Confidence intervals. Part III

Statistics

Anton Afanasev

Higher School of Economics

DSBA 211
December 17, 2022

Seminar Overview

1. Quiz
2. Confidence intervals for difference of population means
   - Independent samples. Variances are unknown, but equal
   - Independent samples. Variances are unknown
3. Confidence intervals for ratio of variances
   - Population means are unknown
   - Population means are known
4. One-sided confidence intervals
5. Practice
6. Confidence interval for correlation coefficient
   - Sample correlation coefficient
   - Fisher transformation
   - Confidence interval

Quiz

Economists have long realized that GNP alone does not measure total welfare of a country. Less tangible factors are important too, such as leisure and freedom from pollution and crime. To get some idea of how these other factors vary among countries, in 1970s a worldwide poll was undertaken (Gallup, 1976). To throw light on issue of crime, the question was asked: "Are you afraid to walk the neighboring streets at night?" The replies were as follows:

	United States	Japan	Latin America
Yes	40%	33%	57%
No	56%	63%	42%
No opinion	4%	4%	1%

Assuming each country's poll was equivalent in accuracy to a simple random sample of $n = 300$ people, find a 95% confidence interval for the difference in percentage answering "yes":

1. Between the United States and Japan.
2. Between the United States and Latin America.

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 as S_p^2 pooled variance:

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

  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:

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

- For simulations refer to the 7^{th} block in the link:

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

Problem 1

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:

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

- 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:

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

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

Confidence intervals for $\mu_x - \mu_y$, σ_y^2/σ_x^2, ρ

Problem 2

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:

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

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

- The first one is narrower and 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:

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

- 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:

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

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

Confidence intervals for $\mu_x-\mu_y$, σ_y^2/σ_x^2, ρ

Problem 3

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:

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

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

Confidence intervals for $\mu_x-\mu_y$, σ_y^2/σ_x^2, ρ

One-sided confidence intervals

- Let's constrain a pivot function with only one value. Critical value should separate α probability in a distribution tail for a confidence interval with level $1 - \alpha$.
- On the example of interval for μ with σ^2 unknown:

$$\mathsf{P}\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{n-1;\,\alpha}\right) = 1 - \alpha.$$

$$\mathsf{P}\left(\mu > \overline{X} - t_{n-1;\,\alpha} \cdot \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

$$(\mu)_{1-\alpha} \in \left(\overline{X} - t_{n-1;\,\alpha} \cdot \frac{S}{\sqrt{n}};\ +\infty\right).$$

- One-sided interval with upper bound similarly:

$$(\mu)_{1-\alpha} \in \left(-\infty;\ \overline{X} + t_{n-1;\,\alpha} \cdot \frac{S}{\sqrt{n}}\right).$$

Problem 4

During the Friday night shift, $n = 28$ mints were selected at random from a production line and weighted. They had average weight of $\bar{x} = 21.45$ grams and $s = 0.31$ gram. Give the lower endpoint of a 90% one-sided confidence interval for μ, the mean weight of all mints.

Problem 5

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.

Problem 6

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 7

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. To Lapland were delivered 200000 bottles poured by the first machine and 100000 bottles poured by the second machine. 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 – sizes of samples of bottles, which are needed to be open, that minimize the total sample size $n + m$.

Problem 8

Let 10.1, 9.7 be a sample from the normal population $X \sim \mathcal{N}\left(10, \sigma^2\right)$. Let 20.1, 19.5, 20.4 be a sample from the normal population $Y \sim \mathcal{N}\left(20, \sigma^2\right)$. Find a two-sided 90% confidence interval for the population standard deviation σ.

Confidence interval for correlation coefficient ρ

- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from population $\mathbf{X} = \begin{pmatrix} X & Y \end{pmatrix}^\top$ with bivariate normal distribution $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

  is a mean vector, and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

  is a covariance matrix.
- We want to estimate correlation coefficient $\rho = \mathsf{Corr}(X, Y)$. Values of parameters μ_X, μ_Y, σ_X and σ_Y are unknown.

