.6; 2.7; 1.7; 1.1; 0.2; 1.6; 2.8; 4.0; 4.7

Considering significance level 0.05**, test that the** $\mu_X = \mu_Y$

$\alpha = 0.05$

$H_0 : \mu_X = \mu_Y$

$H_1 : \mu_X \neq \mu_Y$

$$t = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{\frac{\left(\sum(x_i-\mu_X)^2+\sum(y_i-\mu_Y)^2\right)(n_1+n_2)}{n_1 n_2(n_1+n_2-2)}}} \sim S(n_1 + n_2 - 2)$$

$n_1 = n_2 = 12$

Critical area $\overline{G} = [-\infty, -t_{cr}] \cup [t_{cr}, +\infty]$

Critical values (via the table for a central confidence level)
$[S(t_{cr}, n_1 + n_2 - 2) = S(t_{1-\alpha}, 22) = 0.95] \Rightarrow t_{cr} = t_{0.95} = 2.07$

In [13]:

```
import pandas as pd
```

In [14]:

```
df = pd.DataFrame({
    'toolx': [0.8, 1.9, 3.0, 3.5, 3.8, 2.5, 1.7, 0.9, 1.0, 2.3, 3.3, 3.4],
    'tooly': [1.4, 2.1, 3.1, 3.6, 2.7, 1.7, 1.1, 0.2, 1.6, 2.8, 4.0, 4.7]})
```

In [15]:

```
meanx = df['toolx'].mean()
round(meanx, 2)
```

Out[15]:

```
2.34
```

In [16]:

```
meany = df['tooly'].mean()
round(meany, 2)
```

Out[16]:

```
2.42
```

In [17]:

```
sumx = 0
for i in list(df['toolx']):
    sumx += (i - meanx) ** 2
round(sumx, 2)
```

Out[17]:

```
12.83
```

In [18]:

```
sumy = 0
for i in list(df['tooly']):
    sumy += (i - meany) ** 2
round(sumy, 2)
```

Out[18]:

```
18.78
```

In [19]:

```
t = (2.34 - 2.42) / (
        (12.83 + 18.78)
        *
        (12 + 12)
        /
        (12 * 12 * (12 + 12 ))
    ) ** (1/2)
round(t, 2)
```

Out[19]:

```
-0.17
```

$$t = \frac{2.34 - 2.42}{\sqrt{\frac{(12.83+18.78)(12+12)}{12 \cdot 12(12+12-2)}}} = -0.17$$

$t \notin \overline{G}$ thereby accept H_0

Task 7. The engineer tests his inventions that measures a voltage

Invention X: $1.32; 1.35; 1.32; 1.35; 1.30; 1.30; 1.37; 1.31; 1.39; 1.39.$

Invention Y: $1.35; 1.31; 1.31; 1.41; 1.39; 1.37; 1.32; 1.34.$

Which one has better precision? Considering significance level 0.05, test that the $\sigma_X^2 = \sigma_Y^2$

$\alpha = 0.05$

$H_0 : \sigma_X^2 = \sigma_Y^2$

Check that the invention X and the invention Y has equal precision

$H_1 : \sigma_X^2 \neq \sigma_Y^2$

$F = \frac{\hat{s}_X^2}{\hat{s}_Y^2} \sim F(k_1, k_2)$ where $k_1 = n_1 - 1$ and $k_2 = n_2 - 1$ and $\hat{s}^2$ is the unbiased sample variance

$\hat{s}^2 = \frac{1}{n-1} \sum (x_i - \hat{\mu})^2$

Critical area $\overline{G} = [-\infty, F_1] \cup [F_2, +\infty]$

Critical values
$[F(F_1, k_1, k_2) - F(F_2, k_1, k_2) = F(F_{1-\alpha/2}, 9, 7) - F(F_{\alpha/2}, 9, 7) = 0.95] \Rightarrow F_{0.975} = 0.24, F_{0.025} = 4.82$

In [20]:

```
import numpy as np
```

In [21]:

```
inventionx = np.array([1.32, 1.35, 1.32, 1.35, 1.30, 1.30, 1.37, 1.31, 1.39, 1.39])
inventiony = np.array([1.35, 1.31, 1.31, 1.41, 1.39, 1.37, 1.32, 1.34])
```

In [22]:

```
round(inventionx.var() / inventiony.var(), 2)
```

Out[22]:

```
0.9
```

$F = 0.9$

$F \notin \overline{G}$ thereby assept H_0

Answer: the fact that the inventions have the different precision is not statistically significant