# Quiz

A car rental company is interested in the amount of time its vehicles are out of operation for repair work. A random sample of nine cars showed that over the past year, the numbers of days each had been inoperative were

$$16 \quad 10 \quad 21 \quad 22 \quad 8 \quad 17 \quad 19 \quad 14 \quad 19$$

Stating any assumptions you need to make, find a 90% confidence interval for the mean number of days in a year that all vehicles in the company's fleet are out of the operation.

## Solution:

Let $X$ show a number of days vehicles being out of operation in the past year.
$(1-\alpha)\cdot 100\%$ confidence interval for $\mu$, when $\sigma^2$ is unknown:

$$\mathrm{CI}_{1-\alpha}(\mu) = \overline{X} \pm t_{n-1;\,\alpha/2}\cdot\frac{S}{\sqrt{n}}.$$

A value of sample mean:

$$\overline{x} = \frac{1}{9}\sum_{i=1}^{9} x_i \approx 16.22,$$

and a value of sample standard deviation:

$$s = \sqrt{\frac{1}{8}\sum_{i=1}^{9}(x_i - 16.22)^2} \approx 4.79.$$

From problem statement and just calculated statistics:

$$\begin{aligned}\mathrm{CI}_{90\%}(\mu) &= 16.22 \pm t_{8;\,0.05}\cdot\frac{4.79}{\sqrt{9}} = \\ &= 16.22 \pm 1.86\cdot\frac{4.79}{\sqrt{9}} = \\ &= 16.22 \pm 2.97 = \boxed{(13.25; 19.19)}.\end{aligned}$$

# Problem 1

A random sample was taken of 189 National Basketball Association games in which the score was not tied after one quarter. In 132 of these games, the team leading after one quarter won the game.

(a) Find the 90% confidence interval for the population proportion of all occasions on which the team leading after one quarter wins the game.

(b) Without doing the calculations, state whether a 95% confidence interval for the population proportion would be wider than or narrower than that found in 1.

## Solution:

$(1-\alpha)\cdot 100\%$ confidence interval for $p$:

$$\mathrm{CI}_{1-\alpha}(p) = \widehat{P} \pm z_{\alpha/2}\cdot\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}},$$

when $\widehat{P}$ can be approximated with normal distribution. The condition of approximation:

$$\begin{cases} n\widehat{P} > 5, \\ n\left(1-\widehat{P}\right) > 5. \end{cases}$$

A value of sample proportion:

$$\widehat{p} = \frac{132}{189} \approx 0.698.$$

Conditions of approximation are satisfied:

$$\begin{cases} 189\cdot 0.698 \approx 132 > 5, \\ 189\cdot 0.302 \approx 57 > 5, \end{cases}$$

so the formula for CI can be applied.

(a) From problem statement:

$$\begin{aligned} \mathrm{CI}_{90\%}(p) &= 0.698 \pm z_{0.05}\cdot\sqrt{\frac{0.698\cdot 0.302}{189}} = \\ &= 0.698 \pm 1.645\cdot\sqrt{\frac{0.698\cdot 0.302}{189}} = \\ &= 0.698 \pm 0.055 = \boxed{(0.643;\ 0.753)}. \end{aligned}$$

(b) A 95% CI will be wider than that of 90% since with more confidence we cover greater area of the distribution, and $z_{0.05} < z_{0.025}$, while $\widehat{p}$ and E.S.E.$(\widehat{p})$ are fixed:

$$\mathrm{CI}_{90\%}(p) \subset \mathrm{CI}_{95\%}(p).$$

# Problem 2

Sample of Small Business Center clients considering starting a business were questioned. Of a random sample of 94 males, 50 received assistance in business planning. Of an independent random sample of 68 females, 40 received assistance in business planning. Find a 99% confidence interval for the difference between the population proportion of male and female clients who received assistance in business planning.

## Solution:

Let $X$ be an indicator of receiving assistance by males, and let $Y$ be an indicator of receiving assistance by females.

