

Multiple hypotheses for multiple decisions

J.J. Goeman, A. Solari and T. Stijnen

Abstract Many statistical situations really involve a choice between more than two decisions. We approach two classical examples of three-decision problems from the point of view of the partitioning principle. Applied to the null hypothesis specifying a real-valued parameter equal to zero, it will be shown that it is sometimes possible to make an inference on the sign of the parameter of interest, even when the null hypothesis itself could not be rejected. Relationships with confidence intervals are explored.

Key words: Testing statistical hypotheses, Multiple decision problems.

1 Introduction

One of the attractions of formulating statistical problems in terms of hypothesis testing is the resulting structural simplicity. However, at the same time this reduction to a choice between only two decisions frequently causes complications by creating a class of alternatives which combines too many different elements. In many such cases, if one is willing to partition the class of alternatives into its natural components, and to treat them as additional null hypotheses, one obtains a *multiple testing* problem, which admits a simpler and more natural solution than the apparently less complex single testing problem.

For instance, for a single parameter $\theta \in \mathcal{R}$, rejection of a point null hypothesis $\theta = 0$ and a decision like $\theta \neq 0$ is not very informative. We would like to learn

J.J. Goeman

Leiden University Medical Centre, The Netherlands. e-mail: j.j.goeman@lumc.nl

A. Solari

University of Padua and Leiden University Medical Centre. e-mail: solari@stat.unipd.it

T. Stijnen

Leiden University Medical Centre, The Netherlands. e-mail: t.stijnen@lumc.nl

something more about the true θ -value, for instance, whether θ is negative or positive. Then we can partition $\mathcal{R} \setminus \{0\}$ into $\theta < 0$ and $\theta > 0$, and go for simultaneous testing on the null hypotheses $\theta < 0$, $\theta = 0$ and $\theta > 0$. Therefore, we try to reject as many nulls as possible by controlling the familywise error at α and end up with some subset of non rejected θ 's as the final decision, which can be interpreted as a confidence interval on the coarsened parameter space $\{(-\infty, 0), \{0\}, (0, +\infty)\}$.

Undoubtedly, the method of testing each $\theta \in \mathcal{R}$ with a level- α test, which yields a $(1 - \alpha)$ confidence set for θ , is the most powerful method and yields the most precise decision. Therefore, one may argue that there is no necessity for something like the partitioning principle. However, we agree with Finner and Strassburger [1] that such principle is an helpful tool for the construction of *meaningful decisions*. Indeed, one may consider the application of such principles as a first step in finding a decision strategy, and then relationships with confidence intervals could be explored. We show this by constructing the confidence interval associated with the *three-sided testing* procedure [2], and by investigating its relationship with the classical confidence interval. Finally, another example from Lehmann [3] will be given: although of minimal practical importance, it illustrates the crucial dependence of meaningful decisions on the precise formulation of the testing problem.

2 Three-sided testing

A conventional type of hypothesis concerning a single parameter $\theta \in \mathcal{R}$ is

$$\theta = 0 \quad \text{against} \quad \theta \neq 0.$$

However, in many, perhaps most, situations, if the null hypothesis is rejected, we are interested in deciding whether θ is negative or positive. In that case, there are three possible inferences or decisions, and we can formulate three different null hypotheses about θ :

$$\begin{array}{ll} H_0 : \theta = 0 & \text{(zero)} \\ H_+ : \theta > 0 & \text{(positive)} \\ H_- : \theta < 0 & \text{(negative)} \end{array} \quad (1)$$

When making simultaneous inference on H_0 (against $H_- \cup H_+$), H_+ (against $H_- \cup H_0$) and H_- (against $H_0 \cup H_+$), while retaining full type I error control, we must limit the probability of making any false inference among these three hypotheses. This leads naturally to a procedure that controls the familywise error, the probability of making at least one false rejection, at level α . A procedure is said to control the familywise error in the *strong* sense if it controls the familywise error for every possible value of the parameter(s). A very powerful procedure for strong control of

the familywise error is Finner's *partitioning principle* [1]. Application of the partitioning principle requires that the tested null hypotheses partition the parameter space into disjoint regions. If all null hypotheses are disjoint, the true value of the parameter can lie in at most one of the disjoint regions, so that no two hypotheses can be simultaneously true. The partitioning principle says that in this case all tests of all hypotheses can be simultaneously performed at level α , while still controlling the familywise error at level α . The partitioning principle does not prescribe a test for the hypotheses in the partitioning: any level- α test may be used. We choose one-sided tests for H_+ and H_- , and a two-sided test for H_0 .

