STATISTICAL INFERENCE (MS-C1620). 8.1.–17.4.2025. Aalto University. University Lecturer Pekka Pere.

Re-examination 5.6.2024

You may use a scientific calculator, i.e. a calculator that has operations for trigonometry, power, exponential function, logarithm, and binomial coefficients. No other type of calculator is allowed. In particular, you may not use programmable calculators which can run program code, symbolic calculators which can manipulate symbolic expressions, graphical calculators which are able to plot functions, or calculators with Internet connection. You may have a personally handwritten double-sided A4 at the exam. **Present the intermediate steps of all your calculations, and justify all your answers in detail. A correct answer alone is worth zero points. All tasks are worth six points.** Check your answers. Next double-check them. Best of success to the exam! Note: Some quantiles of the Standard Normal distribution are at the end of the exam sheet.

1. The first publication prize of the Finnish Association for Legal and Social Sciences was awarded 4.12.2015 to the article by Peltonen *et al.* (2014).[1] The main finding of the study was that 5.5% of (biological, step, adoptive, or foster) mothers had committed severe violent acts toward their (biological, step *etc.*) child at least once during the preceding 12 months ($100 \times (141/2708)\% = 5.2\%$ in the table below). The table records the data.

	severe violence		
	yes (n)	no (n)	yes (%)
biological mother	132	2548	4.9
step/adoptive/foster mother	9	19	32.1

a) Calculate a (two-sided) 99% Wald confidence interval for the proportion of biological mothers who commit severe violent acts toward their child.

b) Calculate a (two-sided) 99% Agresti–Caffo confidence interval for the difference of proportions of biological and step/adoptive/foster mothers who commit severe violent acts toward their child. Does the confidence interval cover zero? What do you infer from it? Explain why the Agresti–Caffo confidence interval should be more reliable than the Wald confidence interval for the task of this question.

[1] https://oyy-ry.fi/2015/12/04/oyyn-julkaisupalkinnon-voittaja-valittu/ (read 1.6.2024). K. Peltonen, N. Ellonen, T. Pösö, and S. Lucas (2014): Mothers' Self-Reported Violence toward Their Children: A Multifaceted Risk Analysis. *Child Abuse & Neglect*, 12, 1923–1933.

2. Distribution of the monthly return of the Finnish Stock Market 31.10.1912–31.8.2022 is depicted in the histogram.[2] Economic theory suggests that stock returns should be independent. On the other hand empirical research often reports that stock returns feature a distribution with fat tails and with variance which (loosely speaking) varies from observation to observation. Such features may call for a nonparametric test.
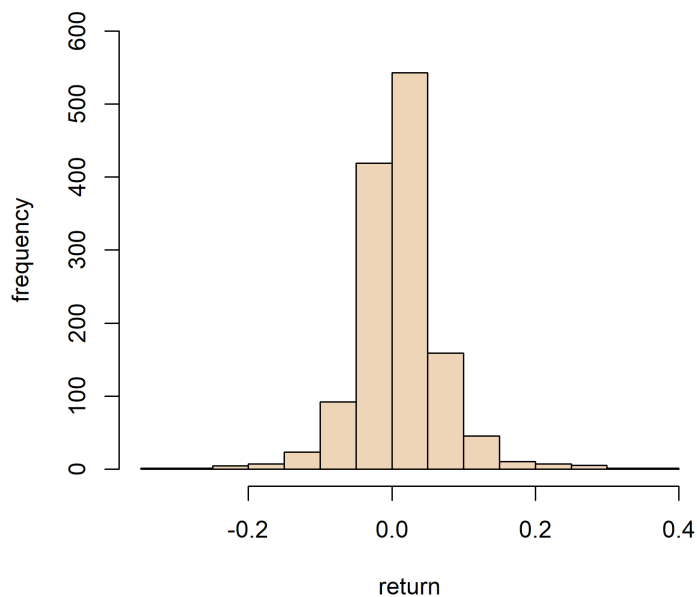
There are 1 303 nonzero observations in the data of which 771 are positive. Test the null hypothesis of zero median of returns with the sign-test.
 a) State the general formula of the test statistic.
 b) Calculate the numerical value of the test statistic.
 c) Conduct a two-sided test of the null hypothesis at significance level 0.001. What do you infer?

