# New relevance and significance measures to replace p-values

Werner A. Stahel[1*]

**1** Seminar for Statistics, ETH, Zurich, Switzerland
* stahel@stat.math.ethz.ch

## Abstract

The p-value has been debated exorbitantly in the last decades, experiencing fierce critique, but also finding some advocates. The fundamental issue with its misleading interpretation stems from its common use for testing the unrealistic null hypothesis of an effect that is precisely zero. A meaningful question asks instead whether the effect is *relevant.* It is then unavoidable that a threshold for relevance is chosen. Considerations that can lead to agreeable conventions for this choice are presented for several commonly used statistical situations. Based on the threshold, a simple quantitative measure of relevance emerges naturally. Statistical inference for the effect should be based on the confidence interval for the relevance measure. A classification of results that goes beyond a simple distinction like "significant / non-significant" is proposed. On the other hand, if desired, a single number called the "secured relevance" may summarize the result, like the p-value does it, but with a scientifically meaningful interpretation.

## 1 Introduction

The p-value is arguably the most used and most controversial concept of applied statistics. Blume *et al.* [1] summarize the shoreless debate about its flaws as follows: "Recurring themes include the difference between statistical and scientific significance, the routine misinterpretation of non-significant p-values, the unrealistic nature of a point null hypothesis, and the challenges with multiple comparisons." They nicely collect 14 citations, and I refrain from repeating their introduction here, but complement the analysis of the problem and propose a solution that both simplifies and extends their's.

The basic cause of the notorious lack of reliability of empirical research, notably in parts of social and medical science, can be found in the failure to ask scientific questions in a sufficiently explicit form, and the p-value problem is intrinsically tied to this flaw. Here is my argument.

Most empirical studies focus on the effect of some treatment, expressed as the difference of a target variable between groups, or on the relationship between two or more variables, often expressed with a regression model. Inferential statistics needs a probabilistic model that describes the scientific question. Usually, this is a parametric model in which the effect of interest appears as a parameter. The question is then typically specified as: "Can we prove that the effect is not zero?"

**The Zero Hypothesis Testing Paradox.** This is, however, not a scientifically meaningful question. When a study is undertaken to find some difference between groups or some influence between variables, the *true* effect—e.g., the difference between two within group expected values—will never be precisely zero. Therefore, the strawman null hypothesis of zero true effect (the "zero hypothesis") could in almost all reasonable applications be rejected if one had the patience and resources to obtain enough observations. Consequently,

the question that is answered mutates to: "Did we produce sufficiently many observations to prove the (alternative) hypothesis that was true on an apriori basis?" This does not seem to be a fascinating task. I call this argument the "Zero Hypothesis Testing Paradox." The problem with the p-value is thus that it is the output of testing an unrealistic null hypothesis and thereby answers a nonsensical scientific question. (Note that the proposal to lower the testing level from $5\,\%$ to $0.5\,\%$ by Benjamin *et al.* [2] is of no help in this respect.)

A sound question about an effect is whether it is large enough to be *relevant*. In other words: Without the specification of a threshold of relevance, the scientific question is void.

Scientists have gladly avoided the determination of such a threshold, because they felt that it would be arbitrary, and have jumped on the train of "Null Hypothesis Significance Testing," that was offered cheaply by statistics. Let us be clear: Avoiding the choice of a relevance threshold means avoiding a scientifically meaningful question.

Given the relevance threshold, the well-known procedures can be applied not only for testing the null hypothesis that the effect is larger than the threshold against the alternative that it is smaller, but also vice versa, proving statistically that the effect is negligible. The result can of course also be ambiguous, meaning that the estimate is neither significantly larger nor smaller than the threshold. I introduce a finer distinction of cases in Section 2.3.

These ideas are well-known under the heading of equivalence testing, and similar approaches have been advocated in connection with the p-value problem, like the "Two One-Sided Tests (TOST)" of Lakens [3], the "Second Generation p-value (SGPV)" by Blume *et al.* [1], or the "Minimum Effect Size plus p-value (MESP)" by Goodman *et al.* [4]. The threshold has been labelled "Smallest Effect Size Of Interest (SESOI)" or "Minimum Practically Significant Distance (MPSD)." I come back to these concepts in Section 2.2.

Using confidence intervals instead of p-values or even "yes-no" results of null hypothesis tests provides the preferable, well-known alternative to null hypothesis testing for drawing adequate inference. Each reader can then judge a result by checking if his or her own threshold of relevance is contained in the interval. Providing confidence intervals routinely would have gone a long way to solving the problem. I come back to this issue in the Discussion (Section 6).

Most probably, the preference to present p-values rather than confidence intervals is due to the latter's slightly more complicated nature. In their usual form, they are given by two numbers that are not directly comparable between applications. I will define a single number, which I call "significance," that characterizes the essence of the confidence interval in a simple and informative way.

In "ancient" times, before the computer produced p-values readily, statisticians examined the test statistics and then compared them to tables of "critical values." In the widespread case that the t test was concerned, they used the t statistic as an informal quantitative measure of significance of an effect by comparing it to the number 2, which is approximately the critical value for moderate to large numbers of degrees of freedom. This will also shine up in the proposed significance measure.

Along the same line of thought, a simple measure of relevance will be introduced. It compares the estimated effect with the relevance threshold. The respective confidence interval is used to distinguish the cases mentioned above, and a single value can be used to characterize the result with the same simplicity as the p-value does it, but with a much more informative interpretation.

# 2 Definitions

The simplest case for statistical inference is the estimation of a constant based on a sample of normal observations. It directly applies to the estimation of a difference between two treatments using paired observations. I introduce the new concepts first for this situation. The problem of assessing a general parameter as well as the application of the concepts for

typical situations—comparison of two or more samples, estimation of proportions, regression and correlation—will be discussed in Section 3.

## 2.1 The generic case

Consider a sample of $n$ statistically independent observations $Y_i$ with a normal distribution,

$$Y_i \sim \mathcal{N}(\vartheta, \sigma^2) \ . \tag{1}$$

The interest is in knowing whether $\vartheta$ is different from $0$ in a relevant manner, where relevance is determined by the relevance threshold $\zeta > 0$. Thus, I want to summarize the evidence for the hypotheses

$$H_0: \ \vartheta \leq \zeta \ , \qquad H_1: \ \vartheta > \zeta \ .$$

(The symbol $\zeta$, pronounced "zeta," delimits the "zero" hypothesis.)

**One sided.** I consider a one-sided hypothesis here. In practice, only one direction of the effect is usually plausible and/or of interest. Even if this is not the case, the conclusion drawn will be one-sided: If the estimate turns out to be significant according to the two-sided test for 0 effect, then nobody will conclude that "the effect is different from zero, but we do not know whether it is positive or negative." Therefore, in reality, two one-sided tests are conducted, and technically speaking, a Bonferroni correction is applied by using the level $\alpha/2 = 0.025$ for each of them. Thus, I treat the one-sided hypothesis and use this testing level.

The point estimate and confidence interval are

$$\widehat{\vartheta} = \overline{Y} = \tfrac{1}{n}\textstyle\sum_i Y_i \ , \qquad \mathsf{CI}_\vartheta = \widehat{\vartheta} \pm \widehat{\omega} \ , \qquad \widehat{\omega} = q\sqrt{\widehat{V}/n} \ , \tag{2}$$

where $\widehat{V}$ is the empirical variance of the sample, $\widehat{V} = \frac{1}{n-1}\sum_i (Y_i - \overline{Y})^2$, and $q$ is the $1 - \alpha/2 = 0.975$ quantile of the appropriate $t$ distribution. Thus, $\widehat{\omega}$ is half the width of the confidence interval and equals the standard error, multiplied by the quantile.

In general problems involving a single effect parameter, the estimated effect usually follows approximately a normal distribution, and these concepts are easily generalized, see Section 3.

**Significance.** The proposed significance measure compares the difference between the estimated effect and the relevance threshold with the half width of the confidence interval,

$$\mathsf{Sig}_\zeta = (\widehat{\vartheta} - \zeta)/\widehat{\omega} \ . \tag{3}$$

The effect is statistically significantly larger than the threshold if and only if $\mathsf{Sig}_\zeta > 1$.

