Confidence intervals. Part II

Statistics

Anton Afanasev

Higher School of Economics

DSBA 221
December 16, 2023

Seminar Overview

1. Quiz
2. Confidence intervals for population proportions
   - Single proportion
   - Difference of proportions
3. Confidence intervals for difference of population means
   - Paired samples
   - Independent samples. Variances are known
   - Independent samples. Variances are unknown, but equal
   - Independent samples. Variances are unknown
4. Confidence intervals for ratio of variances
   - Population means are unknown
   - Population means are known
5. Extra problems

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 10 21 22 8 17 19 14 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.

Confidence interval for population proportion p

- Let $X_1, \ldots, X_n$ be a random sample from a population with probability of "success" p, thus having Bernoulli(p) distribution.
- Point estimate of p is a sample proportion $\widehat{P}$:

$$\widehat{P} = \frac{\sum_{i=1}^{n} X_i}{n}$$

with moments:

$$\mathsf{E}\left(\widehat{P}\right) = p, \qquad \mathsf{V}\left(\widehat{P}\right) = \frac{p(1-p)}{n}.$$

- According to the CLT, for big n pivot function is:

$$h(X_1, \ldots, X_n; p) = \frac{\widehat{P} - p}{\sqrt{p(1-p)/n}} \overset{\text{CLT}}{\sim} \mathcal{N}(0,1).$$

Confidence interval for population proportion p

- Confidence interval of p with confidence level $1-\alpha$ then:

$$\mathsf{P}\left(-z_{\alpha/2} \leq \frac{\widehat{P}-p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right) = 1-\alpha.$$

$$\mathsf{P}\left(\widehat{P} - z_{\alpha/2}\cdot\sqrt{\frac{p(1-p)}{n}} \leq p \leq \widehat{P} + z_{\alpha/2}\cdot\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha.$$

- Let's estimate p with $\widehat{P}$:

$$\mathsf{P}\left(\widehat{P} - z_{\alpha/2}\cdot\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}} \leq p \leq \widehat{P} + z_{\alpha/2}\cdot\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}\right) = 1-\alpha.$$

- Estimated standard error of $\widehat{P}$:

$$\text{E.S.E.}\left(\widehat{P}\right) = \sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}.$$

Confidence interval for population proportion p

- $(1-\alpha)\cdot 100\%$ confidence interval for p can be written as:

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

- Condition of approximation:

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

- For simulations refer to the 1st block in the link:

Confidence intervals for p, $p_X - p_Y$

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.

1. Find the 90% confidence interval for the population proportion of all occasions on which the team leading after one quarter wins the game.
2. 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.

Confidence interval for difference $p_X - p_Y$

- Let $X_1, \ldots, X_{n_X}$ be a random sample from Bernoulli(p_X), and let $Y_1, \ldots, Y_{n_Y}$ be a random sample from Bernoulli(p_Y). Samples are independent.
- Point estimate of $p_X - p_Y$ is obviously $\widehat{P}_X - \widehat{P}_Y$ with pivot:

$$h\left(\{X_i\}_{i=1}^{n_X}, \{Y_j\}_{j=1}^{n_Y}; p_X - p_Y\right) = \\ = \frac{\widehat{P}_X - \widehat{P}_Y - (p_X - p_Y)}{\sqrt{p_X(1-p_X)/n_X + p_Y(1-p_Y)/n_Y}} \overset{\text{CLT}}{\sim} \mathcal{N}(0,1).$$

- Variance of a difference is a sum of variances (if independent):

$$\mathsf{V}\left(\widehat{P}_X - \widehat{P}_Y\right) = \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}.$$

Confidence interval for difference $p_X - p_Y$

- Similarly to the previous case, p_X and p_Y within variance should be estimated in order to get an explicit view of the interval. Estimated standard error of $\widehat{P}_X - \widehat{P}_Y$:

$$\text{E.S.E.}\left(\widehat{P}_X - \widehat{P}_Y\right) = \sqrt{\frac{\widehat{P}_X(1-\widehat{P}_X)}{n_X} + \frac{\widehat{P}_Y(1-\widehat{P}_Y)}{n_Y}}.$$

- Confidence interval of $p_X - p_Y$ with confidence level $1-\alpha$ then:

$$\mathsf{P}\Bigg(\left(\widehat{P}_X - \widehat{P}_Y\right) - z_{\alpha/2} \cdot \text{E.S.E.}\left(\widehat{P}_X - \widehat{P}_Y\right) \leq$$

$$\leq p_X - p_Y \leq$$

$$\leq \left(\widehat{P}_X - \widehat{P}_Y\right) + z_{\alpha/2} \cdot \text{E.S.E.}\left(\widehat{P}_X - \widehat{P}_Y\right) \Bigg) = 1 - \alpha.$$

Confidence interval for difference $p_X - p_Y$

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

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

- Condition of approximation:

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

- For simulations refer to the 2nd block in the link:

Confidence intervals for p, $p_X - p_Y$

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.

