Psychological Test and Assessment Modeling, Volume 54, 2012 (1), 90-99

Hypothesis testing and the error of the third kind

Dieter Rasch[1]

Abstract

In this note it is shown that the concept of an error of the third kind (type-III-error) stems from a wrong understanding of the concept of hypothesis testing in mathematical statistics. If the alternative hypothesis in a statistical testing problem states that the null hypothesis is wrong, then an error of the third kind cannot occur. This error can only be defined in more-decision-problems, i.e. problems where we have to decide for one of more than two possibilities. Further the problem of the power of tests between a null and an alternative hypothesis is discussed. The notation of set theory is used because it allows for a concise notation of this type of statistical testing problems. The notation is explained in an appendix.

Key words: Hypothesis testing, more-decision-problems, Neyman-Pearson Theory, power of a test.

[1] *Correspondence concerning this article should be addressed to:* Dieter Rasch, PhD, University of Life Sciences Vienna; email: renate-rasch@t-online.de

1. Introduction

Kimball (1957) defined "error of the third kind in statistical consulting" as the error of giving the right answer to the wrong problem. Kimball attributed this type of error to poor communication between the consultant and the client. Raiffa (1968) described a type III error as correctly solving the wrong problem.

A somewhat different type of error of the third kind which will be discussed in this paper was suggested by Mosteller (1948) in proposing a non-parametric test for deciding whether one population, out of *k* populations characterized by a location parameter, has shifted too far to the right of the others. He defines it as “the error of correctly rejecting the null hypothesis for the wrong reason.”

This idea can later be found especially in psychological and medical journals. To mention some: Hopkins (1973), Leventhal and Huynh (1996), Sharon and Carpenter (1999), MacDonald (1999), Camdeviren and Mendes (2005), and Mendes (2004, 2007) but also in Elsevier´s Dictionary of Biometry (Rasch, Tiku & Sumpf, 1994).

In Rasch, Tiku and Sumpf we can read that the error of the third kind occurs when we have to select one of three hypotheses, and one of them is called the null hypothesis.

In the Neyman-Pearson test theory, which can best be found in a modern version in Lehmann and Romano (2005), a statistical test is the basis for a decision about which one of two hypotheses can be accepted. One of them is called the null, the other the alternative hypothesis. The hypotheses are a partition of the parameter space into two parts which together cover the whole space and are non-overlapping. In this book we now find no error of the third kind – as we will see in the sequel, such a concept is impossible in the Neyman-Pearson test theory.

We will first give a short repetition of the Neyman-Pearson test theory as it can be found in Lehmann and Romano (2005) and on a lower level in a text book by Rasch, Kubinger and Yanagida (2012) or its German version Kubinger, Rasch and Yanagida (2011) and then discuss the problem of the third kind error.

2. The Neyman-Pearson test theory

A statistical test is in mathematical statistics defined as a rule for a decision based on a random sample from a probability distribution about a conjecture for just this distribution. The conjecture is called the null hypothesis, and the contrary of it, namely the fact that the null hypothesis is wrong, is called the alternative hypothesis. In the parametric case which will be discussed here, we make a partition of the total set of possible parameter values, called the parameter space Ω, into two non-overlapping parts Ω_0 and $\Omega \setminus \Omega_0$; their union is the complete parameter space.

Let $f(x;\theta)$ be either a probability or a density function where θ in general parameter vector taking values in a *t*-dimensional parameter space $\Omega \subseteq \Re^t$ that could be the whole

t-dimensional Euclidian space $\Re^t$. A null hypothesis H_0 is a statement that the parameter θ lies in a subset $\Omega_0 \subset \Omega$ of the parameter space, i.e.

$$H_0 : \theta \in \Omega_0 \subset \Omega$$

The alternative hypothesis H_A then states that θ lies in the other part of Ω namely in $\Omega \setminus \Omega_0$, i.e.

$$H_A : \theta \in \Omega \setminus \Omega_0 = \Omega_A \subset \Omega .$$

There are two types of error, the error of the first kind: *rejecting* H_0 if it is correct and the error of the second kind: *accepting* H_0 if it is wrong. In this theory there is no place for an error of the third kind.