Significance can also be calculated for the common test for zero effect, $\mathsf{Sig}_0 = \widehat{\vartheta}/\widehat{\omega}$. This quantity can be listed in computer output in the same manner as the p-value is given in today's programs, without a requirement to specify $\zeta$. It is much easier to interpret than the p-value, since it is, for a given precision expressed by $\widehat{\omega}$, proportional to the estimated effect $\widehat{\vartheta}$. Furthermore, a standardized version of the confidence interval for the effect is $\mathsf{Sig}_0 \pm 1$,

$$\mathsf{Sig}_0 \pm 1 = \widehat{\omega}\ \mathsf{CI}_\vartheta \ , \qquad \mathsf{CI}_\vartheta = \widehat{\vartheta}\big(1 \pm 1/\mathsf{Sig}_0\big) \ .$$

Nevertheless, it should be clear from the Introduction that $\mathsf{Sig}_0$ should only be used with extreme caution, since it does not reflect relevance.

**Relevance.** An extremely simple and intuitive quantitative measure of relevance is the effect, expressed in $\zeta$ units, $\mathsf{RI} = \vartheta/\zeta$. Its point and interval estimates are

$$\mathsf{RIe} = \widehat{\vartheta}/\zeta \ , \qquad \mathsf{CI}_{\mathsf{RI}} = \mathsf{CI}_{\vartheta}/\zeta \ . \tag{4}$$

I also introduce the "secured relevance" as the lower end of the confidence interval,

$$\mathsf{RIs} = \mathsf{RIe} - \widehat{\omega}^* \ , \qquad \widehat{\omega}^* = \widehat{\omega}/\zeta$$

and the "potential relevance" $\mathsf{RIp} = \mathsf{RIe} + \widehat{\omega}^*$. The effect is called relevant if $\mathsf{RIs} > 1$, that is, if the estimated effect is significantly larger than the threshold.

The estimated relevance RIe is related to $\mathsf{Sig}_\zeta$ by

$$\mathsf{Sig}_\zeta = (\mathsf{RIe} - 1)/\widehat{\omega}^* \ , \qquad \mathsf{RIe} = \mathsf{Sig}_\zeta\, \widehat{\omega}^* + 1 \ .$$

Fig 2 shows several cases of relations between the confidence interval and the effects $0$ and $\zeta$, which can be translated into categories that help interpret results, see Section 2.3.

**Example: Student's sleep data.** Student [5] illustrated his t-test with data measuring the extra sleep evoked by a sleep enhancing drug in 10 patients. The numbers in minutes are $-6$, $6$, $48$, $66$, $96$, $114$, $204$, $264$, $276$, $330$. Their mean is $\widehat{\vartheta} = \overline{Y} = 140$. The p-value for testing the hypothesis of no prolongation is $0.5\,\%$ and the confidence interval extends from $54$ to $226$. The zero significance is obtained from $V = 14,432$, $n = 10$ and $q = 2.26$ with $\widehat{\omega} = 2.26\sqrt{14,432/10} = 86$ as $\mathsf{Sig0} = 140/86 = 1.63$.

If the relevance threshold is one hour, $\zeta = 60$, of extra sleep then $\mathsf{Sig}_\zeta = 80/86 = 0.93$, and the gain is not significantly relevant. This is also seen when calculating the relevance and its confidence interval, $\mathsf{RIe} = 140/60 = 2.33$ and $\mathsf{RIs} = 2.33 - 86/60 = 54/60 = 0.90$, $\mathsf{RIp} = 2.33 + 86/60 = 226/60 = 3.76$. It remains therefore unclear whether the sleep prolongation is relevant. Fig 1 shows the results graphically.

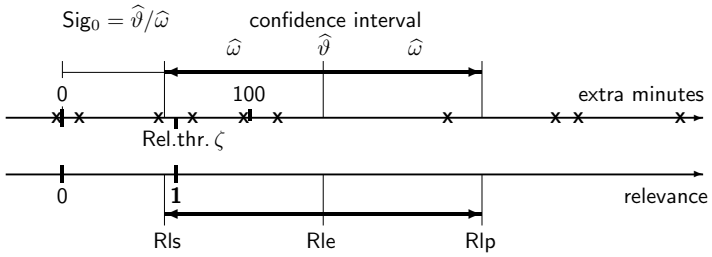


**Fig 1. Estimate, confidence interval and relevance for the sleep data**

## 2.2 Related concepts

**Two one-sided tests (TOST).** Lakens [3] focusses on testing for a negligible effect, advocating the paradigm of equivalence testing. He considers an interval of values that are negligibly different from the point null hypothesis, also called a "thick" or "interval null" [4], [1]. If this interval is denoted as $|\vartheta| \le \zeta$, there is a significantly negligible effect if both hypotheses $\vartheta > \zeta$ and $\vartheta < -\zeta$ are rejected using a one-sided test for each of them. A respective p-value is the larger of the p-values for the two tests.

I have argued for a one-sided view of the scientific problem. With this perspective, the idea reduces to the *one* one-sided test for a negligible effect with significance measure $-\mathsf{Sig}_\zeta$.

**Second Generation P-Value.** The "Second Generation P-Value" SGPV $P_\delta$ has been introduced by Blume *et al.* [1, 6]. In the present notation, $\zeta$ is their $\delta$. The definition of $P_\zeta$ starts from considering the length $O$ of the overlap of the confidence interval with the interval defined by the composite null hypothesis $H_0$. Assume first that $\widehat{\vartheta} > 0$. Then, the overlap measures $O = 2\widehat{\omega}$ if the confidence interval contains the "null interval," that is, if $\widehat{\vartheta} + \widehat{\omega} < \zeta$, and otherwise, $O = \zeta - (\widehat{\vartheta} - \widehat{\omega})$, or 0 if this is negative.

The definition of $P_\zeta$ distinguishes two cases based on comparing $\widehat{\omega}$ to the threshold $\zeta$. If $\widehat{\omega} < 2\zeta$, $P_\zeta = 0$ if there is no overlap, and $P_\zeta = 1$ for complete overlap, $O = 2\widehat{\omega}$. In between, the SGPV is the overlap, compared to the length of the confidence interval,

$$P_\zeta = \frac{O}{2\widehat{\omega}} = \frac{\zeta - (\widehat{\vartheta} - \widehat{\omega})}{2\widehat{\omega}} = \frac{\zeta - \widehat{\vartheta}}{2\widehat{\omega}} + \tfrac{1}{2} = \tfrac{1}{2}\left(1 - \mathsf{Sig}_\zeta\right) \ .$$

In this case, then, $P_\zeta$ is a rescaled, mirrored, and truncated version of the significance at $\zeta$.

Here, I have neglected a complication that arises when the confidence interval covers values below $-\zeta$. The definition of $P_\zeta$ starts from a two-sided formulaton of the problem, $H_0: \ |\vartheta| < \zeta$. Then, the confidence interval can also cover values below $-\zeta$. In this case, the overlap decreases and $P_\zeta$ changes accordingly.

The definition of $P_\zeta$ changes if the confidence interval is too large, specifically, if its length exceeds $2\zeta$. This comes again from the fact that it was introduced with the two-sided problem in mind. In order to avoid small values of $P_\zeta$ caused by a large denominator $2\widehat{\omega}$ in this case, the length of the overlap $O$ is divided by twice the length $2\zeta$ of the "null interval," instead of the length of the confidence interval, $2\widehat{\omega}$, $P_\zeta = O/(4\zeta)$. Then, $P_\zeta$ has a maximum value of $1/2$, which is a deliberate consequence of the definition, as this value does not suggest a "proof" of $H_0$. For a comparison of the SGPV with TOST, see [7].

If the overlap is empty, $P_\zeta = 0$. In this case, the concept of SGPV is supplemented with the notion of the "$\delta$ gap,"

$$\mathsf{Gap}_\zeta = (\widehat{\vartheta} - \zeta)/\zeta = \mathsf{Rle} - 1 \ .$$

Since the significance and relevance measures are closely related to the Second Generation P-Value and the $\delta$ gap, one might ask why still new measures should be introduced. Here is why:

- An explicit motivation for the SGPV was that it should resemble the traditional p-value by being restriced to the 0-1 interval. I find this quite undesirable, as it perpetuates the misinterpretation of $P$ as a probability. Even worse, the new concept is further removed from such an interpretation than the old one, for which the problem "Find a correct statement including the terms p-value and probability" still has a (rather abstract) solution.