To be more concrete, we focus on the example of the t -test. In a one-sample situation, suppose we have sampled n i.i.d. observations $\{X_i\}_{i=1}^n$ from a normal distribution $\mathcal{N}(\theta, \sigma^2)$, so that we have an estimate $\bar{X}$ of θ , which has standard error $S/\sqrt{n}$. The composite hypotheses H_0 , H_+ and H_- , which we have to test, are built from point null hypotheses of the form $H_{0,u} : \theta = u$, for every $u \in \mathcal{R}$. Suppose we have a t -test statistic

$$T_u = \sqrt{n} \cdot \frac{\bar{X} - u}{S},$$

for every such hypothesis $H_{0,u}$. Let t_q for $q \in [0, 1]$ be the q th quantile of the Student's t distribution of T_u under $H_{0,u}$. Based on the test statistics T_u and quantiles t_q we can formulate the three-sided testing procedure:

Procedure 1 (Three-sided testing)

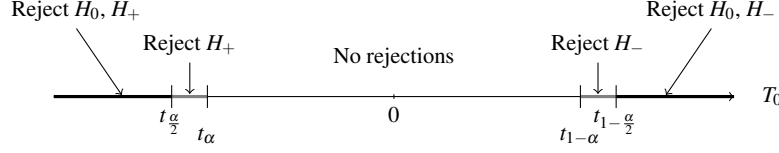
Simultaneously do the following tests:

1. reject H_+ if $T_0 \leq t_\alpha$
2. reject H_- if $T_0 \geq t_{1-\alpha}$, and
3. reject H_0 if either $T_0 \geq t_{1-\alpha/2}$ or if $T_0 \leq t_{\alpha/2}$.

Three-sided testing in this case is essentially the classical test of a point null hypothesis, but with an additional inference on the sign of the parameter of interest θ . Figure 1 shows the inference that results from each realized value of this statistic. The rejection region for H_0 is the familiar rejection region with bounds $t_{\alpha/2}$ and $t_{1-\alpha/2}$. Inference on the sign of θ , through rejection of H_+ or H_- comes for free whenever H_0 is rejected.

The most surprising result, however, is that for $t_{\alpha/2} < T_0 \leq t_\alpha$ we may reject H_+ , inferring that $\theta \leq 0$, even though we are not able to reject H_0 . This additional inference, and its counterpart for $t_{1-\alpha} < T_0 \leq t_{1-\alpha/2}$, are free and can be added to any inference on a point null hypothesis without any additional cost in terms of false positives. The additional inference does, however, complicate the calculation of confidence intervals somewhat.

A confidence interval is an inverted test. To make a confidence interval, we test $H_{0,u} : \theta = u$ for every $u \in \mathcal{R}$, and record which ones are rejected. The collection of values of u for which we failed to reject $H_{0,u}$ constitutes the confidence interval

Fig. 1 Illustration of the rejection regions based on the t-test.

for θ . By definition, every confidence interval defined in this way has the correct coverage probability.

Note that the definition of the confidence interval does not require the test that is done for $H_{0,u}$ for different values of u to be similar in any way. As long as each test for each value of u is a valid level α test, the required $1 - \alpha$ coverage of the confidence interval is guaranteed. If we let the type of test for each u be dictated by the membership of u in H_0 , H_+ or H_- , by choosing right-sided tests for all $u < 0$, left-sided tests for all $u > 0$, and two-sided tests for $u = 0$, the resulting confidence interval is given by

$$l = \begin{cases} \bar{x} + \frac{s}{\sqrt{n}}t_{\alpha} & \text{if } \bar{x} < -\frac{s}{\sqrt{n}}t_{\alpha} \\ 0 \text{ (inclusive)} & \text{if } -\frac{s}{\sqrt{n}}t_{\alpha} \leq \bar{x} \leq -\frac{s}{\sqrt{n}}t_{\alpha/2} \\ 0 & \text{if } -\frac{s}{\sqrt{n}}t_{\alpha/2} < \bar{x} \end{cases}$$