$(1-\alpha)\cdot 100\%$ confidence interval for $p_X - p_Y$:

$$\mathrm{CI}_{1-\alpha}\left(p_X - p_Y\right) = \left(\widehat{P}_X - \widehat{P}_Y\right) \pm z_{\alpha/2}\cdot\sqrt{\frac{\widehat{P}_X(1-\widehat{P}_X)}{n_X} + \frac{\widehat{P}_Y(1-\widehat{P}_Y)}{n_Y}},$$

when $\widehat{P}_X$ and $\widehat{P}_Y$ can be approximated with normal distribution. The condition of approximation:

$$\begin{cases} n_X\widehat{P}_X > 5, \\ n_X\left(1-\widehat{P}_X\right) > 5; \end{cases} \qquad \begin{cases} n_Y\widehat{P}_Y > 5, \\ n_Y\left(1-\widehat{P}_Y\right) > 5. \end{cases}$$

Values of sample proportions:

$$\widehat{p}_X = \frac{50}{94} \approx 0.53, \qquad \widehat{p}_Y = \frac{40}{68} \approx 0.59.$$

Conditions of approximation are satisfied:

$$\begin{cases} 94\cdot\dfrac{50}{94} = 50 > 5, \\ 94\cdot\dfrac{44}{94} = 44 > 5; \end{cases} \qquad \begin{cases} 68\cdot\dfrac{40}{68} = 40 > 5, \\ 68\cdot\dfrac{28}{68} = 28 > 5. \end{cases}$$

so the formula for CI can be applied.

From problem statement:

$$\begin{aligned} \mathrm{CI}_{99\%}\left(p_X - p_Y\right) &= 0.53 - 0.59 \pm z_{0.005}\cdot\sqrt{\frac{0.53\cdot 0.47}{94} + \frac{0.59\cdot 0.41}{68}} = \\ &= 0.53 - 0.59 \pm 2.57\cdot\sqrt{\frac{0.53\cdot 0.47}{94} + \frac{0.59\cdot 0.41}{68}} = \\ &= -0.06 \pm 0.202 = \boxed{(-0.262;\ 0.142)}. \end{aligned}$$

There is no highly significant evidence that proportions of males and females, who received assistance in business planning, are different (0 is in the CI).

# Problem 3

A new training programme is designed to improve the performance of 100-metre runners. A random sample of nine 100-metre runners were trained according to this programme and, in order to assess its effectiveness, they participated in a run before and after completing this training programme. The times (in seconds) for each runner were recorded and are shown below. The aim is to determine whether this training programme is effective in reducing the average times of the runners.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Before training | 12.5 | 9.6 | 10.0 | 11.3 | 9.9 | 11.3 | 10.5 | 10.6 | 12.0 |
| After training | 12.3 | 10.0 | 9.8 | 11.0 | 9.9 | 11.4 | 10.8 | 10.3 | 12.1 |

Compute an 80% confidence interval for the difference in the means of the times.

## Solution:

Let $X$ show results of 100-metre runs before training, and let $Y$ show results of 100-metre runs after training.

The samples pair follows "before – after" behavior, so we should construct a new sample of differences $d_i = x_i - y_i$:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Before training ($x_i$) | 12.5 | 9.6 | 10.0 | 11.3 | 9.9 | 11.3 | 10.5 | 10.6 | 12.0 |
| After training ($y_i$) | 12.3 | 10.0 | 9.8 | 11.0 | 9.9 | 11.4 | 10.8 | 10.3 | 12.1 |
| Difference ($d_i$) | 0.2 | −0.4 | 0.2 | 0.3 | 0 | −0.1 | −0.3 | 0.3 | −0.1 |

$(1-\alpha)\cdot 100\%$ confidence interval for $\mu_X - \mu_Y$ in a case of paired samples:

$$\mathrm{CI}_{1-\alpha}\left(\mu_X - \mu_Y\right) = \overline{D} \pm t_{n-1;\,\alpha/2} \cdot \frac{S_D}{\sqrt{n}}.$$

A value of sample mean:

$$\bar{d} = \frac{1}{9}\sum_{i=1}^{9} d_i \approx 0.011,$$

and a value of sample standard deviation:

$$s_D = \sqrt{\frac{1}{8}\sum_{i=1}^{9}\left(d_i - 0.011\right)^2} \approx 0.307.$$

From problem statement and just calculated statistics:

$$\begin{aligned}\mathrm{CI}_{80\%}\left(\mu_X - \mu_Y\right) &= 0.011 \pm t_{8;\,0.1} \cdot \frac{0.307}{\sqrt{9}} = \\ &= 0.011 \pm 1.397 \cdot \frac{0.307}{\sqrt{9}} = \\ &= 0.011 \pm 0.143 = \boxed{(-0.132; 0.154)}.\end{aligned}$$

There is no significant evidence (with confidence level of 80%) that the programme is helpful (0 is in the CI).