Inclusion of zero

- If $0 \in \text{CI}_{1-\alpha}\left(p_X - p_Y\right)$, then we can't claim with confidence $(1-\alpha)\cdot 100\%$ that either $p_X > p_Y$ or $p_X < p_Y$.
- If $0 \notin \text{CI}_{1-\alpha}\left(p_X - p_Y\right)$:

$\alpha \leq 1\%$:
There is a highly significant evidence that p_X is greater (less) than p_Y.

$1\% < \alpha \leq 5\%$:
There is a moderately significant evidence that p_X is greater (less) than p_Y.

$5\% < \alpha \leq 10\%$:
There is a weakly significant evidence that p_X is greater (less) than p_Y.

$\alpha > 10\%$:
There is no significant evidence that p_X is greater (less) than p_Y.

Confidence interval for difference $\mu_X - \mu_Y$

Paired samples

- Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}\left(\mu_X, \sigma^2\right)$, and let $Y_1, \ldots, Y_n$ be a random sample from $\mathcal{N}\left(\mu_Y, \sigma^2\right)$.
- Samples follow "before – after" behavior. They show states of the same n entities in different time instances.
  1. $n_X = n_Y = n$,
  2. $\sigma_X = \sigma_Y = \sigma$,
  3. samples are not independent.
- Let's construct a new sample of differences:

$$D_i = X_i - Y_i.$$

- D_i follows normal distribution as a linear combination of normals.
- $\overline{D}$ is an unbiased estimator of $\mu_X - \mu_Y$, since

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \overline{X} - \overline{Y},$$

which also follows normal distribution.

Confidence interval for difference $\mu_X - \mu_Y$

Paired samples

- Unbiased sample variance for D_i:

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(D_i - \overline{D}\right)^2$$

- According to Fisher's lemma:

$$\frac{(n-1)S_D^2}{2\sigma^2} \sim \chi^2_{n-1}.$$

- Pivot function (similarly to interval for μ with unknown σ):

$$h\left(D_1, \ldots, D_n; \mu_X - \mu_Y\right) = \frac{\overline{D} - (\mu_X - \mu_Y)}{S_D/\sqrt{n}} \sim t_{n-1}.$$

Confidence interval for difference $\mu_X - \mu_Y$

Paired samples

- Confidence interval of $\mu_X - \mu_Y$ with confidence level $1 - \alpha$ then:

$$\mathsf{P}\left(-t_{n-1;\,\alpha/2} \leq \frac{\overline{D} - (\mu_X - \mu_Y)}{S_D/\sqrt{n}} \leq t_{n-1;\,\alpha/2}\right) = 1 - \alpha.$$

$$\mathsf{P}\left(\overline{D} - t_{n-1;\,\alpha/2} \cdot \frac{S_D}{\sqrt{n}} \leq \mu_X - \mu_Y \leq \overline{D} + t_{n-1;\,\alpha/2} \cdot \frac{S_D}{\sqrt{n}}\right) = 1 - \alpha.$$

- $(1 - \alpha) \cdot 100\%$ confidence interval for $\mu_X - \mu_Y$ can be written as:

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

- For simulations refer to the 1st block in the link:

Confidence intervals for $\mu_X - \mu_Y$

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.

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	

1. Find a 90% confidence interval for the difference $\mu_X - \mu_Y$.
2. Formulate the assumptions you have made.

Confidence interval for difference $\mu_X - \mu_Y$

Independent samples, σ_X and σ_Y are known

- Let $X_1, \ldots, X_{n_X}$ be a random sample from $\mathcal{N}\left(\mu_X, \sigma_X^2\right)$, and let $Y_1, \ldots, Y_{n_Y}$ be a random sample from $\mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$.
- Samples are independent. Population variances σ_X and σ_Y are known.
- Point estimator of $\mu_X - \mu_Y$ is naturally:

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

- Pivot function is standardized $\overline{X} - \overline{Y}$:

$$h\left(\{X_i\}_{i=1}^{n_X}, \{Y_j\}_{j=1}^{n_Y}; \mu_X - \mu_Y\right) = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0, 1).$$

Confidence interval for difference $\mu_X - \mu_Y$

Independent samples, σ_X and σ_Y are known

- Confidence interval of $\mu_X - \mu_Y$ with confidence level $1 - \alpha$ then:

$$\mathsf{P}\left(-z_{\alpha/2} \leq \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \leq z_{\alpha/2} \right) = 1 - \alpha.$$

$$\mathsf{P}\left(\overline{X} - \overline{Y} - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \leq \mu_X - \mu_Y \leq \overline{X} - \overline{Y} + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right) = 1 - \alpha.$$

- $(1 - \alpha) \cdot 100\%$ confidence interval for $\mu_X - \mu_Y$ can be written as:

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

- For simulations refer to the 2nd block in the link:

Confidence intervals for $\mu_X - \mu_Y$

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.

