Confidence intervals. Part II

Statistics

Anton Afanasev

Higher School of Economics

DSBA 211
December 14, 2022

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

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 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:

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

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

- Lat'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:

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

- Condition of approximation:

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

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

Confidence intervals for σ^2, p, $p_x - p_y$, $\mu_x - \mu_y$

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:

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

- Condition of approximation:

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

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

Confidence intervals for σ^2, p, $p_x - p_y$, $\mu_x - \mu_y$

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

$\alpha \leq 1\%$:
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 state 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:

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

- For simulations refer to the 5^{th} block in the link:

Confidence intervals for σ^2, p, $p_x - p_y$, $\mu_x - \mu_y$

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.

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:

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

- For simulations refer to the 6^{th} block in the link:

Confidence intervals for σ^2, p, $p_x - p_y$, $\mu_x - \mu_y$

Problem 4

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.

Look at the time!