# Problem 4

Let $Y$ and $X$ be people's reaction time (in seconds) to a green and red lights. Eight individuals participated in the experiment, results are in the table:

| Individual | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Green ($Y$) | 0.43 | 0.32 | 0.58 | 0.46 | 0.27 | 0.41 | 0.38 | 0.61 |
| Red ($X$) | 0.30 | 0.23 | 0.41 | 0.53 | 0.24 | 0.36 | not available | |

(a) Find a 90% confidence interval for the difference $\mu_X - \mu_Y$.

(b) Formulate the assumptions you have made.

## Solution:

(a) The samples pair follows "before – after" behavior, so we should construct a new sample of differences $d_i = x_i - y_i$. But $y_7$ and $y_8$ have no counterparts in $X$-sample, so we can not calculate differences for them. Unfortunately, they have to be eliminated from consideration.

| Individual | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Green ($y_i$) | 0.43 | 0.32 | 0.58 | 0.46 | 0.27 | 0.41 |
| Red ($x_i$) | 0.30 | 0.23 | 0.41 | 0.53 | 0.24 | 0.36 |
| Difference ($d_i$) | −0.13 | −0.09 | −0.17 | 0.07 | −0.03 | −0.05 |

$(1-\alpha)\cdot 100\%$ confidence interval for $\mu_X - \mu_Y$ in a case of paired samples:

$$\mathrm{CI}_{1-\alpha}\left(\mu_X - \mu_Y\right) = \overline{D} \pm t_{n-1;\,\alpha/2} \cdot \frac{S_D}{\sqrt{n}}.$$

A value of sample mean:

$$\bar{d} = \frac{1}{6}\sum_{i=1}^{6} d_i \approx -0.0667,$$

and a value of sample standard deviation:

$$s_D = \sqrt{\frac{1}{5}\sum_{i=1}^{6}\left(d_i + 0.0667\right)^2} \approx 0.0843.$$

From problem statement and just calculated statistics:

$$\begin{aligned}\mathrm{CI}_{90\%}\left(\mu_X - \mu_Y\right) &= -0.0667 \pm t_{5;\,0.05} \cdot \frac{0.0843}{\sqrt{6}} = \\ &= -0.0667 \pm 2.015 \cdot \frac{0.0843}{\sqrt{6}} = \\ &= -0.0667 \pm 0.0694 = \boxed{(-0.136; 0.003)}.\end{aligned}$$

There is no significant evidence (with confidence level of 90%) between reaction times to green and red colors (0 is in the CI).

(b) Differences $D_i$ should be independent and normally distributed.

# Problem 5

For random sample of 190 firms that revalued their fixed assets, the mean ratio of debt to tangible assets was 0.517 and he sample standard deviation was 0.148. For an independent random sample of 417 firms that did not revalue their fixed assets, the mean ratio of debt to tangible assets was 0.489 and the sample standard deviation was 0.159. Find a 99% confidence interval for the difference between the two population means.

## Solution:

Let $X$ show a ratio of debt to tangible assets of firms that revalued their fixed assets, and Let $Y$ show a ratio of debt to tangible assets of firms that did not revalue their fixed assets.