Confidence interval for difference $\mu_X - \mu_Y$

Independent samples, $\sigma_X = \sigma_Y = \sigma$ is unknown

- Let $X_1, \ldots, X_{n_X}$ be a random sample from $\mathcal{N}\left(\mu_X, \sigma^2\right)$, and let $Y_1, \ldots, Y_{n_Y}$ be a random sample from $\mathcal{N}\left(\mu_Y, \sigma^2\right)$.
- Samples are independent. Population variances σ_X and σ_Y are unknown, but it's known that they are equal $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.
- Point estimator of $\mu_X - \mu_Y$:

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}\right)$$

- Let Z be standardized $\overline{X} - \overline{Y}$:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim \mathcal{N}(0, 1).$$

It can't be pivot function, since σ is unknown.

Confidence interval for difference $\mu_X - \mu_Y$

Independent samples, $\sigma_X = \sigma_Y = \sigma$ is unknown

- More observations we use – more accurate will be estimation of the sample variance. Let's apply Fisher's lemma to both samples and call their sum Q:

$$Q = \underbrace{\frac{(n_X - 1)S_X^2}{\sigma^2}}_{\sim \chi^2_{n_X - 1}} + \underbrace{\frac{(n_Y - 1)S_Y^2}{\sigma^2}}_{\sim \chi^2_{n_Y - 1}} \sim \chi^2_{n_X + n_Y - 2}.$$

  Summation to one chi-squared variable is possible due to independence of sample variances.

- Let's denote pooled variance as S_p^2:

$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

  with $n_X + n_Y - 2$ degrees of freedom.

Confidence interval for difference $\mu_X - \mu_Y$

Independent samples, $\sigma_X = \sigma_Y = \sigma$ is unknown

- S_p^2 is an approximation of unknown σ^2 and basically is a weighted average between S_X^2 and S_Y^2, with weights being their degrees of freedom.
- Pivot function with t-distribution is:

$$h\left(\{X_i\}_{i=1}^{n_X}, \{Y_j\}_{j=1}^{n_Y}; \mu_X - \mu_Y\right) = \frac{Z}{\sqrt{Q/(n_X + n_Y - 2)}} =$$

$$= \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}} \sim t_{n_X+n_Y-2}.$$

Confidence interval for difference $\mu_X - \mu_Y$

Independent samples, $\sigma_X = \sigma_Y = \sigma$ is unknown

- Confidence interval of $\mu_X - \mu_Y$ with confidence level $1 - \alpha$ then:

$$\mathsf{P}\left(-t_{n_X+n_Y-2;\,\alpha/2} \le \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \le t_{n_X+n_Y-2;\,\alpha/2} \right) = 1 - \alpha.$$

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

- $(1 - \alpha) \cdot 100\%$ confidence interval for $\mu_X - \mu_Y$ can be written as:

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

- For simulations refer to the 3rd block in the link:

Confidence intervals for $\mu_X - \mu_Y$

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.

Confidence interval for difference $\mu_X - \mu_Y$

Independent samples, $\sigma_X \neq \sigma_Y$ are unknown

- Let $X_1, \ldots, X_{n_X}$ be a random sample from $\mathcal{N}\left(\mu_X, \sigma_X^2\right)$, and let $Y_1, \ldots, Y_{n_Y}$ be a random sample from $\mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$.
- Samples are independent. Population variances σ_X and σ_Y are unknown.
- Standardized $\overline{X} - \overline{Y}$ can't be a pivot function, since σ_X^2 and σ_Y^2 are unknown:

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0, 1).$$

- Let's estimate σ_X^2 and σ_Y^2 with sample variances S_X^2 and S_Y^2.

Confidence interval for difference $\mu_X - \mu_Y$

Independent samples, $\sigma_X \neq \sigma_Y$ are unknown

- Quantity

$$\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}$$

is almost applicable to Fisher's lemma, but not exactly. So, it is almost distributed as χ_k^2, where k is the closest possible number of degrees of freedom.