- The new p-value was constructed to share with the classical one the property that small values signal a large effect. This is a counter-intuitive aspect that leads to confusion for all beginners in statistics. In contrast, larger effects lead to larger significance (and, of course, larger relevance).

- Taking these arguments together, the problems with the p-value are severe enough to prefer a new concept with a new name and more direct and intuitive interpretation rather than advocating a new version of p-value that will be confused with the traditional one.

- The definition of the SGPV is unnecessarily complicated, since it is intended to correspond to the two-sided testing problem, and only quantifies the undesirable case of ambiguous results. It deliberately avoids to quantify the strength of evidence in the two cases in which either $H_0$ or $H_1$ is accepted.

## 2.3 Classification of results

There is a wide consensus that statistical inference should *not* be reported simply as "significant" or "non-significant." Nevertheless, communication needs words. I therefore propose to distiguish the cases that the effect is shown to be relevant (Rlv), that is, $H_1 : \vartheta > \zeta$ is "statistically proven," or negligible (Ngl), that is, $H_0 : \vartheta \leq \zeta$ is proven, or the result is ambiguous (Amb), based on the significance measure $\mathrm{Sig}_\zeta$ or on the secured and potential relevance Rls and Rlp (Rls $> 1$ for Rlv, Rlp $< 1$ for Ngl and Rls $\leq 1 \leq$ Rlp for Amb).

For a finer classification, the significance for a zero effect, $\mathrm{Sig}_0$, is also taken into account. This may even lead to a contradiction (Ctr) if the estimated effect is significantly negative. Fig 2 shows the different cases with corresponding typical confidence intervals, and Table 1 lists the respective significance and relevance ranges. Similar figures have appeared in [1, Fig. 2] and [4, Fig. 1] and before, with different interpretations.

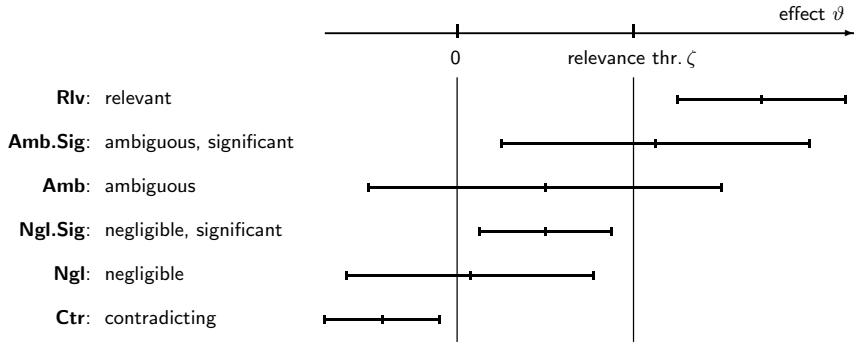


**Fig 2. Classification of cases based on a confidence interval and a relevance threshold**

**Table 1.** Classification of cases defined by ranges of significance and relevance measures. $s$ and $r$ are the place holders for the column headings.

| Case | $\mathrm{Sig}_0$ | $\mathrm{Sig}_\zeta$ | Rls | Rlp |
|---|---|---|---|---|
| Rlv | $s >> 1$ | $s > 1$ | $r > 1$ | $r >> 1$ |
| Amb.Sig | $s > 1$ | $-1 < s < 1$ | $0 < r < 1$ | $r > 1$ |
| Amb | $-1 < s < 1$ | $-1 < s < 1$ | $r < 0$ | $r > 1$ |
| Ngl.Sig | $s > 1$ | $s < -1$ | $0 < r < 1$ | $0 < r < 1$ |
| Ngl | $-1 < s < 1$ | $s < -1$ | $r < 0$ | $0 < r < 1$ |
| Ctr | $s < -1$ | $s << -1$ | $r << 0$ | $r < 0$ |

# 3 Generalization and more models

## 3.1 General model and two-sample problem

Let us now discuss a general parametric model. To make the notation transparent, the two-sample problem is discussed in parallel as an example.

Consider $n$ statistically independent observations following the parametric model

$$\underline{Y}_i \sim \mathcal{F}\big(\underline{\theta}, \underline{\phi}_i; \underline{x}_i\big) \ , \tag{5}$$

where $\underline{\theta}$ is the parameter of interest, $\underline{\phi}_i$ denotes nuisance parameters, and the distribution $\mathcal{F}$ may vary between observations depending on covariates $\underline{x}_i$. These variables may be multidimensional.

The model for comparing two treatments arises when $x_i = 1$ if observation $i$ received treatment 1, and $x_i = 0$ otherwise; $\theta$ is the difference of expected values between the two groups; and the nuisance parameters are the expected value $\phi^{(1)} = \mu_0$ of $Y_i$ for treatment $k = 0$ and the standard deviation of the observations, $\phi^{(2)} = \sigma$. Then,

$$Y_i \sim \mathcal{N}\big(\mu_0 + \theta x_i \, , \sigma^2\big) \ .$$

The problem is to draw inference about the effect $\underline{\theta}$. There is a "null value" $\underline{\theta}_0$ and a threshold $\zeta$ for a relevant effect. For ease of notation, assume $\zeta > 0$.

Inference is based on an estimator $\widehat{\underline{\theta}}$ of $\underline{\theta}$. Assume that its distribution is approximately (multivariate) normal,

$$\widehat{\underline{\theta}} \approx\sim \mathcal{N}_p(\underline{\theta}, \mathbf{V}/n) \ , \tag{6}$$

where the "single observation" variance-covariance matrix $\mathbf{V}$ may depend on all nuisance parameters $\underline{\phi}_i$ and design vectors $\underline{x}_i$, $i = 1, ..., n$, and $p$ is the dimension of $\underline{\theta}$. It may also depend on the parameter of interest, $\underline{\theta}$, but this case needs additional discussion. These assumptions usually hold for the Maximum Likelihood Estimator of $[\underline{\theta}, \underline{\phi}]$, $\mathbf{V}$ being the "$\theta$ part" of the inverse Fisher Information of a single observation.

In the two samples problem with $n_0$ observations in group $k = 0$ and $n_1$, in group $k = 1$,

$$\begin{aligned} \widehat{\theta} &= \tfrac{1}{n_1}\textstyle\sum_i Y_i x_i - \tfrac{1}{n_0}\sum_i Y_i(1 - x_i) \\ V &= (1/\nu_0 + 1/\nu_1)\,\sigma^2 \, , \qquad \nu_k = n_k/n \ . \end{aligned}$$

**Effect scale.** In several models, it appears useful to consider a transformed version of the parameter of interest as the effect, since the transformation leads to a more generally interpretable measure and may have more appealing properties, as in the next subsection. Therefore, the original parameter of interest is denoted as $\theta$ or as popular in the model, and the transformed version will be considered as the effect, $\vartheta = g(\theta)$.

In order to obtain a standardized version of an effect measure that does not depend on units of measurement, the effect can be standardized,

$$\vartheta = \theta \Big/ \sqrt{V}$$

in the one-dimensional case. (For the mulitvariate case, see Section 3.6.) Note that the single observation variance is used here, which makes the definition a parameter of the model, independent of the number of observations. It still depends on the estimator of the parameter (and the design in regression models, see below) through $V$. One may therefore use the inverse Fisher information for the effect, which equals the variance of the Maximum Likelihood Estimator, instead of the $V$ defined by the estimator actually used.

If the variance depends on the effect parameter, this standardization is of limited value. Therefore, a variance stabilizing transformation may be appropriate. If $V$ is constant, the confidence interval for the standardized effect is

$$\widehat{\vartheta} \pm q/\sqrt{n} \,,$$

where $q$ is the appropriate quantile of the normal or a t distribution.