In this problem population variances are unknown, but considered samples are large: $n_X = 190$ and $n_Y = 417$. So practically we can use normal pivot functions, since:

$$t_{n-1;\,\alpha/2} \underset{n\to\infty}{\longrightarrow} z_{\alpha/2},$$
$$S^2 \underset{n\to\infty}{\longrightarrow} \sigma^2.$$

$(1-\alpha)\cdot 100\%$ confidence interval for $\mu_X - \mu_Y$ in a case of independent samples, when variances are unknown, but with large samples:

$$\mathrm{CI}_{1-\alpha}\left(\mu_X - \mu_Y\right) = \left(\overline{X} - \overline{Y}\right) \pm z_{\alpha/2} \cdot \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}.$$

From problem statement:

$$\begin{aligned}\mathrm{CI}_{99\%}\left(\mu_X - \mu_Y\right) &= 0.517 - 0.489 \pm z_{0.005} \cdot \sqrt{\frac{0.148^2}{190} + \frac{0.159^2}{417}} = \\ &= 0.517 - 0.489 \pm 2.57 \cdot \sqrt{\frac{0.148^2}{190} + \frac{0.159^2}{417}} = \\ &= 0.028 \pm 0.034 = \boxed{(-0.006; 0.062)}.\end{aligned}$$

There is no highly significant evidence of difference between mean ratios of debt to tangible assets of firms that revalued their fixed assets and that did not (0 is in the CI).

# Problem 6

A company sends a random sample of twelve of its salespeople to a course designed to increase their motivation and hence, presumably, their effectiveness. In the following year, these people generated sales with an average of \$435,000 and sample standard deviation \$56,000. During the same period, an independently chosen random sample of fifteen salespeople who had not attended the course obtained sales with average value \$408,000 and standard deviation \$43,000. Assuming that the two population distributions are normal and have the same variance, find a 95% confidence interval for the difference between their means.

## Solution:

Let $X$ show a value of sales, generated by a salesperson who attended motivational courses, and let $Y$ show a value of sales, generated by a salesperson who did not attend those courses.
$(1-\alpha)\cdot 100\%$ confidence interval for $\mu_X - \mu_Y$ in a case of independent samples with unknown but identical variance:

$$\mathrm{CI}_{1-\alpha}\left(\mu_X - \mu_Y\right) = \left(\overline{X} - \overline{Y}\right) \pm t_{n_X+n_Y-2;\,\alpha/2} \cdot S_p \cdot \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}.$$

A value of pooled sample standard deviation:

$$s_p = \sqrt{\frac{(n_X-1)\cdot s_X^2 + (n_Y-1)\cdot s_Y^2}{n_X+n_Y-2}} = \sqrt{\frac{11\cdot 56000^2 + 14\cdot 43000^2}{25}} \approx 49145.5.$$

From problem statement and just calculated statistics:

$$\begin{aligned}\mathrm{CI}_{95\%}\left(\mu_X - \mu_Y\right) &= 435000 - 408000 \pm t_{25;\,0.025} \cdot 49145.5 \cdot \sqrt{\frac{1}{12} + \frac{1}{16}} = \\ &= 435000 - 408000 \pm 2.06 \cdot 49145.5 \cdot \sqrt{\frac{1}{12} + \frac{1}{16}} = \\ &= 27000 \pm 39210 = \boxed{(-12210; 66210)}.\end{aligned}$$

There is no moderately significant evidence of difference between mean sales of salespersons, who attended motivational courses and who did not (0 is in the CI).

# Problem 7

Grandmother Martha has got bored of her reliable tomatoes variety "Bull's Heart" and has decided to try out new one, called "De Barao", which was conveniently advised by her best friend Agnia. Martha wants to find out if the time, required for tomatoes to ripen, is smaller for the new variety, but she doesn't know how to do it.

Fortunately, her smart grandson Michael had been studying mathematical statistics in the HSE for a last year. He asked her to give him a few samples, constituted from ripening times of each tomatoes variety. Martha has sent the following list (in days):

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Bull's Heart | 109 | 110 | 107 | 108 | 114 | 111 | 109 | 114 | 119 | 107 |
| De Barao | 105 | 115 | 100 | 89 | 113 | 87 | 91 | 100 | | |

At which maximal level of confidence Michael can claim that "De Barao"'s ripening time is different from that of "Bull's Heart"?

## Solution:

Let $X$ show a ripening time of "Bull's Heart", and let $Y$ show a ripening time of "De Barao".