- The pivot function is approximately t_k-distributed:

$$h\left(\{X_i\}_{i=1}^{n_X}, \{Y_j\}_{j=1}^{n_Y}; \mu_X - \mu_Y\right) = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \sim t_k.$$

- But what is the value k for degrees of freedom?

Confidence interval for difference $\mu_X - \mu_Y$

Independent samples, $\sigma_X \neq \sigma_Y$ are unknown

- Relation on k is given by:

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

- Since k should be integer:

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

Confidence interval for difference $\mu_X - \mu_Y$

Independent samples, $\sigma_X \neq \sigma_Y$ are unknown

- Confidence interval of $\mu_X - \mu_Y$ with confidence level $1 - \alpha$ then:

$$\mathsf{P}\left(-t_{k;\,\alpha/2} \leq \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{S_X^2}{n_X} + \dfrac{S_Y^2}{n_Y}}} \leq t_{k;\,\alpha/2} \right) \approx 1 - \alpha.$$

$$\mathsf{P}\left(\overline{X} - \overline{Y} - t_{k;\,\alpha/2} \cdot \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} \leq \mu_X - \mu_Y \leq \overline{X} - \overline{Y} + t_{k;\,\alpha/2} \cdot \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} \right) \approx 1 - \alpha.$$

- $(1 - \alpha) \cdot 100\%$ confidence interval for $\mu_X - \mu_Y$ can be written as:

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

- For simulations refer to the 4^{th} block in the link:

Confidence intervals for $\mu_X - \mu_Y$

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"?

When do we assume that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$?

- If population variances are unknown, we can choose between 2 intervals:

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

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

- The first one gives better results, if our assumption is correct that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.
- We can assume that if

$$\frac{1}{2} < \frac{S_X}{S_Y} < 2.$$

Confidence interval for ratio σ_Y^2/σ_X^2

Population means μ_X and μ_Y are unknown

- Let $X_1, \ldots, X_{n_X}$ be a random sample from $\mathcal{N}\left(\mu_X, \sigma_X^2\right)$, and let $Y_1, \ldots, Y_{n_Y}$ be a random sample from $\mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$. Values of μ_X and μ_Y are unknown.
- Unbiased estimators of σ_X^2 and σ_Y^2 are:

$$S_X^2 = \frac{1}{n_X - 1} \sum_{i=1}^{n_X} \left(X_i - \overline{X}\right)^2 \quad \text{and} \quad S_Y^2 = \frac{1}{n_Y - 1} \sum_{i=1}^{n_Y} \left(Y_i - \overline{Y}\right)^2.$$

- According to Fisher's lemma:

$$\frac{(n_X - 1)S_X^2}{\sigma_X^2} \sim \chi^2_{n_X - 1} \quad \text{and} \quad \frac{(n_Y - 1)S_Y^2}{\sigma_Y^2} \sim \chi^2_{n_Y - 1}.$$

- If $Q \sim \chi_p^2$ and $R \sim \chi_k^2$ and they are independent, then

$$F = \frac{Q/p}{R/k} \sim F_{p,k}.$$

Confidence interval for ratio σ_Y^2/σ_X^2

Population means μ_X and μ_Y are unknown

- Thus, dividing one χ^2 onto another, we get pivot function:

$$h\left(\{X_i\}_{i=1}^{n_X}, \{Y_j\}_{j=1}^{n_Y}; \frac{\sigma_Y^2}{\sigma_X^2}\right) = \frac{S_X^2}{S_Y^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} \sim F_{n_X-1,\, n_Y-1}.$$

- Confidence interval of σ_Y^2/σ_X^2 with confidence level $1-\alpha$ then:

$$\mathsf{P}\left(F_{n_X-1,\, n_Y-1;\, 1-\alpha/2} \leq \frac{S_X^2}{S_Y^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} \leq F_{n_X-1,\, n_Y-1;\, \alpha/2}\right) = 1-\alpha.$$

$$\mathsf{P}\left(F_{n_X-1,\, n_Y-1;\, 1-\alpha/2} \cdot \frac{S_Y^2}{S_X^2} \leq \frac{\sigma_Y^2}{\sigma_X^2} \leq F_{n_X-1,\, n_Y-1;\, \alpha/2} \cdot \frac{S_Y^2}{S_X^2}\right) = 1-\alpha.$$

Confidence interval for ratio σ_Y^2/σ_X^2

Population means μ_X and μ_Y are unknown

- $(1-\alpha)\cdot 100\%$ confidence interval for σ_Y^2/σ_X^2 can be written as:

$$\text{CI}_{1-\alpha}\left(\frac{\sigma_Y^2}{\sigma_X^2}\right) = \left(F_{n_X-1,\, n_Y-1;\, 1-\alpha/2}\cdot\frac{S_Y^2}{S_X^2};\; F_{n_X-1,\, n_Y-1;\, \alpha/2}\cdot\frac{S_Y^2}{S_X^2}\right).$$

