Psychological Test and Assessment Modeling, Volume 64, 2022 (2), 127-130

Note: The use of incorrect statistics-based formulations and phrases in papers of psychological research work

Dieter Rasch[1]

Preamble

Due to acting regularly as a reviewer of submitted papers in severals journals of Psychology I often realize a hardly statistician correct approach of the authors when trying for statistical testing. Although, there are a lot of Psychologists in the scientific community who are very competent with respect to the issues given in the following, it once more seems worthwhile to advice others fundamentally in Statistics.

In this Note we first remind the reader of the principles of the *Neyman-Pearson* concept of statistics (cf. Lehmann,1995/2008). Second, referring to the two-sample *t*-test we illustrate how psychological experiments or surveys can be planned in order to fulfil relevant precision requirements – and spare a lot of observations.

Fundamentally, statistical tests deal with hypotheses about certain parameters of the distribution used as a model of the empirical distribution of the observations of the (psychological) character in question. Take for instance the intelligence quotient; then to apply statistics means assuming that the empirical distribution of this character can be well modelled as a random variable by a theoretical distribution – this often is the normal distribution and we are interested in the parameter "average of the variable", that is the expectation of the theoretical distribution.

Now, any statistical test is based on one or more random samples of elements from the well-defined population(s) which deliver the observations of an experiment or a survey. According to the *Neyman-Pearson* concept they decide between two hypotheses, the null-hypothesis and the alternative hypothesis. The former claims that the interesting parameter has a certain value (or that the difference(s) of the interesting parameter in different samples are null), the latter claims that the null-hypothesis is wrong. Accepting one of these hypotheses does not at all imply that the accepted hypothesis is true; but only creates the belief that it is true. That is: By random sampling

[1] *Correspondence concerning this article should be addressed to:* Dieter Rasch, University of Natural Resources and Life Sciences, Vienna, Austria, d_rasch@t-online.de

we accept one of the hypotheses but matter of fact the decision of accepting the one and rejecting the other hypothesis can be wrong. As a consequence there are two kinds of possible errors. The error of the first kind (i.e. type-I-error) happens if we reject the null-hypothesis although it is actually true; the error of the second kind (i.e. type-II-error) happens if we accept the null-hypothesis although it is actually false.

Statistically important are the probabilities of these errors. They are called risks, the type-I-risk and the type-II-risk. Commonly in statistics the type-I-risk is symbolized by the Greek letter α, the type-II-risk by the Greek letter β . As α and β represent specific (*a priori* settled) values, there is however, as sometimes used in psychology, neither an "α risk" nor a "β risk". Even worse, we found in a submitted (but rejected) paper the following formulation: *The α error should be fixed to .05 the β error to .2. We recommend to choose $\beta - \alpha < .2$.* Apart from the fact, that here errors are mixed up with their probabilities this recommendation means only nonsense: From the point of planning a study, α and β are completely independent from each other but have to be determined only with respect to the given content.

Now, as concerns the test of comparing two expectations (also called means), tests have been developed for the case of a continuous character's normal distribution in two independently random samples from two populations. The matter is to test the null-hypothesis H_0: $\mu_1 = \mu_2$ (μ_1 , μ_2 being expectations and the population mean, respectively, of the two populations in question) against a two-sided alternative hypothesis H_A: $\mu_1 \neq \mu_2$.

In a paper submitted (but not published) to this journal the authors wrote erroneously: "*If one of the two means is a constant μ_0 then the two-sample problem becomes a one-sample problem."* This is completely wrong. At first as well μ_1 as μ_2 are constants but they are unknown. What the authors meant is that one of the expectation is known. But then we really have a one-sample problem but this does not arise from the two-sample problem.

We now have to test the null-hypothesis H_0: $\mu = \mu_0$ as opposed to the two-sided alternative H_A: $\mu \neq \mu_0$. Then the best test is the one-sample *t*-test.

Although, tests of H_0: $\mu_1 = \mu_2$ have been developed under the presumption of normal distributions they can be applied even in case of non-normality, as they are robust against non-normality (see Rasch & Guiard, 2004, and Rasch, Teuscher, & Guiard, 2007). On the other side, the well-known two-sample *t*-test is very sensitive against non-homogeneity of the variances of the two distributions. That is, if they are not equal, the nominal type-I-risk does not hold. For this reason it is at any rate to prefer the *Welch*-test, instead. This test and the belonging R-program can be found in Rasch, Kubinger and Yanagida (2011) at pages 206-208. Rasch, Kubinger, and Moder (2011) worked out that the idea of pre-testing the variances' equality leads to non-calculable overall type-I- and type-II-risk.

In any case, planning a study should always be done before data were sampled. That is, the sample sizes should be calculated in advance in that way, that a defined relevant

effect size ([standardized] mean difference $\frac{\mu_1 - \mu_2}{\sigma}$) is – by a given type-I-risk α – actually only not detected with a given type-II-risk β. Such a calculation can be done by an R-program described in Rasch, Kubinger, and Yanagida (2011). Using this, that leads for several examples to the sample sizes n given in Table 1. This table can be applied for the *Welch*-test by using for σ the presumably smaller of the two standard deviations.

Table 1: Sample sizes of each of the two samples for comparing two expectations for a two-sided alternative hypothesis

α	β	$\frac{\mu_1 - \mu_2}{\sigma}$	n
.05	.05	.5	106
		1	28
	.1	.5	86
		1	23
	.2	.5	64
		1	17
.1	.1	.5	70
		1	18
	.2	.5	51
		1	14
	.25	.5	44
		1	12

Conclusion

Researchers in Psychology should always take care in using correct statistical expressions and formulations but may not apply commonly used slovenly verbalisms which disclose them as not statistical firm and statistically not well-knowing. By the way they are advised to use instead of the two-sample *t*-test at any rate the *Welch*-test – whilst the one-sample *t* -test is without doubt the best method on their disposal.

Acknowledgement

The author thanks the chief editor for helpful comments.

References

Lehmann, E. L. (1959/2008) Testing Statistical Hypothesis, New York: Wiley.

Rasch D. & Guiard, V. (2004). The Robustness of Parametric Statistical Methods. Psychology Science, 46, 175-208.

Rasch, D., Teuscher, F., & Guiard, V. (2007). How Robust are tests for two independent samples? Journal of. Statistical Planning. Inference, 137, 2706-2720.

Rasch, D., Kubinger, K. D., & Moder, K (2011). The two-sample t test: pre-testing its assumptions does not pay off. Statistical Papers, 52, 219 – 231.

Rasch, D., Kubinger, K. D., & Yanagida, T. (2011) Statistics in Psychology using R and SPSS. New York: Wiley.