In the case of two samples, a very popular way to standardize the difference between the groups is Cohen[8]'s $d$

$$d = \theta/\sigma \,.$$

The standardized effect $\vartheta$ is related to $d$ by

$$\vartheta = d/\sqrt{1/\nu_0 + 1/\nu_1} = d\sqrt{\nu_0\nu_1} \,.$$

If the two groups are equally frequent, $\nu_0 = \nu_1 = 1/2$, then $d = 2\vartheta$.

Cohen's $d$ and the effect $\vartheta$ compare the difference between the groups to the variation $\sigma$ of the target variable within groups. This makes sense if $\sigma$ measures the natural standard variation between observation units. It is not well justified if it includes measurement error, since this would change if more precise measurements were obtained, for example, by averaging over several repeated measurements. In this case, the standardized effect is not defined by the scientific question alone, but also by the study design.

Even though $d$ and $\vartheta$ have been introduced in the two samples framework, they also apply to a single sample, since the effect in this case is the difference between its expected value and a potential population that has an expectation of zero. Remember that the effect and its threshold are defined as a function of parameters (a single one in this case), not of their estimates.

## 3.2 Proportions

When a proportion is estimated, the model is, using $\mathcal{B}$ to denote the binomial distribution,

$$\begin{array}{rcll} Y_i & \sim & \mathcal{B}(1,p) \,, & \widehat{p} = S/n \,, \qquad S = \sum_i Y_i \sim \mathcal{B}(n,p) \\ \widehat{p} & \approx\sim & \mathcal{N}(p, V_p/n) \,, & V_p = p\,(1-p) \,. \end{array}$$

For this model, the variance $V_p$ depends on the parameter of interest. As a consequence, the confidence intervals derived from the asymptotic approximation are not suitable for small to moderate sample sizes—more precisely, for small $np$ or $n(1-p)$. Exact confidence intervals are well-known and resolve the problem. However, choosing a relevance threshold needs more attention. It may be plausible to say that a difference of $0.05$ is relevant if $p$ is around $1/2$, but such a difference is clearly too high if $p$ is itself around $0.05$ or below. Thus, the relevance threshold should depend on the effect itself. The choice of a relevance threshold is discussed in Section 4.

**Variance stabilizing transformation.** A variance stabilizing transformation helps to make the general procedures more successful. Here,

$$\vartheta = g(p) = \operatorname{asin}(\sqrt{p})/\,(\pi/2)$$

is the useful transformation. (The division by $\pi/2$ entails a range from $0$ to $1$.) It leads to

$$\widehat{\vartheta} = g(S/n) \approx\sim \mathcal{N}(\vartheta, V/n) \,, \qquad V = 1/\pi^2 \,.$$

**Risk.** Risks usually have low probabilities of occurring. Good practice focusses on logarithmically transformed risks, even more clearly when comparing or modelling them: When a treatment changes a risk, the effect is naturally assessed in terms of a percentage change it entails. This translates into a change on the log scale that is independent of the probability $p$. Thus, the effect measure should be $\vartheta = \log(p)$. The variance transforms to $V \approx 1/p = e^{-\vartheta}$ and again depends on the effect $\vartheta$.

**Logit transformation.** When larger probabilities are studied, it is appropriate to modify the logarithm into the logit transformation, leading to the log-odds instead of the probability $p$ as the effect parameter,

$$\vartheta = \log\left(\frac{p}{1-p}\right) \ , \qquad \widehat{\vartheta} = \log\left(\frac{S+0.5}{n-S+0.5}\right) \ ,$$

where the expression for $\widehat{\vartheta}$ is called empirical logit and avoids infinite values for $S = 0$ and $S = n$. The variance is $\mathrm{var}\left(\widehat{\vartheta}\right) \approx V/n$, where the single observation variance $V$ is

$$V = \frac{1}{p(1-p)} = 2 + e^{\vartheta} + e^{-\vartheta} \ .$$

**Comparing two proportions.** Log-odds are again suitable for a comparison between two proportions $p_0$ and $p_1$. They lead to the log-odds ratio,

$$\vartheta = \log\left(\frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0}\right) = \log(p_1/(1-p_1)) - \log(p_0/(1-p_0)) \ .$$

For such comparisons, paired observations are not popular. Therefore, consider two groups, $k = 0, 1$, with $n_0 = n\nu_0$ and $n_1 = n\nu_1$ observations. Using the difference of empirical logits to estimate $\vartheta$ leads to

$$V = \frac{1}{\nu_0 \, p_0(1-p_0)} + \frac{1}{\nu_1 \, p_1(1-p_1)} \ .$$

Again, the variance stabilizing transformation for $p$ could be used, treating $\vartheta = g(p_1) - g(p_2)$ as the effect, but retaining the desirable properties of the log-odds ratio appears more important.

## 3.3 Simple regression and correlation

**Normal response.** In applications of the common simple regression model,

$$Y_i = \alpha + \beta x_i + \varepsilon_i \ , \qquad \varepsilon_i \sim \mathcal{N}\big(0, \sigma^2\big) \ ,$$

the slope is almost always the parameter of interest, $\theta = \beta$, the nuisance parameters being $\underline{\phi} = [\alpha, \sigma]$. The least squares estimator and its "single observation variance" are

$$\begin{aligned} \widehat{\theta} &= \tfrac{1}{n-1} \textstyle\sum_i (Y_i - \overline{Y})(x_i - \bar{x}) / \,\mathsf{MSX} \ , \qquad \mathsf{MSX} = \tfrac{1}{n-1} \textstyle\sum_i (x_i - \bar{x})^2 \\ V_\theta &= \sigma^2 / \mathsf{MSX} \ . \end{aligned}$$

(To be precise, $V_\theta$ corresponds to (6) if $n$ is replaced by $n-1$.)

In order to make the coefficient comparable between studies, the standardized coefficient $\beta^*$ has been introduced as the amount of change in the target variable, in units of its (marginal) standard deviation $\sqrt{\mathsf{MSY}}$, induced by increasing the predictor $x$ by once its standard deviation, $\delta x = \sqrt{\mathsf{MSX}}$, that is, $\widehat{\beta}^* = \widehat{\beta}\,\sqrt{\mathsf{MSX}}\big/\sqrt{\mathsf{MSY}}$. Here, I prefer to measure the effect in units of the error standard deviation $\sigma$, since this effect is not limited by 1, and therefore the relevance measure will not be limited either. Thus, I introduce the "coefficient effect" as

$$\vartheta = \beta\,\sqrt{\mathsf{MSX}}/\sigma \ , \qquad V = (n-1)\,\mathrm{var}\Big(\widehat{\vartheta}\Big) = 1 \ .$$

(Thus, $\widehat{\vartheta} = \widehat{\beta}^* \sqrt{\mathsf{MSY}}/\widehat{\sigma}$.)

In principle, the effect in this situation should measure the effect of a relevant change $\delta x$ in the predictor $x$ on the target variable $Y$. In the absence of a plausible $\delta x$ and a natural unit of measurement for $Y$ coming from the scientific context, a reasonable choice is to set $\delta x$ equal to the standard deviation of $x$, and $\sigma$ is used as a unit of measurement, leading to $\vartheta$ as the effect scale. It should, however, be noted that the standardized coefficient depends on the standard deviation of the predictor and thus on the design of the experiment in a fixed design situation. In this sense, it does not conform to the principle of focussing on an effect parameter of the model that is independent of choices for obtaining data to estimate it.

Clearly, the two samples problem discussed above is a special case of simple regression, and the effect $\vartheta$ introduced for that problem agrees with the effect defined here.

**Correlation.** Before displaying the formulas for a correlation, let us discuss its suitability as an effect. The related question is: "Is there a (monotonic, or even linear) relationship between the variables $Y^{(1)}$ and $Y^{(2)}$?" According to the basic theme, we need to insert the word "relevant" into this question. But this does not necessarily make the question relevant. What would be the practical use of knowing that there is a relationship? It may be that

- there is a causal relationship; then, the problem is one of simple regression, as just discussed, since the relationship is then asymmetic, from a cause $x$ the a response $Y$;
- one of the variables should be used to infer ("predict") the values of the other; again a regression problem;
- in an exploratory phase, the causes of a relationship may be indirect, both variables being related to common causes, and this should lead to further investigations; this is then a justified use of the correlation as a parameter, which warrants its treatment here.