3. Keinänen and Tukiainen (2010) find the following equation by the method of least squares:

$$y = 2441.8 + 40.3x + \hat{\varepsilon},$$

$$\mathsf{R}^2 = 0.182, \ n = 258.$$

Here y is fine in euros levied by court to a labour union, x is economic loss in hundreds of thousands of euros to a company due to an illegitimate strike. $\hat{\varepsilon}$ is the residual of the model, and n is sample size.[3] Errors ε are assumed to follow Normal distribution $\mathsf{N}(0, \sigma^2)$ and be independent of each other. The observations and the fitted regression line are graphed in the figure (sakko = fine, taloudellinen menetys = economic loss).

a) Interpret the model carefully (the intercept, the regression coefficient, and R^2).

b) What is the sample correlation between x and y?

c) An observation stands out from the rest in the figure. By what term are such observations referred to? Should such observations always be deleted from the data before analysing the data statistically? Explain.

taloudellinen menetys (100 000 e)

---
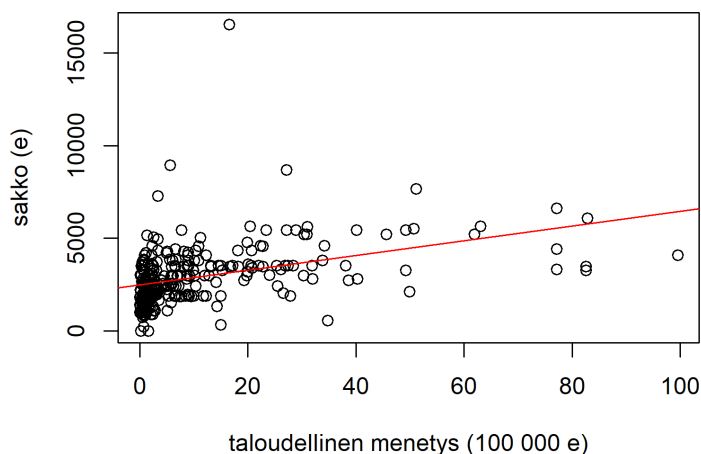
[3]A. Keinänen ja T. Tukiainen (2010): Laittomista työtaistelutoimenpiteistä tuomittujen hyvityssakkojen määräytyminen työtuomioistuinkäytännössä. Edilex 2010/15 (`http://www.edilex.fi/lakikirjasto/asiantuntijakirjoitukset/7008` (read 7.5.2016). Example of units of x: If $x = 30$ then the economic loss is $30 \times 100\,000 = 3\,000\,000$ euros.

4. Dallas *et al.* (2018, 2020) tested if word order matters. The experimentees were given menus with calorie information either to the left or to the right of the food items or not at all. Hebrew is read from the right to the left so the researchers hypothesized that Hebrew speaking experimentees are tempted to choose a food item with less calories if calories are stated to the right rather than to the left of the food items.[4] The researchers conducted an experiment to test the hypothesis: A menu was given by random assignment to Hebrew speaking experimentees. There were three kinds of menus: the menu either had calorie information to the left of each food item, had calorie information to the right of each food item, or had no calorie information. The means and standard deviations of the calories of the food choices the three groups made are in the table. The researchers carried out a standard one-way ANOVA. The outcome was F-statistic 2.92 with p-value 0.06.

	calories (kcal)		
	to the left	to the right	not reported
mean	1428.24	1308.66	1441.45
standard deviation	377.02	420.14	382.43
sample size	85	86	81

a) Are the group means in accordance with the theory of the researchers? (Statistical significance is not yet paid attention to.)

b) What are the degrees of freedom of the F-statistic?

c) The corresponding F-test is carried out at significance level 0.05. What is the outcome of the test?

A hint to any of the questions: The 0.925th, 0.95th, 0.975th, 0.99th, 0.995th, 0.999th, and 0.9995th quantiles of the Standard Normal distribution are 1.440, 1.645, 1.960, 2.326, 2.576, 3.090, and 3.290, respectively.

---

[4]Dallas, S.K., P.J. Liu, and P.A. Ubel (2018): Don't Count Calorie Labeling Out: Calorie Counts on the Left Side Items Lead to Lower Calorie Food Choices. *Journal of Consumer Psychology*, 29, 60–69. S.K. Dallas *et al.* (2020): Corrigendum. *Journal of Consumer Psychology*, 30, 571. In a previous test English speaking experimentees were tempted to choose a food item with less calories if the calorie information was printed to the left before the food item in line with the theory of the researchers.