Values of sample mean:

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i \approx 110.8, \qquad \overline{y} = \frac{1}{8}\sum_{i=1}^{8} y_i \approx 100,$$

and values of sample standard deviation:

$$s_X = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(x_i - 110.8)^2} \approx 3.82, \qquad s_Y = \sqrt{\frac{1}{7}\sum_{i=1}^{8}(y_i - 100)^2} \approx 10.62.$$

The ratio of sample standard deviations prevents us from assuming that $\sigma_X = \sigma_Y$:

$$\frac{s_Y}{s_X} \approx 2.78 > 2.$$

$(1-\alpha)\cdot 100\%$ confidence interval for $\mu_X - \mu_Y$ in a case of independent samples with unknown variances:

$$\mathrm{CI}_{1-\alpha}\left(\mu_X - \mu_Y\right) = \left(\overline{X} - \overline{Y}\right) \pm t_{k;\,\alpha/2} \cdot \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}},$$

where

$$k = \left\lfloor \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}} \right\rfloor.$$

A number of degrees of freedom:

$$k = \left\lfloor \frac{\left(\frac{3.82^2}{10} + \frac{10.62^2}{8}\right)^2}{\frac{\left(\frac{3.82^2}{10}\right)^2}{9} + \frac{\left(\frac{10.62^2}{8}\right)^2}{7}} \right\rfloor \approx \lfloor 8.45 \rceil = 8.$$

From problem statement and just calculated statistics:

$$\begin{aligned} \mathrm{CI}_{1-\alpha}\left(\mu_X - \mu_Y\right) &= 110.8 - 100 \pm t_{8;\,\alpha/2} \cdot \sqrt{\frac{3.82^2}{10} + \frac{10.62^2}{8}} = \\ &= 10.8 \pm t_{8;\,\alpha/2} \cdot 3.95. \end{aligned}$$

"De Barao"'s ripening time is significantly different from that of "Bull's Heart" if 0 will not be included in $\mathrm{CI}_{1-\alpha}\left(\mu_X - \mu_Y\right)$. Since the value of point estimate $\overline{x} - \overline{y}$ is positive, both sides of of $\mathrm{CI}_{1-\alpha}\left(\mu_X - \mu_Y\right)$ should also be positive. An inequality on lower boundary (upper one is redundant):

$$\begin{aligned} 0.8 - t_{8;\,\alpha/2} \cdot 3.95 &> 0, \\ t_{8;\,\alpha/2} &< \frac{10.8}{3.95} \approx 2.73. \end{aligned}$$

Using table values of $t$-distribution:

$$\begin{aligned} t_{8;\,0.01} &= 2.898, \\ t_{8;\,0.025} &= 2.306, \end{aligned}$$

so the closest value of $\alpha$ which satisfy inequality:

$$\begin{aligned} \frac{\alpha}{2} &= 0.025, \\ \alpha &= 0.05. \end{aligned}$$

The confidence level:

$$1 - \alpha = \boxed{0.95}.$$

# Problem 8

A candy maker produces mints that have a label weight of 20.4 grams. For quality assurance, $n = 16$ mints were selected at random from the Wednesday morning shift, resulting in the statistics $\overline{x} = 21.95$ and $s_x = 0.197$. On Wednesday afternoon $m = 13$ mints were selected at random, giving $\overline{y} = 21.88$ and $s_y = 0.318$. Find a 90% confidence interval for the $\sigma_x/\sigma_y$, the ratio of the standard deviations of the mints produced by the morning and by the afternoon shifts, respectively.

## Solution:

$(1-\alpha)\cdot 100\%$ confidence interval for $\sigma_x^2/\sigma_y^2$ with unknown means:

$$\mathrm{CI}_{1-\alpha}\left(\frac{\sigma_x^2}{\sigma_y^2}\right) = \left(\frac{1}{F_{n_x-1,\, n_y-1;\, \alpha/2}}\cdot\frac{S_x^2}{S_y^2};\ F_{n_y-1,\, n_x-1;\, \alpha/2}\cdot\frac{S_x^2}{S_y^2}\right).$$

From problem statement:

$$\begin{aligned}\mathrm{CI}_{90\%}\left(\frac{\sigma_x^2}{\sigma_y^2}\right) &= \left(\frac{1}{F_{15,\, 12;\, 0.05}}\cdot\frac{0.197^2}{0.318^2};\ F_{12,\, 15;\, 0.05}\cdot\frac{0.197^2}{0.318^2}\right) = \\ &= \left(\frac{1}{2.62}\cdot\frac{0.197^2}{0.318^2};\ 2.48\cdot\frac{0.197^2}{0.318^2}\right) = \\ &= (0.146; 0.952).\end{aligned}$$