The probability of both these errors depends on the unknown parameter $\theta \in \Omega$; they are called the risk of the first kind $\alpha(\theta), \theta \in \Omega_0$ and the risk of the second kind $\beta(\theta), \theta \in \Omega_A$ correspondingly.

3. Test power

The power of a test is in mathematical statistics defined as the probability $\pi(\theta)$ of rejecting the null hypothesis as a function of θ, and not as stated for instance by Leventhal and Huynh (1996) and in many other psychological texts, as the probability of rejecting a *wrong* null hypothesis. Thus power is defined as

$$\pi(\theta) = \begin{matrix} \alpha(\theta), \text{if } \theta \in \Omega_0 \\ 1 - \beta(\theta), \text{if } \theta \in \Omega_A \end{matrix} .$$

and not as $1 - \beta(\theta)$, if $\theta \in \Omega_A$. Maybe some people would say that this is not very important and only a question of different definitions. But the more general concept used in mathematical statistics is needed to define an unbiased test. A statistical test is called unbiased if its power is never below the significance level α. This concept is needed if we discuss uniformly most powerful tests – UMP tests for short. An UMP test is one whose power function is in no part of the parameter space exceeded by the power function of any other test. The test of the hypothesis about the mean of a normal distribution with known variance (as well as the corresponding t-test for unknown variance) rejecting the null hypothesis $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ if with the sample mean $\overline{x}$ the value of $z = \frac{\overline{x} - \mu_0}{\sigma}$ is either smaller than the $\alpha/2$-quantile or larger than the (1-$\alpha/2$)-quantile of the standard normal distribution is no UMP-test. This can be shown by an example. The test by rejecting the null hypothesis $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ if the value of z is smaller than the α-quantile of the standard normal distribution has a larger power if $\mu > \mu_0$ and this for any α ($0<\alpha<1$). We demonstrate this for $\alpha = 0.05$ by Table 1.

Table 1:

Power $\pi(\mu)$ of the test for $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ for $\alpha = 0.05$, if H_0 is rejected as long as $z < -1.96$ or $z > 1.96$ and power $\pi^*(\mu)$ of the test for $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ for $\alpha = 0.05$, if H_0 is rejected as long as $z > -1.645$. (both tests have the significance level 0.05).

μ	π(μ)	π*(μ)
-1	0.9793	0.9907
-0.6	0.6700	0.7749
-0.2	0.1259	0.1991
0	0.05	0.05
0.2	0.1259	0.0072
0.6	0.6700	0.0000
1	0.9793	0.0000

Of course is the second test biased and nobody would use it but it contradicts the UMP-property of the first test. This test is the UMP-test in the class of all unbiased tests and therefore called a uniformly most powerful unbiased test – UMPU-test.

4. Likelihood ratio tests

If the type of $f(x;\theta)$ is known and considered as a function of θ for observed x it is called a likelihood function. Then a likelihood ratio test can be constructed which has some advantages (good properties) such as often being a uniformly most powerful (unbiased) test.

If we take a random sample $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$[2] of size n where each $\mathbf{x}_i$ is distributed as $f(x;\theta)$ with realization $(x_1, x_2, ..., x_n)$ a likelihood ratio test is either based on the (realized) test statistic

$$L = \frac{\underset{\theta \in \Omega_0}{\max} \prod_{i=1}^{n} f(x_i, \theta)}{\underset{\theta \in \Omega_A}{\max} \prod_{i=1}^{n} f(x_i, \theta)}$$

or on

[2] Random variables are bold print

$$\Lambda = \frac{\max\limits_{\theta \in \Omega_0} \prod\limits_{i=1}^{n} f(x_i,\theta)}{\max\limits_{\theta \in \Omega} \prod\limits_{i=1}^{n} f(x_i,\theta)}.$$

The possible values of L (or of Λ) are split into two subsets, the acceptance region and the critical region. If the test statistic falls into the acceptance region, the null hypothesis is accepted but otherwise rejected.

