Homework assignment
L3: Validity

Assignment date: **20.10.2020**
Deadline: **26.10.2020 23:59**
Slides: http://www.cs.cas.cz/martinkova/NMST570
Note: Send answers and `R` script to hladka@cs.cas.cz and martinkova@cs.cas.cz (in CC)
Include NMST570 in subject of your e-mails

Name:

Lecture presentation

Watch lecture presentation (online Zoom, or video shared on course webpage) and provide answer(s) to question(s) posed in the presentation. Note that the question slide is not included in the PDF file shared on course webpage.

Reading with Perusall (alternative)

It is possible to skip up to 4 HW assignments and to provide satisfactory feedback (10 relevant annotations, each may gain up to 1 point) to readings instead (Chapter 4 and relevant `R` code this week).

---

1 Reading with Perusall

Provide 1 annotation in Czech or in English to assigned reading (Chapter 4 and relevant `R` code this week) [1].

2 Validity in `ShinyItemAnalysis`

Ex. 2.1 Run `ShinyItemAnalysis` (online or locally), change data to HCI and answer following questions:

1. Which two items do correlate with the item 9 the most? What are the values of correlations? **(Validity/Correlation structure)** [1]

2. Read wording of these items in the supplement of the McFarland et al. (2017):

   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5459253/bin/supp_16.2.ar35_CombinedSupMats.pdf

   and try to explain why is the correlation between them increased? [1]

3. The HCI dataset includes criterion variable – indicator whether student plans to major in the life sciences. What is the correlation between this variable and total scores? Briefly interpret. **(Validity/Criterion validity)** [0.75]

3 Correction for restriction of range

Ex. 3.1 2,254 examinees were given theoretical test with 65 dichotomous items. Those examinees who passed the theoretical test (had score at least 52 points) were given practical test which consists of five parts. To pass the practical test, examinee needed to succeed in all five parts (i.e., gain score of 5 points).
Download data available at

http://www.cs.cas.cz/hladka/documents/test_theory_practice.RData

and create `R` script to answer following questions.

1. How many examinees passed theoretical test? What is the pass rate of the theoretical test? [0.75]

2. How many examinees passed practical test? What is the pass rate of the practical test in restricted sample (i.e., examinees who passed theoretical test)? [0.75]

3. Calculate correlation between scores of theoretical and practical tests in restricted sample (i.e., examinees who passed theoretical test) [1]

4. Use formula to correct for restriction of range to estimate correlation between scores of theoretical and practical tests in unrestricted sample:

$$\mathrm{cor}(X,Y) = \frac{\sigma_X \mathrm{cor}(x,y)}{\sqrt{\sigma_X^2 \mathrm{cor}(x,y)^2 + \sigma_x^2 - \sigma_x^2 \mathrm{cor}(x,y)^2}},$$

where X and Y are scores of unrestricted sample, x and y are scores of restricted sample, σ_X^2 and σ_x^2 are variances of X and x. [1.25]

HINT: You can use `attach(data)` to simply use variables by calling their name (e.g., `score_theory` instead of `data$score_theory` or `data[, "score_theory"]`). When missing values are present, you can add `na.rm = TRUE` in some functions to account for them (e.g., `mean(x, na.rm = TRUE)`). When calculating covariance or correlation, you need to use argument `use = "complete.obs"` (e.g., `cov(x, y, use = "complete.obs")`).

Ex. 3.2 Assume that score of the first test $X \sim \mathcal{N}(5,1)$. Score of the second test Y is linearly dependent on the score of the first test by formula $Y = 2X - 1 + e$, where $e \sim \mathcal{N}(0,1)$, and X and e are independent. Theoretical correlation between scores X and Y is $\frac{2}{\sqrt{5}} = 0.894$.

1. Using `R` create a script to generate X and Y of sample size 1,000. Use `set.seed(1)` for reproducibility. [0.5]
2. Calculate correlation between generated scores X and Y. Compare to theoretical results. [0.5]
3. Consider only those respondents who achieved
   - at least in 85th percentile in the first test
   - at most in 15th percentile in the first test
   - at most in 15th percentile or at least in 85th percentile in the first test

   Recalculate the estimate of correlation, apply formula for correction for restriction of the range, compare results and briefly comment. [1.5]

Scenario	Correlation in restricted sample	Correlation after correction
$\geq$ 85th percentile		
$\leq$ 15th percentile		
$\leq$ 15th percentile or $\geq$ 85th percentile		

HINT: For calculation of percentiles use function `quantile()`. For example use `quantile(X, 0.85)` to calculate 85th percentile of score X and `X[X >= quantile(X, 0.85)]` to compute scores X at least in 85th percentile. Use `&` for combining multiple conditions.

4 Provide feedback

Here you can provide feedback on lecture, lab session and/or materials (slides, video presentation, HW assignment, `ShinyItemAnalysis` application, etc.) [1pt bonus] :)