Confidence interval for $\sigma_x/\sigma_y$ then:

$$\begin{aligned}\mathrm{CI}_{90\%}\left(\frac{\sigma_x}{\sigma_y}\right) &= (\sqrt{0.146}; \sqrt{0.952}) = \\ &= \boxed{(0.383; 0.975)}.\end{aligned}$$

# Problem 9

A factory operates with three machines of type $A$ and two machine of type $B$. The weekly repair costs $X$ for type $A$ machines are normally distributed with mean $\mu_1$ and variance $\sigma^2$. The weekly repair costs $Y$ for machines of type $B$ are also normally distributed but with mean $\mu_2$ and variance $3\sigma^2$. The expected repair cost per week for factory is thus $3\mu_1 + 2\mu_2$. You are given a random sample $x_1 = 100$, $x_2 = 120$, $x_3 = 95$ on costs of type $A$ machines and an independent random sample $y_1 = 200$, $y_2 = 280$ on costs for type $B$ machines. Construct a 95% confidence interval for expected repair cost per week $3\mu_1 + 2\mu_2$.

## Solution:

In order to find a pivot function for the parameter $3\mu_1 + 2\mu_2$ let's use its point estimate

$$3\overline{X} + 2\overline{Y}$$

since

$$\overline{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{3}\right), \qquad \overline{Y} \sim \mathcal{N}\left(\mu_2, \frac{3\sigma^2}{2}\right),$$

and

$$\mathsf{E}\left(3\overline{X} + 2\overline{Y}\right) = 3 \cdot \mathsf{E}\left(\overline{X}\right) + 2 \cdot \mathsf{E}\left(\overline{Y}\right) = 3\mu_1 + 2\mu_2.$$

A variance of the point estimate (samples are independent):

$$\mathsf{V}\left(3\overline{X} + 2\overline{Y}\right) = 3^2 \cdot \mathsf{V}\left(\overline{X}\right) + 2^2 \cdot \mathsf{V}\left(\overline{Y}\right) = 9 \cdot \frac{\sigma^2}{3} + 4 \cdot \frac{3\sigma^2}{2} = 9\sigma^2.$$

Thus, a distribution of the point estimate:

$$3\overline{X} + 2\overline{Y} \sim \mathcal{N}\left(3\mu_1 + 2\mu_2, 9\sigma^2\right).$$

A possible pivot function is:

$$Z = \frac{3\overline{X} + 2\overline{Y} - (3\mu_1 + 2\mu_2)}{3\sigma} \sim \mathcal{N}(0, 1),$$

but, unfortunately, we have an unknown parameter $\sigma$ within. Let's approximate it with appropriate sample standard deviation, and make pivot function follow a $t$-distribution.

Sample variances of each sample:

$$S_X^2 = \frac{1}{2}\sum_{i=1}^{3}\left(X_i - \overline{X}\right)^2, \qquad S_Y^2 = \sum_{j=1}^{2}\left(Y_j - \overline{Y}\right)^2.$$

According to Fisher's lemma they are distributed as follows:

$$\frac{2S_X^2}{\sigma^2} \sim \chi_2^2, \qquad \frac{S_Y^2}{3\sigma^2} \sim \chi_1^2.$$

Due to samples independence, their sum is also $\chi^2$-distributed:

$$Q = \frac{2S_X^2}{\sigma^2} + \frac{S_Y^2}{3\sigma^2} = \frac{2S_X^2 + \frac{1}{3}S_Y^2}{\sigma^2} \sim \chi_3^2.$$

A new pivot function:

$$T_3 = \frac{Z}{\sqrt{Q/3}} = \frac{3\overline{X} + 2\overline{Y} - (3\mu_1 + 2\mu_2)}{3\sigma} \cdot \frac{1}{\sqrt{\frac{2S_X^2 + \frac{1}{3}S_Y^2}{3\sigma^2}}} =$$

$$= \frac{3\overline{X} + 2\overline{Y} - (3\mu_1 + 2\mu_2)}{\sqrt{6S_X^2 + S_Y^2}} \sim t_3.$$