So far this is a very general formulation of the statistical test problem. For those not very familiar with this abstract description, we consider the special case that the parameter space is the real line and $t = 1$.

5. Some examples

Examples where the parameter space is the real line are:

Tests about the expectation of a normal distribution with parameter space $\Omega = (-\infty,\infty) \subseteq \Re^1$,

Tests about a probability with parameter space $\Omega = [0,1] \subset \Re^1$,

Tests about the variance of a normal distribution with parameter space $\Omega = [0,\infty) \subset \Re^1$.

Let us base the discussion upon the test about a probability p for some event.

If the null hypothesis states:

$H_0 : p = 0$ the alternative must be $H_A : 0 < p \le 1$ and is called one-sided (or directional).

If the null hypothesis states:

$H_0 : p = 1$ the alternative must be $H_A : 0 \le p < 1$ and is again one-sided.

If the null hypothesis states:

$H_0 : p = p_0; 0 < p_0 < 1$ the alternative must be $H_A : p \ne p_0$ and is called two-sided.

Here the two kinds of error are exactly defined.

One-sided alternatives can also be defined if $0 < p_0 < 1$, for instance as $H_0 : p \le p_0; 0 < p_0 < 1$ then $H_A : p > p_0; 0 < p_0 < 1$ or $H_0 : p \ge p_0; 0 < p_0 < 1$ and $H_A : p < p_0; 0 < p_0 < 1$.

In the case $H_0 : p \le p_0; 0 < p_0 < 1$ the power function is a sigmoid curve starting at 0 for p = 0 and tending to 1 if p = 1. At the value $p = p_0$ it reaches the value α = $\max\limits_{p \le p_0} \alpha(p) = \alpha$ and this is called the significance level.

The problem of an error of the third kind stems from leaving the basic concept of the Neyman-Pearson theory. If we formulate the hypotheses in this way:

$$H_0 : p = p_0 ; 0 < p_0 < 1 ,\ H_A : p > p_0 ,$$

a part of the parameter space belongs to neither the null nor the alternative hypothesis. When we do this, we leave the Neyman-Pearson test theory and switch over to a quite different decision problem, namely the decision between at least three hypotheses. Because most applicants of statistical tests accept the Neyman-Pearson approach, an error of the third kind can not be determined and is not needed at all.

We discuss this situation in the next paragraph.

The complete set of possibilities must in the case above be written:

$$H_0 : p = p_0 ; 0 < p_0 < 1 ,\ H_A^- : p > p_0 , H_A^+ : p < p_0 \text{ -}$$

6. The three-decision problem and an error of the third kind

Leventhal and Huynh's position (1996) that we have to make a decision between three hypotheses does not fit with the Neyman-Pearson test theory, which more generally is a two-decision problem with the statements "*the null hypothesis is accepted*" or "*the null hypothesis is rejected.*" The null hypothesis is some statement about a probability distribution, and the alternative hypothesis is the statement that the null hypothesis is wrong. When we are interested in larger parameter values (as in Mosteller, 1948) we nowadays prefer the selection procedure of Bechhofer (1954) as described in Rasch, Kubinger and Yanagida (2011). Selecting the best out of *a* distributions (populations) then means to select the population with the largest means. Selection procedures have only one risk, namely that of selecting (based on samples) a population incorrectly as the best one.

We first consider the general situation of a three-decision problem before we – for simplicity – again go back to probabilities.

We split the parameter space in three non-overlapping subsets Ω_0, Ω_1 and Ω_2 [3].