The Pearson correlation is

$$\begin{aligned} \rho &= \frac{\mathcal{E}\big((Y^{(1)} - \mu^{(1)})(Y^{(2)} - \mu^{(2)})\big)}{\sqrt{\mathcal{E}\big((Y^{(1)} - \mu^{(1)})^2\big)\,\mathcal{E}\big((Y^{(2)} - \mu^{(2)})^2\big)}} \ , \qquad \mu^{(k)} = \mathcal{E}\Big(Y^{(k)}\Big) \\ \widehat{\rho} &= S_{12} \Big/ \sqrt{S_{11} S_{22}} \ , \qquad S_{jk} = \textstyle\sum_i (Y_i^{(j)} - \overline{Y}^{(j)})(Y_i^{(k)} - \overline{Y}^{(k)}) \ . \end{aligned}$$

Fisher's well-known variance stabilizing transformation provides the natural way to treat the case of a simple linear correlation,

$$\vartheta = g(\rho) = \tfrac{1}{2}\log((1+\rho)/(1-\rho)) \ , \quad \widehat{\vartheta} = g(\widehat{\rho}) \ , \quad n\,\mathrm{var}\Big(\widehat{\vartheta}\Big) \approx 1/(1-3/n) \approx V = 1 \ . \quad (7)$$

It is worth noting that it defines a logistic scale, going to infinity when the parameter $\rho$ approaches its extreme values $1$ or $-1$. When large correlations are compared, the effect as measured by the difference of $\vartheta$ values is approximately
$\vartheta = \vartheta_1 - \vartheta_0 \approx \frac{1}{2}\log((1-\rho_0)/(1-\rho_1))$, that is, it compares the complements to the correlation on a relative (logarithmic) scale.

## 3.4 Multiple regression and analysis of variance

This and the following subsections are technically more involved. Readers are encouraged to continue with Section 4 in a first run.

In the multiple regression model, the predictor is multivariate,

$$Y_i = \alpha + \underline{x}_i^{\top}\underline{\beta} + \varepsilon_i \ , \qquad \varepsilon_i \sim \mathcal{N}\big(0, \sigma^2\big) \ . \quad (8)$$

The model also applies to (fixed effects) analysis of variance or general linear models, where a categorical predictor variable (often called a factor) leads to a group of components in the predictor vector $\underline{x}_i$.

Since we set out to ask scientifically relevant questions, a distinction must be made between two fundamentally different situations in which the model is proposed.

- In technical applications, the $\underline{x}$ values are chosen by the experimenter and are therefore fixed numbers. Then, a typical question is whether changing the values from an $\underline{x}_0$ to $\underline{x}_1$ evokes a relevant change in the target variable $Y$. This translates into the relevance of single coefficients $\beta_j$ or of several of them.

- In the sciences, the values of the predictor variables are often also random, and there is a joint distribution of $\underline{X}$ and $Y$. A very common type of question asks whether a predictor variable or a group of them have a relevant influence on the target variable. The naive interpretation of influence here is that, as in the foregoing situation, an increase of the variable $X^{(j)}$ by one unit leads to a change given by $\beta_j$ in the target variable $Y$. However, this is not necessarily true since even if such an intervention may be possible, it can cause changes in the other predictors that lead to a compensation or an enhancement of the effect described by $\beta_j$. Thus, the question if $\beta_j$ is relevantly different from 0 is of unclear scientific merit.

  A legitimate use of the model is prediction of $Y$ on the basis of the predictors. Then, one may ask if a preditor or a group of them reduce the prediction error by a relevant amount.

  It is of course also legitimate to use the model as a description of a dataset. Then, statistical inference is not needed, and there is a high risk of over-interpretation of the outputs obtained from the fitting functions.

- An intermediate situation can occur if the researcher can select observation units that differ mainly in the values of a given subset of predictor variables. Then, any remaining predictors should be excluded from the model, and the situation can be interpreted, with caution, as in the experimental situation.

**Fixed design.** Let us first consider the experimental situation, where the effect of interest is a part of $\underline{\beta}$. If it reduces to a single coefficient $\beta_j$, the other components are part of $\underline{\phi}$, and the formulas for simple regression generalize in a straightforward way,

$$\widehat{\beta}_j = \left(\mathbf{C}\mathbf{X}^\top \underline{Y}\right)_j \ , \quad \mathbf{C} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \ , \qquad V_j = n\,\sigma^2 \mathbf{C}_{jj} \ ,$$

where $\mathbf{X}$ is the design matrix including a column of ones for the intercept term. The standardized coefficient, measuring the effect of increasing $x^{(j)}$ by one standard deviation $s_j$ of $x^{(j)}$ is now $\beta_j^* = \beta_j s_j / \sqrt{\mathsf{MSY}}$, where $s_j$ is the standard deviation of the predictor $X^{(j)}$. Again, I prefer the standardization by the standard deviation of the random deviations $\varepsilon$,

$$\vartheta_j = \beta_j s_j / \sigma \ . \tag{9}$$

If a categorical predictor is in the focus, a contrast between its levels may be identified as the effect of interest. For example, a certain group may be supposed to have higher values for the target variable than the average of the other groups. Then, the problem can be cast in the same way as the single coefficient.

Often, several parameters are of interest. When they have an independent meaning, like the coefficients of several predictors that can be varied independently in an experiment, they are best treated as single coefficients in turn, applying modifications required by multiple testing. However, in case of a categorical predictor and also as a deliberate choice, it may be more adequate to consider the coefficients together as a multivariate effect, and I come back to this view below (Section 3.6). Alternatively, the following approach can be followed.

**Random design.** The prediction error for predicting $Y_0$ for a given predictor vector $\underline{x}_0$ is a function of $\underline{x}_0$, the design $\mathbf{X}$ used for estimation of $\underline{\beta}$, and the variance $\sigma^2$ of the random deviations. In order to simplify the situation, the predictor vector is set to all of those used in the estimation and the squared prediction errors are averaged. This average still depends on the design, which we assume to be random here, and on the number of observations used for estimation. A further simplification just considers the remaining prediction error neglecting estimation of $\underline{\beta}$, which reduces to $\sigma^2$.

In the sequel, I will use the multiple correlation $R$, related to the variances of the random deviatons and of $Y$ by

$$R^2 = 1 - \sigma^2/\mathrm{var}(Y) \ , \qquad \sigma^2 = (1 - R^2)\,\mathrm{var}(Y) \ .$$

The problem considered here asks for comparing a given "full" model, with random deviation variance $\sigma_f^2$, to a "reduced" model in which some components of $\underline{x}$ are dropped—or the respective coefficients set to zero, leading to a variance $\sigma_r^2$. A comparison of variances—or other scale parameters for that matter—is best done at the logarithmic scale, since relative differences are a natural way of expressing such differences (cf. Section 4). Then, an effect measure is

$$\vartheta_{\mathsf{pred}} = \log(\sigma_r/\sigma_f) = \tfrac{1}{2}\log(\theta) \ , \qquad \theta = \frac{\sigma_r^2}{\sigma_f^2} = \frac{1 - R_r^2}{1 - R_f^2} \ . \tag{10}$$

For simple analysis of variance, equivalent to comparison of several groups, $\theta$ reduces to $\theta = 1/(1 - R_f^2)$, where $R_f^2$ is the fraction of the target variable's variance explained by the grouping, called $\eta^2$ in [9] and is between 0 and 1.

Note that $\vartheta = \widetilde{g}(R_r) - \widetilde{g}(R_f)$, where

$$\widetilde{g}(R) = -\tfrac{1}{2}\,\log\bigl(1 - R^2\bigr) \ .$$

It is related to Fisher's z transformation $g$ for correlations (7) by $\widetilde{g}(R) = g(R) - \log(1 + R)$ and shows the same behavior for large $R$.