So a desired confidence interval takes the form:

$$\mathrm{CI}_{95\%}(3\mu_1 + 2\mu_2) = 3\overline{X} + 2\overline{Y} \pm t_{3;\,0.025} \cdot \sqrt{6S_X^2 + S_Y^2}.$$

Values of sample mean:

$$\overline{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = 105, \qquad \overline{y} = \frac{1}{2}\sum_{j=1}^{2} y_j = 240.$$

Values of sample variance:

$$s_X^2 = \frac{1}{2}\sum_{i=1}^{3} (x_i - 105)^2 = 175, \qquad s_Y^2 = \sum_{j=1}^{2} (y_j - 240)^2 = 3200.$$

From just calculated statistics:

$$\mathrm{CI}_{95\%}(3\mu_1 + 2\mu_2) = 3 \cdot 105 + 2 \cdot 240 \pm 3.182 \cdot \sqrt{6 \cdot 175 + 3200} =$$

$$= 795 \pm 207.4 = \boxed{(587.6; 1002.4)}.$$

# Problem 10

There are two machines bottling kvass. Accuracy (standard deviation) of the machine is $\sigma = 2$ ml. Amount of kvass in a bottle, poured by the first machine, is a random variable with normal distribution $X \sim \mathcal{N}(\mu_1, \sigma^2)$, and $Y \sim \mathcal{N}(\mu_2, \sigma^2)$ for the second machine. 200000 bottles of kvass poured by the first machine were delivered to Lapland, and also 100000 bottles of kvass poured by the second machine were delivered there. It is necessary to estimate at confidence 95% and precision $\pm 100$ liters the expected amount of kvass $200000\mu_1 + 100000\mu_2$, delivered to Lapland. Find $n, m$ – sample sizes of bottles to be opened that minimize the total sample size $n + m$.

## Solution:

Let $N = 200000$ be a number of bottles of kvass, poured by the first machine, and let $M = 100000$ be a number of bottles of kvass, poured by the second machine.

A point estimate of $N\mu_1 + M\mu_2$ is:

$$N\overline{X} + M\overline{Y}$$

since

$$\overline{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n}\right), \qquad \overline{Y} \sim \mathcal{N}\left(\mu_2, \frac{\sigma^2}{m}\right),$$

and

$$\mathsf{E}\left(N\overline{X} + M\overline{Y}\right) = N \cdot \mathsf{E}\left(\overline{X}\right) + M \cdot \mathsf{E}\left(\overline{Y}\right) = N\mu_1 + M\mu_2.$$

A variance of the point estimate (samples are independent):

$$\mathsf{V}\left(N\overline{X} + M\overline{Y}\right) = N^2 \cdot \mathsf{V}\left(\overline{X}\right) + M^2 \cdot \mathsf{V}\left(\overline{Y}\right) = N^2 \cdot \frac{\sigma^2}{n} + M^2 \cdot \frac{\sigma^2}{m} = \sigma^2\left(\frac{N^2}{n} + \frac{M^2}{m}\right).$$

Thus, a distribution of the point estimate:

$$N\overline{X} + M\overline{Y} \sim \mathcal{N}\left(N\mu_1 + M\mu_2, \sigma^2\left(\frac{N^2}{n} + \frac{M^2}{m}\right)\right).$$

A desired confidence interval with confidence 95% then:

$$\mathrm{CI}_{95\%}\left(N\mu_1 + M\mu_2\right) = N\overline{X} + M\overline{Y} \pm z_{0.025} \cdot \sigma \cdot \sqrt{\frac{N^2}{n} + \frac{M^2}{m}}.$$

From problem statement we know the required accuracy of estimation $e = 100$ liters:

$$e = z_{0.025} \cdot \sigma \cdot \sqrt{\frac{N^2}{n} + \frac{M^2}{m}}.$$

Let's denote the coefficient of standard error to standard deviation ratio as $k$:

$$k = \frac{\text{S.E.}\left(N\overline{X} + M\overline{Y}\right)}{\sigma} = \sqrt{\frac{N^2}{n} + \frac{M^2}{m}},$$

so we have a constraint on $n$ and $m$:

$$\frac{N^2}{n} + \frac{M^2}{m} = k^2 = \frac{e^2}{z^2_{0.025} \cdot \sigma^2} = \frac{100^2}{1.96^2 \cdot 0.002^2} \approx 650770512.$$