Then we can formulate three hypotheses:

$$H_0 : \theta \in \Omega_0 \subset \Omega ,$$

$$H_1 : \theta \in \Omega_1 \subset \Omega$$

$$H_2 : \theta \in \Omega_2 \subset \Omega .$$

[3] $\Omega_0 \cap \Omega_1 = \varnothing, \Omega_0 \cap \Omega_2 = \varnothing, \Omega_1 \cap \Omega_2 = \varnothing; \Omega_0 \cup \Omega_1 \cup \Omega_2 = \Omega$

We now can choose one of three decisions, namely:

$D_0 : \theta \in \Omega_0 \subset \Omega$,

$D_1 : \theta \in \Omega_1 \subset \Omega$

$D_2 : \theta \in \Omega_2 \subset \Omega$.

Each of these three decisions may either be correct or wrong as it is shown in the following Table 2.

Table 2:
Consequences of three decisions under three true situations

		True situation		
		$\theta \in \Omega_0 \subset \Omega$	$\theta \in \Omega_1 \subset \Omega$	$\theta \in \Omega_2 \subset \Omega$
Decision	$D_0 : \theta \in \Omega_0 \subset \Omega$	Correct decision	Error A of the second kind	Error B of the second kind
	$D_1 : \theta \in \Omega_1 \subset \Omega$	Error A of the first kind	Correct decision	Error A of the third kind
	$D_2 : \theta \in \Omega_2 \subset \Omega$	Error B of the first kind	Error B of the third kind	Correct decision

We see that in a three-decision problem six kinds of errors are possible.

Before we discuss this in more detail we go back to our special case with tests about probabilities.

Here our null hypothesis is

$H_0 : p = p_0 = \Omega_0 ; 0 < p_0 < 1$ and the two other options are

$H_1 : p < p_0 = \Omega_1 ; 0 < p_0 < 1$

$H_2 : p > p_0 = \Omega_2 ; 0 < p_0 < 1$.

In Table 3 the consequences of the three decisions about a probability can be found.

If we accept H_0 (first row in Table 3) but erroneously because one of the two other options is correct ($p < p_0$ or $p > p_0$), then we have an error of the second kind (A or B) in analogy to hypothesis testing in the Neyman-Pearson approach. If we accept H_1 and H_0 is correct, we have a type of an error of the first kind. But if we accept H_1 and H_2 is correct, we get an error which could be called an error of the third kind. If we accept H_2 and H_0 is correct, we have another type of an error of the first kind. The concept of an error of the third kind is reasonable if we have to make decisions between three options; one of them is called the null hypothesis H_0 .

Table 3:
Consequences of three decisions under three true situations for a probability

		True situation		
		$p = p_0; 0 < p_0 < 1$	$p < p_0; 0 < p_0 < 1$	$p > p_0; 0 < p_0 < 1$
Decision	$D_0 : p = p_0$	Correct decision	Error A of the second kind	Error B of the second kind
	$D_1 : p < p_0$	Error A of the first kind	Correct decision	Error A of the third kind
	$D_2 : p > p_0$	Error B of the first kind	Error B of the third kind	Correct decision

There is no need to change the definition of the power and in connection with the discussion of UMPU-tests the original definition is needed. It is still the probability of rejecting H_0 as a function of θ in the whole parameter space and in our examples of p over the interval [0,1]. Modified power definitions as in Leventhal., and Huynh (1996), MacDonald.(1999), Mendes (2004), and Mendes, M. (2007) stem from a definition of the power function (see section 3) contrary to Lehmann and Romano and general to that in mathematical statistics.

7. Conclusion

As long as we base statistical inference on the Neyman-Pearson test theory, an error of the third kind cannot occur. When we consider a problem of decision between three options where one of them is called the H_0, besides errors of the first and the second kinds, another class of errors is possible which we can call errors of the third kind. The definition of power function is in all cases the same, namely the probability of rejecting H_0. When we have to make a decision between four and more options the number of kinds of errors is increasing rapidly.

8. Acknowledgement

The author thanks two referees for helpful comments.

9. References

Camdeviren, H., & Mendes, M. (2005). A Simulation Study for Type III Error Rates of Some Variance Homogeneity Tests. *Pakistan Journal of Statistics, 21*(2), 223-234.

Hopkins, B. (1973). Educational Research and Type III Errors. *The Journal of Experimental Education, 41,* 31-32.