The effect is estimated by plugging in $\widehat{\sigma}_f$ and $\widehat{\sigma}_r$. The distribution can be characterized by noting that

$$\widehat{\theta} = \frac{(\mathsf{SSE} + \mathsf{SSRed})/\nu_r}{\mathsf{SSE}/\nu_f} = \frac{\nu_f}{\nu_r}\left(1 + F\frac{\nu}{\nu_f}\right) = (\nu_f + \nu F)/\nu_r \approx 1 + \nu F/n \;,$$

where SSE and SSRed are the sums of squares of the error term and for the reduction of the model, $\nu_f$ and $\nu_r$ are the residual degrees of freedom for the full and reduced model, respectively, $\nu = \nu_r - \nu_f$, and $F$ is the usual statistic with an F distribution with $\nu$ and $\nu_f$ degrees of freedom. It is worthwile to note that

$$\nu F = \mathsf{SSRed}/\widehat{\sigma}^2 = \underline{\widehat{\beta}}_a^\top \widehat{\mathrm{var}}\Big(\underline{\widehat{\beta}}_a\Big)^{-1} \underline{\widehat{\beta}}_a = (n-1)\,\widehat{\vartheta}_a^{*2} \;, \tag{11}$$

where $\underline{\beta}_a$ collects the $\nu$ coefficients of the additional predictor variables in the full model and $\widehat{\vartheta}_a^*$ is the estimate of the respective standardized effect norm to be introduced below (15) (the proof is given in the Appendix). Let $\vartheta_a^*$ be defined by

$$\vartheta_a^{*2} = \underline{\beta}_a^\top \mathrm{var}\Big(\underline{\widehat{\beta}}_a\Big)^{-1} \underline{\beta}_a \,/\, n \;, \tag{12}$$

the corresponding squared norm of the true $\underline{\beta}_a$. I call it the "drop effect" of the term(s) defining $\underline{\beta}_a$. It is related to the prediction error effect by

$$\vartheta_{\mathsf{pred}} = \tfrac{1}{2}\log\big(1 + \vartheta_a^{*2}\big) \approx \tfrac{1}{2}\vartheta_a^{*2} \;, \tag{13}$$

the approximation being useful for reasonably small $\vartheta_a^*$.

The effect measure $\vartheta_a^*$ and the corresponding $\vartheta_{\mathsf{pred}}$ can be calculated for the comparison between the full model and the reductions obtained by dropping each term in turn. For continuous predictors, this leads to alternative measures of effect, $\vartheta_j^*$ and $\vartheta_{\mathsf{pred},j}$, to the one defined by the standardized coefficient introduced for fixed designs. In this case, the square root $\vartheta_j^*$ of $\vartheta_j^{*2}$ in (12) shall carry the sign of the coefficient. It is then related to $\vartheta_j$ by

$$\vartheta_j^* = \vartheta_j \sqrt{1 - R_j^2} \;, \tag{14}$$

where $R_j$ is the multiple correlation between predictor $X^{(j)}$ and the other predictors (see Appendix), and it can be interpreted as the effect on the response (in $\sigma$ units) of increasing the predictor $X^{(j)}$, orthogonalized on the other predictors, by one of its standard deviations. If the predictor $X^{(j)}$ is orthogonal to the others, $\vartheta_j$ and $\vartheta_j^*$ coincide.

The distribution of $\widehat{\vartheta}_a^{*2}$ is an F distribution according to (11), with non-centrality $\lambda = n\vartheta_a^{*2}$. A confidence interval cannot be obtained from asymptotic results since the F distribution with low numerator degrees of freedom and low non-centrality is skewed and its variance depends on the exptected value. Therefore, a confidence interval for its non-centrality must be obtained by finding numerical solutions for $\lambda$ in $q^{F(\nu,\nu_f,\lambda)}(\alpha) = F$, for $\alpha = 0.975$ and $= 0.025$. The respective values are then transformed to confidence limits of $\vartheta_{\mathsf{pred}}$ by (13).

## 3.5 Other regression models

**Logistic regression.** For a binary response variable $Y$, logistic regression provides the most well established and successful model. It reads

$$g(P(Y=1)) = \alpha + \underline{x}_i^\top \underline{\beta} + \varepsilon_i \ , \qquad g(p) = \log(p/(1-p)) \ .$$

The parameters of interest are again the coefficients $\beta_j$. The model emerges if the (latent) variable $Z$ follows the ordinary regression model (8) with an random deviation $\varepsilon$ following a standard logistic distribution instead of the normal one, and the observed response $Y$ is a binary classification of it, $Y = 1$ if $Z > c$ for some $c$. Since the definition of an effect should be as independent as possible of the way the model is assessed through observations, the standardized coefficients should be the same in the model for $Z$ and for $Y$. Thus, $\vartheta_j = \beta_j s_j/\sigma$ with a suitable $\sigma$. Since the logistic distribution with scale parameter $\sigma = 5/3$ has $P(|Z| < 1) = 0.67$ like the standard normal distribution, this value is suggested, and

$$\vartheta_j = 0.6\, \beta_j s_j \ .$$

In case of overdispersion, this needs to be divided by the square root of respective parameter $\phi$.

The argument also applies to proportional odds logistic regression for ordered response variables.

In other generalized linear models, like Poisson regression for responses quantifying frequencies, I do not find a plausible version of $\sigma$ and suggest to use $\vartheta_j = \beta_j s_j$.

**Classification.** A classical subject of multivariate statistics is discriminant analysis as introduced by R.A. Fisher using as as example the dataset on iris flowers that has become the most well-known dataset in history. The data follows the model (8) with multivariate $\underline{Y}_i$ and $\underline{\varepsilon}_i$ and predictors $\underline{x}_i$ corresponding to the categorical variable "Species." The interest is not in the multivariate differences between the expected values of the target variables for the three species but in the ability to determine the correct group from the variables' values. If there were only two groups, the problem is better cast by regarding the binary variable "group" as random and the characteristics of the observations—orchids in the example—as predictors and applying the model of logistic regression. For more than two groups, this generalizes to a multinomial regression and leads to a problem of multiple comparisons. This complication goes beyond the scope of the present paper.

## 3.6 Multivariate effects

The general model (6) includes the case of a multivariate parameter of interest $\underline{\theta}$. The test for the null hypothesis $\underline{\theta} = \underline{0}$ is the well-known Chisquared test. The question then arises what a relevant effect should be in this context. A suitable answer is that an effect is relevant if a suitable norm of it exceeds a certain threshold.

A variance standardized effect is determined by a square root of $\mathbf{V}^{-1}$ as

$$\underline{\vartheta} = \mathbf{B}\,\underline{\theta} \ , \qquad \mathbf{B}^\top \mathbf{B} = \mathbf{V}^{-1} \ ,$$

such that $\mathrm{var}(\underline{\vartheta}) = \mathbf{I}$. The context may suggest a suitable root, often the Cholesky factor or the symmetric one.

The standardized effect's (Euclidean) norm $\vartheta^* = \|\underline{\vartheta}\|$ equals the Mahalanobis norm $\Delta$ of $\underline{\theta}$ given by the covariance matrix $\mathbf{V}$. The range of irrelevant effects is then given by

$$\vartheta^{*2} = \Delta^2(\underline{\theta}, \mathbf{V}) = \underline{\theta}^\top \mathbf{V}^{-1} \underline{\theta} < \zeta^2 \ , \tag{15}$$

and the confidence region, by

$$\left\{ \underline{\theta} \mid n\ \Delta^2\Big(\widehat{\underline{\theta}} - \underline{\theta}, \mathbf{V}\Big) \le q \right\} = \left\{ \underline{\vartheta} \mid n\ \|\widehat{\underline{\vartheta}} - \underline{\vartheta}\|^2 \le q \right\} \ ,$$

where $q$ is the $1-\alpha = 0.95$ quantile of the Chisquared or the appropriate F distribution. The two do not intersect if $\Delta(\underline{\theta}, \mathbf{V}) > \zeta + \sqrt{q/n}$ in which case the effect is clearly relevant, case Rlv (Section 2.3). The confidence region is contained in the ellipsoid of irrelevant effects if $\Delta(\underline{\theta}, \mathbf{V}) \leq \zeta - \sqrt{q/n}$, called case Ngl.