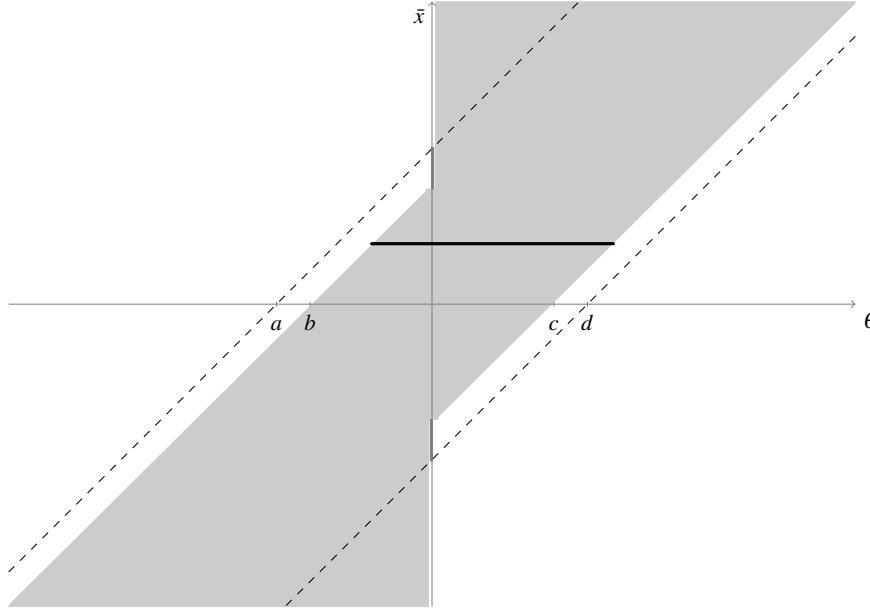
$$u = \begin{cases} 0 & \text{if } \bar{x} < -\frac{s}{\sqrt{n}}t_{1-\alpha/2} \\ 0 \text{ (inclusive)} & \text{if } -\frac{s}{\sqrt{n}}t_{1-\alpha/2} \leq \bar{x} \leq -\frac{s}{\sqrt{n}}t_{1-\alpha} \\ \bar{x} + \frac{s}{\sqrt{n}}t_{1-\alpha} & \text{if } -\frac{s}{\sqrt{n}}t_{1-\alpha} < \bar{x} \end{cases}$$

The lower bound of this confidence interval is never above 0, and the upper bound is never below 0. Note how a change in $\bar{x}$ can change the boundary of the confidence interval at zero from open to closed or vice versa. A graphical illustration of this confidence interval is given in Figure 2.

It is easy to see that the confidence interval described in this section is consistent with three-sided testing of Procedure 1. The three-sided testing procedure is not consistent with the classical confidence interval. It may reject H_+ , for example, even when the classical confidence interval contains positive values of θ . The three-sided testing procedure brings its own confidence interval, which is somewhat peculiar from the classical perspective, but nonetheless a valid confidence interval with the correct coverage probability. This confidence interval is sometimes narrower than the classical one, and sometimes broader.

As a final remark, simultaneous testing on H_0 , H_+ and H_- links closely with the notion of directional errors or “Type III” errors. Directional errors may occur if we do a two-sided test of H_0 , followed up, if H_0 was rejected, by a post hoc inference

Fig. 2 Confidence intervals consistent with three-sided testing (shading), compared with the classical confidence interval (dashed lines), based on the t-test. The thick gray lines indicate the “inclusive” part of the confidence interval. The thick black line illustrates how a confidence interval may be read off the graph for an observed value of $\bar{x}$. Legend: $a = \frac{s}{\sqrt{n}}t_{\alpha/2}$, $b = \frac{s}{\sqrt{n}}t_{\alpha}$, $c = \frac{s}{\sqrt{n}}t_{1-\alpha}$, $d = \frac{s}{\sqrt{n}}t_{1-\alpha/2}$.



of the sign of θ . A directional error occurs if we correctly reject H_0 , but make an incorrect inference on the sign of θ .

3 Lehmann's example

Lehmann [3] described a situation in which the best symmetric test based on average likelihood is the most-powerful test and the best symmetric test based on maximum likelihood is the least-powerful symmetric test.