Kimball, A. W. (1957). Errors of the Third Kind in Statistical Consulting. *Journal of the American Statistical Association, 52,* 133-142.

Kimmel, H. D. (1957). Three Criteria for the Use of One-Tailed Tests. *Psychological Bulletin, 54,* 351-353.

Kubinger, K. D, Rasch, D., & Yanagida, T. (2011). *Statistik in der Psychologie – vom Einführungskurs bis zur Dissertation* [Statistics in Psychology]. Göttingen: Hogrefe.

Lehmann, E. L., & Romano, J. P. (2005). *Testing Statistical Hypotheses*. New York: Springer.

Leventhal, L., & Huynh, C. L. (1996). Directional Decisions for Two-Tailed Tests: Power, Error Rates and Sample Size. *Psychological Methods, 1*(3), 278-292.

MacDonald, P. (1999). Power, Type I, and Type III Error Rates of Parametric and Nonparametric Statistical Tests. *The Journal of Experimental Education, 67,* 367-379.

Mendes, M. (2004). The Comparison of ANOVA F and K-Statistic (KANOVA) with Respect to Type III Error Rate. *Journal of Agriculture Sciences, 10*(2), 1-4.

Mendes, M. (2007). The Effect of Non-Normality on Type III Error for Comparing Independent Means. *Journal of Applied Quantitative Methods, 2*(4), 444-454.

Mosteller, F. (1948). A k-sample slippage test for an extreme population. *Annals of Mathematical Statististics, 19,* 58-65

Raiffa, H. (1968). *Decision analysis*. Reading, MA: Addison-Wesley.

Rasch, D., Kubinger, K. D., & Yanagida, T. (2012). Statistics in Psychology using R and SPSS. Chichester: Wiley.

Rasch, D., Tiku, M. L., & Sumpf, D. (1994). *Elsevier´s Dictionary of Biometry*. Amsterdam: Elsevier.

10. Appendix: Fundamentals of set theory

Set theory begins with a fundamental relation between an object and a set. In our case the objects are parameters *θ.* If *θ* is an element of Ω, we write $\theta \in \Omega$. A derived relation between two sets is the subset relation, also called set inclusion. If all the members of set *A* are also members of set *B*, then *A* is a subset of *B*, denoted $A \subseteq B$. From this definition, it is clear that a set is a subset of itself. Above we find $\Omega \subseteq \Re^t$ what means that Ω is a part of the *t*-dimensional Euclidian space $\Re^t$ or this space itself like in the case $\Omega = \left(-\infty,\infty\right) \subseteq \Re^1 = \Re^1$ but in $\Omega = \left[0,1\right] \subset \Re^1$ the interval [0,1] is a proper subspace of $\left(-\infty,\infty\right) = \Re^1$. A proper subset *A* of *B* $A \subset B$ means then at least one element of *B* may not belong to *A*. Another example is $\Omega_0 \subset \Omega$. A set, containing no element is called an empty set *∅.* Just as in arithmetic set theory features operations on sets . The union of the sets *A* and *B*, denoted $A \cup B$, is the set of all objects that are a member of *A*, or *B*, or both. In our article we find $;\Omega_0 \cup \Omega_1 \cup \Omega_2 = \Omega$ what means that any element in *Ω* either belongs to $;\Omega_0, \Omega_1$ or Ω_2 . The intersection of the sets *A* and *B*, denoted *A* ∩ *B*, is the set of all objects that are members of both *A* and *B*. Above we find

$\Omega_0 \cap \Omega_1 = \varnothing, \Omega_0 \cap \Omega_2 = \varnothing, \Omega_1 \cap \Omega_2 = \varnothing$ what means that all pairs have empty intersections (they are non-overlapping).

A set difference of *B* and *A*, denoted $B \setminus A$ is the set of all members of *B* that are not members of *A*. In the expression above $H_A : \theta \in \Omega \setminus \Omega_0 = \Omega_A \subset \Omega$ the set Ω_A is the set difference between Ω and Ω_0.