Note that in this treatment of the problem, the alternative hypothesis is no longer one-sided for the parameter of interest itself—although it is, for the Mahalanobis norm—, since there is no natural ordering in the multivariate space. This shows an intrinsic difficulty of the present approach in this case. However, the limitation mirrors the difficulty of asking scientifically relevent questions to begin with: What would be an effect that leads to new scientific insight?

In order to fix ideas, let us consider a multivariate regression model. A scientific question may concern an intrinsically multivariate target variable. For example, $\underline{Y}$ may be a characterization of color or of shape, and the multivariate regression model may describe the effect of a treatment on the expected value of $\underline{Y}$. In the case of a single predictor, e.g., in a two-groups situation, the parameter of interest $\underline{\theta}$ in (6) has a direct interpretation as the difference of colors, shapes or the like, and a range of relevant differences may be determined using a norm that characterizes distinguishable colors or shapes, which will be different from $\mathbf{V}$. In more general situations, it seems difficult to define the effect in a way that leads to a practical interpretation.

If the target variable $\underline{Y}$ measures different aspects of interest, like quality, robustness and price of a product or the abundance of different species in an environment, the scientific problem itself is a composite of problems that should be regarded in their own right and treated as univariate problems in turn.

# 4 Relevance thresholds

The arguments in the Introduction have lead to the molesting requirement of choosing a threshold of relevance, $\zeta$. Ideally, such a choice is based on the specific scientific problem under study. However, researchers will likely hesitate to take such a decision and to argue for it. Conventions facilitate such a burden, and it is foreseeable that rules will be invented and adhered to sooner or later, analogously to the ubiquitous fixation of the testing level $\alpha = 5\,\%$. Therefore, some considerations about simple choices of the relevance threshold in typical situations follow here.

**Relative effect.** General intuition may often lead to an agreeable threshold expressed as a percentage. For example, for a treatment to lower blood pressure, a reduction by 10 % may appear relevant according to common sense. Admittedly, this value is as arbitrary as the $5\,\%$ testing level. Physicians should determine if such a change usually entails a relevant effect on the patients' health, and subsequently, a corresponding standard might be generally accepted for treatments of high blood pressure.

When percentage changes are a natural way to describe an effect, it is appropriate to express it formally on the log scale, like $\vartheta = \mathcal{E}\big(\log\big(Y^{(1)}\big)\big) - \mathcal{E}\big(\log\big(Y^{(0)}\big)\big)$ in the two samples situation. Then, one might set $\zeta = 0.1$ for a $10\,\%$ relevance threshold for the change.

**Log-percent.** To be more precise, let the "log-percent" scale for relative effects be defined as $100 \cdot \vartheta$ and indicate it as, e.g., $8.4\,\%\ell$. For small percentages, the ordinary "percent change" and the "log-percent change" are approximately equal. The new scale has the advantage of being symmetric in the two values generating the change, and therefore, the discussion whether to use the first or the second as a basis is obsolete. A change by $100\,\%\ell$ equals an increase of $100\,\%\,(e-1) = 171\,\%$ ordinary percent, or a decrease by $100\,\%\,(1-1/e) = 63\,\%$ in reverse direction. Using this scale, the suggested threshold is $\zeta = 10\,\%\ell$.

**One and two samples, regression coefficients.** An established "small" value of Cohen's $d$ is $20\,\%$ ([8]). It may serve as the threshold for $d$. Since $d = 2\vartheta$ in the case of equal group sizes, this leads to $\zeta = 10\,\%$ for $\vartheta$, which can be used also for unbalanced groups, a single sample as well as regression coefficients according to the discussion in the foregoing section. It also extends to drop effects for terms with a single degree of freedom. However, this threshold transforms to a tiny effect $\vartheta_{\mathsf{pred}}$ of $0.5\,\%\ell$ on the difference in lengths of prediction intervals according to (13). A threshold of $5\,\%\ell$ seems be more appropriate here. This shows again that the scientific question should guide the choice of the effect scale and of the relevance threshold!

**Correlation.** In the two samples situation, considering the $x_i$ as random,

$$\rho^2 = \nu_0\nu_1 d^2/(1 + \nu_0\nu_1 d^2)\ , \tag{16}$$

and the threshold of $20\,\%$ on Cohen's $d$ leads approximately again to $\zeta = 0.1$ (see Appendix for the calculation). However, if correlations are compared between each other rather than to zero, a transformed correlation is more suitable as an effect measure. If the Fisher transformation is used, then the same threshold can be applied, since $\vartheta = g(\rho) \approx \rho$ for $\rho \leq 0.1$. Since $g$ is a logarithmic transformation, I write $\zeta = 10\,\%\ell$.

**Proportions.** The comparison of two proportions is a special case of logistic regression, with $\beta$ equal to the log odds ratio and MSX $= \nu_0\nu_1$ as for the two samples case. If the threshold for coefficient effects, $10\,\%$, is used and the two groups have the same size, this leads to a threshold of $\zeta = 33\,\%\ell$ for the log odds ratio, which appears quite high in this situation.

On the other hand, for low risks, the recommendation for relative effects applies. For larger probabilities $p$, the transformation turns into the logit, $\vartheta = \log(p/(1-p))$, and "log-percent" turn into "logit-percent." The threshold $\zeta = 10\,\%\ell$ may still be used in this scale. Back-transformation to probabilities $p$ leads to a change from $p = 0.5$ to $p = 0.525$ being relevant, and from $25\,\%$ to $27\,\%$,from $10\,\%$ to $10.9\,\%$, and from $2\,\%$ to $2.2\,\%$.

**Log-linear models.** Several useful models connect the logarithm of the expected response with a linear combination of the predictors, notably Poisson regression with the logarithm as the canonical link function, log-linear models for frequencies, and Weibull regression, a standard model for reliability and survival data. Here, the consideration of a relative effect applies again. An increase of 0.1 in the linear predictor leads to an increase of $10\,\%$ in the expected value, and therefore, $\zeta = 10\,\%\ell$ seems appropriate for the standardized coefficients $\vartheta_j = \beta_j s_j$.

**Summary.** The scales and thresholds for the different models that are recommended here for the case that the scientific context does not suggest any choices are listed in Table 2.

# 5 Description of results

It is common practice to report the statistical significance of results by a p-value in parenthesis, like "The treatment has a significant effect ($p = 0.04$)," and estimated values are often decorated with asterisks to indicate their p-values in symbolized form. If such short descriptions are desired, secured relevance values should be given. If RIs $> 1$, the effect is relevant, if it is $> 0$, it is significant in the traditional sense, and these cases can be distingished in even shorter form in tables by plusses or an asterisk as symbols as follows: $^*$ for significant, that is, RIs $> 0$; $^+$ for relevant (RIs $> 1$); $^{++}$ for RIs $> 2$; and $^{+++}$ for RIs $> 5$. To make these indications well-defined, the relevance threshold $\zeta$ must be declared either for a whole paper or alongside the indications, like "RIs $= 1.34$ ($\zeta = 10\,\%\ell$)."

**Table 2.** Models, recommended effect scales and relevance thresholds

| Problem | Basic model | Effect $\vartheta = g(\theta)$ | Rel. thresh. $\zeta$ |
|---|---|---|---|
| One, or two paired samples | $\mathcal{N}(\mu, \sigma^2)$ | $\mu/\sigma$ | $10\,\%$ |
| Two independent samples | $\mathcal{N}(\mu_k, \sigma^2)$ | $d = (\mu_1 - \mu_0)/\sigma$<br>$\vartheta = (\mu_1 - \mu_0)\sqrt{\nu_0\nu_1}/\sigma$ | $20\,\%$<br>$10\,\%$ |
| Regression<br>coefficients<br>prediction error | $Y_i = \alpha + \underline{x}_i^\top \underline{\beta} + \varepsilon_i$<br>$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ | <br>$\underline{\beta}_j \sqrt{\mathsf{MSX}^{(j)}}/\sigma$<br>$-\frac{1}{2}\log(1 - R^2)$ | <br>$10\,\%$<br>$0.5\,\%\ell$ or $5\,\%\ell$ |
| Logistic regression | $g(Y_i = 1) = \alpha + \underline{x}_i^\top \underline{\beta}$ | $\underline{\beta}_j 0.6\sqrt{\mathsf{MSX}^{(j)}}/\sqrt{\phi}$ | $10\,\%\ell$ |
| Relative Difference | $\log(Y) \sim \mathcal{N}(\mu_k, \sigma^2)$ | $\log(\mu_1/\mu_0)$ | $10\,\%\ell$ |
| Proportion | $\mathcal{B}(n, p)$ | $\log(p/(1-p))$ | $33\,\%\ell$ or $10\,\%\ell$ |
| Correlation | $\underline{Y} \sim \mathcal{N}_2(\underline{\mu}, \mathbf{\Sigma})$<br>$\rho = \mathbf{\Sigma}_{12}/\sqrt{\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22}}$ | $\frac{1}{2}\log((1+\rho)/(1-\rho))$ | $10\,\%\ell$ |

**Examples.** The first examples are taken from the first "manylabs" project about replicability of findings in psychology ([10]), since for that study, the scientific questions had been judged to deserve replication and full data for the replication is easily available.