- Critical value $F_{n_X-1,\, n_Y-1;\, 1-\alpha/2}$ is rarely given in tables of F-distribution, since it's closer to a left tail. Using relation of quantile functions for inverse F-distributed variables:

$$\text{CI}_{1-\alpha}\left(\frac{\sigma_Y^2}{\sigma_X^2}\right) = \left(\frac{1}{F_{n_Y-1,\, n_X-1;\, \alpha/2}}\cdot\frac{S_Y^2}{S_X^2};\; F_{n_X-1,\, n_Y-1;\, \alpha/2}\cdot\frac{S_Y^2}{S_X^2}\right).$$

- For simulations refer to the 1st block in the link:

Confidence intervals for σ_Y^2/σ_X^2

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 $\bar{x} = 21.95$ and $s_x = 0.197$. On Wednesday afternoon $m = 13$ mints were selected at random, giving $\bar{y} = 21.88$ and $s_y = 0.318$. Find a 90% confidence interval for the σ_x/σ_y, the ratio of the standard deviations of the mints produced by the morning and by the afternoon shifts, respectively.

Confidence interval for ratio σ_Y^2/σ_X^2

Population means μ_X and μ_Y are known

- Let $X_1, \ldots, X_{n_X}$ be a random sample from $\mathcal{N}\left(\mu_X, \sigma_X^2\right)$, and let $Y_1, \ldots, Y_{n_Y}$ be a random sample from $\mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$. Values of μ_X and μ_Y are known.
- Unbiased estimators of σ_X^2 and σ_Y^2 are:

$$\varsigma_X^2 = \frac{1}{n_X}\sum_{i=1}^{n_X}(X_i - \mu_X)^2 \quad \text{and} \quad \varsigma_Y^2 = \frac{1}{n_Y}\sum_{i=1}^{n_Y}(Y_i - \mu_Y)^2 .$$

- Distributions:

$$\frac{n_X\varsigma_X^2}{\sigma_X^2} \sim \chi^2_{n_X} \quad \text{and} \quad \frac{n_Y\varsigma_Y^2}{\sigma_Y^2} \sim \chi^2_{n_Y}.$$

- Pivot function:

$$h\left(\{X_i\}_{i=1}^{n_X}, \{Y_j\}_{j=1}^{n_Y}; \frac{\sigma_Y^2}{\sigma_X^2}\right) = \frac{\varsigma_X^2}{\varsigma_Y^2}\cdot\frac{\sigma_Y^2}{\sigma_X^2} \sim F_{n_X,\, n_Y}.$$

Confidence interval for ratio σ_Y^2/σ_X^2

Population means μ_X and μ_Y are known

- Confidence interval of σ_Y^2/σ_X^2 with confidence level $1-\alpha$ then:

$$\mathsf{P}\left(F_{n_X,\,n_Y;\,1-\alpha/2} \le \frac{\varsigma_X^2}{\varsigma_Y^2}\cdot\frac{\sigma_Y^2}{\sigma_X^2} \le F_{n_X,\,n_Y;\,\alpha/2}\right) = 1-\alpha.$$

$$\mathsf{P}\left(F_{n_X,\,n_Y;\,1-\alpha/2}\cdot\frac{\varsigma_Y^2}{\varsigma_X^2} \le \frac{\sigma_Y^2}{\sigma_X^2} \le F_{n_X,\,n_Y;\,\alpha/2}\cdot\frac{\varsigma_Y^2}{\varsigma_X^2}\right) = 1-\alpha.$$

- $(1-\alpha)\cdot 100\%$ confidence interval for σ_Y^2/σ_X^2 can be written as:

$$\mathrm{CI}_{1-\alpha}\left(\frac{\sigma_Y^2}{\sigma_X^2}\right) = \left(\frac{1}{F_{n_Y,\,n_X;\,\alpha/2}}\cdot\frac{\varsigma_Y^2}{\varsigma_X^2};\; F_{n_X,\,n_Y;\,\alpha/2}\cdot\frac{\varsigma_Y^2}{\varsigma_X^2}\right).$$

- For simulations refer to the 2nd block in the link:

Confidence intervals for σ_Y^2/σ_X^2

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 μ_1 and variance σ^2. The weekly repair costs Y for machines of type B are also normally distributed but with mean μ_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 $3\mu_1 + 2\mu_2$.

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


That's all Folks