Sample correlation coefficient

- Correlation coefficient is defined as:

$$\rho = \frac{\mathsf{Cov}(X, Y)}{\sigma(X) \cdot \sigma(Y)} = \frac{\mathsf{E}\left(X - \mathsf{E}(X)\right)\left(Y - \mathsf{E}(Y)\right)}{\sqrt{\mathsf{E}\left(X - \mathsf{E}(X)\right)^2} \cdot \sqrt{\mathsf{E}\left(Y - \mathsf{E}(Y)\right)^2}}.$$

- Natural point estimator of ρ is a Pearson correlation coefficient, defined right from the view of ρ itself:

$$\widehat{\rho} = \frac{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \cdot \sum\limits_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}},$$

where S_{XX} and S_{YY} are short notations for corrected sums of squares, S_{XY} is a corrected sum of cross-products.

Sample correlation coefficient

- Usually, it's more convenient to have corrected sums in the following form:

$$\widehat{\rho} = \frac{\sum\limits_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sqrt{\left(\sum\limits_{i=1}^{n} X_i^2 - n\overline{X}^2\right)\left(\sum\limits_{i=1}^{n} Y_i^2 - n\overline{Y}^2\right)}},$$

since it uses only 5 different statistics – $\overline{X}$, $\overline{Y}$, $\sum\limits_{i=1}^{n} X_i^2$, $\sum\limits_{i=1}^{n} Y_i^2$ and $\sum\limits_{i=1}^{n} X_i Y_i$, and they are given sometimes in problem statement explicitly.

Variance-stabilizing transformations

- The distribution of $\widehat{\rho}$ is a complex hypergeometric function, and moreover, its variance depends on ρ.
- Let's transform $\widehat{\rho}$ into another variable, which will be well-approximated with table function and will have constant variance. Such transformations are called variance-stabilizing.

Example (Poisson distribution)

- Let $P \sim \text{Poisson}(\lambda)$ with $\mathsf{E}(P) = \lambda$ and $\mathsf{V}(P) = \lambda$.
- In order to allow analysis of variances techniques, we want all sources to have identical variance.
- Using Anscombe transform: $Q = 2\sqrt{P + \frac{3}{8}}$,

  new variable will have $\mathsf{E}(Q) \approx \sqrt{\lambda + \frac{3}{8} - \frac{1}{4\sqrt{\lambda}}}$ and $\mathsf{V}(Q) \approx 1$ for larger parameters λ.

Fisher z-transform

- Fisher z-transform of a sample correlation coefficient $\widehat{\rho}$:

$$\widehat{Z} = \operatorname{artanh}(\widehat{\rho}),$$

where "artanh" is an inverse hyperbolic tangent function, equivalent to

$$\operatorname{artanh}(x) \equiv \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right).$$

- Transformed variable is very close to normal distribution with constant variance:

$$\widehat{Z} \overset{\text{approx}}{\sim} \mathcal{N}\left(\operatorname{artanh}(\rho), \frac{1}{n-3}\right).$$

- Approximation becomes better for larger n, though it's good for any $n > 3$.

Confidence interval for correlation coefficient ρ

- Confidence interval for transformed ρ:

$$(\operatorname{artanh}(\rho))_{1-\alpha} \in \operatorname{artanh}(\widehat{\rho}) \pm z_{\alpha/2} \cdot \frac{1}{\sqrt{n-3}}.$$

- Applying inverse Fisher transform $\left(\widehat{\rho} = \tanh\left(\widehat{Z}\right)\right)$ to those endpoints gives a result for ρ:

$$(\rho)_{1-\alpha} \in \left(\tanh\left(\operatorname{artanh}(\widehat{\rho}) - z_{\alpha/2} \cdot \frac{1}{\sqrt{n-3}}\right); \right.$$
$$\left. \tanh\left(\operatorname{artanh}(\widehat{\rho}) + z_{\alpha/2} \cdot \frac{1}{\sqrt{n-3}}\right)\right),$$

  where "tanh" is a hyperbolic tangent function

$$\tanh(x) \equiv \frac{e^{2x}-1}{e^{2x}+1}.$$

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

  Confidence intervals for $\mu_x - \mu_y$, σ_y^2/σ_x^2, ρ

Problem 9

The sample from bivariate normal distribution with random variables X and Y is following:

X	1.59	−2.20	−0.06	−1.45	−1.02	−2.59	−1.14	−3.25
Y	3.24	0.44	−1.14	5.40	2.09	5.33	1.25	8.72

Find 90% confidence interval for a population correlation coefficient ρ.

Look at the time!