A constrained minimization problem:

$$\widehat{n}, \widehat{m} = \arg\min_{n,m}(n+m),$$
$$\text{s.t. } \frac{N^2}{n} + \frac{M^2}{m} = k^2.$$

Let's build a Lagrange function $\mathcal{L}$:

$$\mathcal{L} = n + m - \lambda\left(\frac{N^2}{n} + \frac{M^2}{m} - k^2\right).$$

By necessary condition of extremum:

$$\begin{cases} \left.\dfrac{\partial \mathcal{L}}{\partial n}\right|_{n=\widehat{n},m=\widehat{m},\lambda=\widehat{\lambda}} = 1 + \widehat{\lambda} \cdot \dfrac{N^2}{\widehat{n}^2} = 0, \\ \left.\dfrac{\partial \mathcal{L}}{\partial m}\right|_{n=\widehat{n},m=\widehat{m},\lambda=\widehat{\lambda}} = 1 + \widehat{\lambda} \cdot \dfrac{M^2}{\widehat{m}^2} = 0, \\ \left.\dfrac{\partial \mathcal{L}}{\partial \lambda}\right|_{n=\widehat{n},m=\widehat{m},\lambda=\widehat{\lambda}} = k^2 - \dfrac{N^2}{\widehat{n}} - \dfrac{M^2}{\widehat{m}} = 0. \end{cases} \quad \Longrightarrow \quad \begin{cases} \widehat{n} = \dfrac{N(N+M)}{k^2}, \\ \widehat{m} = \dfrac{M(N+M)}{k^2}, \\ \widehat{\lambda} = -\dfrac{(N+M)^2}{k^4}. \end{cases}$$

From problem statement:

$$\widehat{n} = \frac{200000 \cdot (200000 + 100000)}{650770512} \approx 92.2,$$
$$\widehat{m} = \frac{100000 \cdot (200000 + 100000)}{650770512} \approx 46.1.$$

In order to have $\widehat{n}$ and $\widehat{m}$ integers and simultaneously not to decrease the accuracy of estimation $e$, we should take the ceiling of each number:

$$\boxed{\begin{array}{l} \widehat{n} = 93, \\ \widehat{m} = 47. \end{array}}$$

Here we need to prove that values of $\widehat{n}$ and $\widehat{m}$ are indeed arguments of minimum. Second derivatives:

$$\left.\frac{\partial^2 \mathcal{L}}{\partial n^2}\right|_{n=\widehat{n},m=\widehat{m}} = -\frac{2\widehat{\lambda}N^2}{\widehat{n}^3}, \quad \left.\frac{\partial^2 \mathcal{L}}{\partial n \partial m}\right|_{n=\widehat{n},m=\widehat{m}} = \left.\frac{\partial^2 \mathcal{L}}{\partial m \partial n}\right|_{n=\widehat{n},m=\widehat{m}} = 0, \quad \left.\frac{\partial^2 \mathcal{L}}{\partial m^2}\right|_{n=\widehat{n},m=\widehat{m}} = -\frac{2\widehat{\lambda}M^2}{\widehat{m}^3}.$$

By sufficient condition of the minimum, let's see if the Hessian matrix $\mathcal{H}$ is positive-definite (all minors are positive):

$$\mathcal{H} = \begin{pmatrix} \left.\frac{\partial^2 \mathcal{L}}{\partial n^2}\right|_{n=\widehat{n},m=\widehat{m}} & \left.\frac{\partial^2 \mathcal{L}}{\partial n \partial m}\right|_{n=\widehat{n},m=\widehat{m}} \\ \left.\frac{\partial^2 \mathcal{L}}{\partial m \partial n}\right|_{n=\widehat{n},m=\widehat{m}} & \left.\frac{\partial^2 \mathcal{L}}{\partial n^2}\right|_{n=\widehat{n},m=\widehat{m}} \end{pmatrix} = \begin{pmatrix} \frac{2k^2}{N(N+M)} & 0 \\ 0 & \frac{2k^2}{M(N+M)} \end{pmatrix} \succ 0.$$

Q.E.D.