A more specific illustration of Lehmann's example can be formulated as follows: assume a single observation $0 < X < 1$ from $f(x)$ and consider testing

$$f(x) \equiv 1 \quad \text{versus} \quad f(x) = (1 + \gamma)x^\gamma \quad \text{or} \quad f(x) = (1 + \gamma)(1 - x)^\gamma$$

with $\gamma > 0$ known. But suppose it is important not only to reject the null hypothesis but also to decide in that case whether $f(x)$ is the increasing one or the decreasing

one. Again, there are three possible decisions, and we can formulate three different null hypotheses about $f(x)$:

$$\begin{aligned} H_{=} &: f(x) \equiv 1 && \text{(uniform)} \\ H_{\uparrow} &: f(x) = (1 + \gamma)x^\gamma && \text{(increasing)} \\ H_{\downarrow} &: f(x) = (1 + \gamma)(1 - x)^\gamma && \text{(decreasing)} \end{aligned} \quad (2)$$

Let $f_{=}$, $f_{\downarrow}$ and $f_{\uparrow}$ denote $f(x)$ under $H_{=}$, $H_{\downarrow}$ and $H_{\uparrow}$, respectively. Lehmann [3] showed that when $\gamma > 1$ ($f_{\downarrow}$ and $f_{\uparrow}$ convex), the most-powerful symmetric test of $H_{=}$ against $H_{\uparrow} \cup H_{\downarrow}$ is an extreme-tails test based on both maximum and average likelihood criteria, i.e.

$$\text{reject } H_{=} \text{ if } 0 \leq x \leq \alpha/2 \text{ or } 1 - \alpha/2 \leq x \leq 1$$

When $0 < \gamma < 1$ ($f_{\downarrow}$ and $f_{\uparrow}$ concave), the most-powerful symmetric test of $H_{=}$ against $H_{\uparrow} \cup H_{\downarrow}$ is a central test based the average likelihood criterion, i.e.

$$\text{reject } H_{=} \text{ if } .5 - \alpha/2 \leq x \leq .5 + \alpha/2$$

and that the least powerful symmetric test is the extreme-tails test based on the maximum likelihood ratio criterion. For testing $H_{\downarrow}$ against $H_{=} \cup H_{\uparrow}$ or $H_{\uparrow}$ against $H_{=} \cup H_{\downarrow}$, an one-tail test is the most powerful test:

$$\begin{aligned} \text{reject } H_{\downarrow} &\text{ if } 1 - \alpha^{1/(\gamma+1)} \leq x \leq 1 && \text{(right - tail test)} \\ \text{reject } H_{\uparrow} &\text{ if } 0 \leq x \leq \alpha^{1/(\gamma+1)} && \text{(left - tail test)} \end{aligned}$$

We can formulate a procedure, which depends on whether $\gamma < 1$ or $\gamma > 1$:

Procedure 2 ($\gamma > 0$ known)

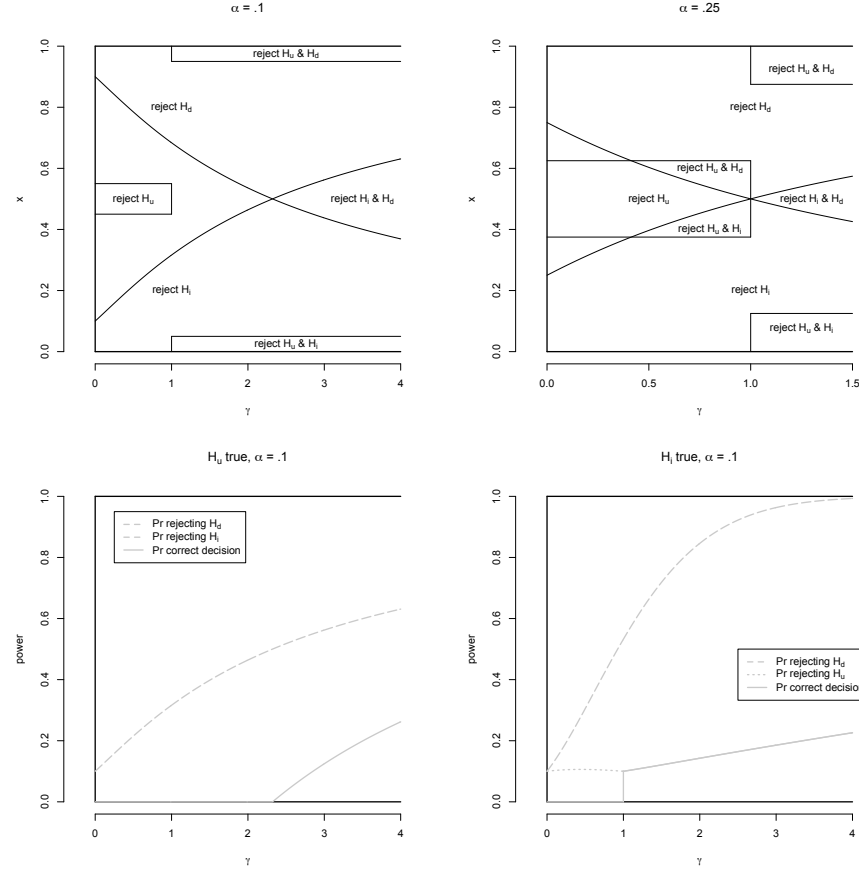
Simultaneously do the following tests:

1. $\text{reject } H_{=} \text{ if } \begin{cases} .5 - \alpha/2 \leq x \leq .5 + \alpha/2 & \text{when } 0 < \gamma < 1 \\ 0 \leq x \leq \alpha/2 \text{ or } 1 - \alpha/2 \leq x \leq 1 & \text{when } \gamma > 1 \end{cases}$
2. $\text{reject } H_{\downarrow} \text{ if } 1 - \alpha^{1/(\gamma+1)} \leq x \leq 1$, and
3. $\text{reject } H_{\uparrow} \text{ if } 0 \leq x \leq \alpha^{1/(\gamma+1)}$.

We consider the following generalizations of power for multiple testing: (b) probability of rejecting a particular false hypothesis, and (b) probability of rejecting all false hypotheses. To see the effects of the procedure, we summarize the rejection regions and powers as a function of θ and with $\alpha = .1$ and $\alpha = .25$ in Figure 3. Note that the probability of rejecting a true hypothesis is smaller than the probability of rejecting a false hypothesis, i.e. each test in (2) is unbiased. However, the

probability of making the correct inference could be lower than α . Note also that for $\alpha > .25$ or γ very large, it could be possible to reject all null hypotheses.

Fig. 3 Rejection regions with $\alpha = .1$ and $\alpha = .25$ and power functions with $\alpha = .1$ when either $H_{=}$ or $H_{\uparrow}$ is true.



Shaffer [4] illustrated the divergence between test procedures based on two-sided ($H_{=}$ vs $H_{\downarrow} \cup H_{\uparrow}$) and *directional hypothesis-pair* ($H_{=} \cup H_{\downarrow}$ vs $H_{\uparrow}$ and $H_{=} \cup H_{\uparrow}$ vs $H_{\downarrow}$) formulations when $0 < \gamma < 1$. However, the directional hypothesis-pair procedure is designed for controlling Type I error at α when $H_{=}$ is true and Type III error at α when $H_{=}$ is false, but the resulting inference is either of no rejections, reject both $H_{=}$ and $H_{\downarrow}$, or reject both $H_{=}$ and $H_{\uparrow}$.

Acknowledgements The work of AS was supported by the EX 60% research grant 60A15-7791/09. We are grateful to the Department of Medical Statistics and Bioinformatics of the LUMC for support, advice and assistance.

References

1. Finner, H, Strassburger, K.: The Partitioning Principle: A Powerful Tool in Multiple Decision Theory. *The Annals of Statistics*. **30**, 1194–1213. (2002)
2. Goeman, J.J., Solari, A.P., Stijnen, T.: Three-sided Hypothesis Testing: Simultaneous Testing of Superiority, Equivalence and Inferiority. Preprint available from <http://www.msbi.nl/gorman>. Cited 15 Sep 2009
3. Lehmann, E.L.: On likelihood ratio tests. *IMS Lecture Notes - Monograph Series*. 2nd Lehmann Symposium Optimality **49**, 1–8. (2006)
4. Shaffer, J.P.: Recent developments towards optimality in multiple hypothesis testing. *IMS Lecture Notes - Monograph Series*. 2nd Lehmann Symposium Optimality **49**, 1–8. (2006)