The original studies were replicated in each of 36 institutions. Here, I pick the replication at Penn State University of the following item: "Students were asked to guesstimate the height of Mount Everest. One group was 'anchored' by telling them that it was more than 2000 feet, the other group was told that it was less than 45,500 feet. The hypothesis was that respondents would be influenced by their 'anchor,' such that the first group would produce smaller numbers than the second" ([11]). The true height is 29,029 feet.

According to the discussion in Section 4, the data is analyzed here on the logarithmic scale, and the threshold of $10\,\%\ell$ is applied. The data, reduced to the first 20 observations for simplicity, are given in Table 3.

**Table 3.** Data for the anchoring example in 1,000 feet

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| group "low", $n_0 = 8$ | 2.3 | 2.7 | 3 | 3 | 3.1 | 6 | 12 | 15 | | | | |
| group "high", $n_1 = 12$ | 25 | 32 | 34 | 40 | 40 | 40 | 42.7 | 43.5 | 44 | 45 | 45 | 45.5 |

The group means of the log values are $1.52$ and $3.67$ (corresponding to $4{,}560$ and $39{,}190$ feet) and the standard error for their difference is $0.216$ on $18$ degrees of freedom. This leads to a confidence interval of $\widehat{\vartheta} \pm \widehat{\omega} = 2.15 \pm (2.10 \cdot 0.216) = [1.70, 2.60]$ and $\mathsf{Sig0} = \widehat{\vartheta}/\widehat{\omega} = 4.74$. The relevance is $\mathsf{Rle} = 100 \cdot \widehat{\vartheta}/\zeta = 2.15/0.1 = 21.5$ with interval limits of $\mathsf{Rls} = \mathsf{Rle} - \widehat{\omega}/\zeta = 17.0$ and $\mathsf{Rlp} = \mathsf{Rle} + \widehat{\omega}/\zeta = 26.0$. The single value notation is $\mathsf{Rls} = 17.0$ ($\zeta = 10\,\%\ell$). This is an extremely clear effect.

A second study asked if a positive or negative formulation of the same options had an effect on the choice ([12]). Confronted with a new contagious disease, the government has a choice between action A that would save 200 out of 600 people or action B which would save all 600 with probability 1/3. The negative description was that either (A) 400 would die or (B) all 600 would die with probability 2/3. I report the results for Penn State (US) and Tilburg (NL) universities. The data is summarized in Table 4, and the effect, significance, and relevance, in Table 5. The secured relevance is $\mathsf{Rls} = 4.16$ ($\zeta = 10\,\%\ell$) and

$10.1$ ($\zeta = 10\,\%\ell$) for the two institutions, the effect is thus clearly relevant. One may ask if there is a relevant (!) difference between these two studies, with a view of applying the notions of this paper to the theme of replicability. This will be done in a forthcoming paper.

**Table 4.** Data for the second example

| | PSU | | | Tilburg | | |
|---|---|---|---|---|---|---|
| | $n$ | A | B | $n$ | A | B |
| negative | 48 | 16 | 32 | 34 | 6 | 28 |
| positive | 47 | 30 | 17 | 46 | 29 | 17 |

**Table 5.** Results for the second example. Relevance threshold $10\,\%\ell$.

| | effect | | | signif. | relevance | | |
|---|---|---|---|---|---|---|---|
| | $\widehat{\vartheta}$ | low | high | Sig0 | Rle | Rls | Rlp |
| PSU | **1.26** | 0.42 | 2.10 | 1.49 | 12.6 | **4.2** | 21.0 |
| Tilburg | **2.08** | 1.01 | 3.14 | 1.95 | 20.8 | **10.1** | 31.4 |

The third example is a multiple regression problem. The dataset reflects the blasting activity needed for digging a freeway tunnel beneath a Swiss city. Since blasting can cause damage in houses located at a small `distance` from the point of blasting, the `charge` should be adjusted to keep the `tremor` in the basement of such a house below a threshold $y_0$. The logarithmic `tremor` is modelled as a linear function of the logarithmic `distance` and `charge`, an additive adjustment to the house where the measurements are taken (factor `location`), and `time`, a rescaled calendar day. Only part of the data for 3 locations are used here, see Table 6.

**Table 6.** Data for the blasting example.

| charge | dist | loc'n | time | tremor | charge | dist | loc'n | time | tremor |
|---|---|---|---|---|---|---|---|---|---|
| 4.760 | 62 | loc1 | 0.5562 | 4.07 | 3.640 | 55 | loc1 | 0.7644 | 4.31 |
| 4.848 | 58 | loc1 | 0.5699 | 0.71 | 3.708 | 61 | loc1 | 0.7699 | 4.43 |
| 5.824 | 55 | loc1 | 0.5890 | 6.71 | 3.812 | 46 | loc2 | 0.7726 | 10.67 |
| 6.656 | 50 | loc1 | 0.6082 | 12.23 | 3.725 | 69 | loc4 | 0.7808 | 2.00 |
| 6.656 | 42 | loc1 | 0.6274 | 10.55 | 3.305 | 67 | loc1 | 0.7836 | 2.51 |
| 4.368 | 37 | loc1 | 0.6384 | 16.90 | 3.744 | 50 | loc2 | 0.7863 | 7.91 |
| 5.200 | 33 | loc1 | 0.6548 | 16.90 | 3.725 | 65 | loc4 | 0.7863 | 3.47 |
| 4.998 | 31 | loc1 | 0.6685 | 14.99 | 3.725 | 55 | loc2 | 0.7918 | 5.63 |
| 4.998 | 49 | loc2 | 0.6712 | 8.39 | 3.870 | 61 | loc4 | 0.8000 | 2.36 |
| 5.236 | 29 | loc1 | 0.6849 | 16.42 | 4.765 | 60 | loc2 | 0.8055 | 6.59 |
| 5.593 | 44 | loc2 | 0.6877 | 12.23 | 1.248 | 59 | loc4 | 0.8055 | 1.70 |
| 1.190 | 30 | loc1 | 0.6904 | 5.03 | 4.644 | 62 | loc2 | 0.8082 | 5.15 |
| 4.998 | 41 | loc2 | 0.6932 | 12.23 | 5.285 | 56 | loc4 | 0.8110 | 5.39 |
| 4.998 | 31 | loc1 | 0.7041 | 14.27 | 5.285 | 69 | loc2 | 0.8219 | 5.27 |
| 5.712 | 38 | loc2 | 0.7068 | 23.38 | 0.624 | 53 | loc4 | 0.8247 | 1.07 |
| 4.680 | 35 | loc1 | 0.7123 | 13.91 | 3.986 | 73 | loc2 | 0.8274 | 5.03 |
| 4.702 | 36 | loc2 | 0.7233 | 14.15 | 2.490 | 51 | loc4 | 0.8301 | 4.43 |
| 4.784 | 39 | loc1 | 0.7260 | 9.95 | 4.390 | 79 | loc2 | 0.8411 | 4.43 |
| 5.824 | 36 | loc2 | 0.7288 | 13.43 | 4.390 | 50 | loc4 | 0.8438 | 5.99 |
| 4.160 | 43 | loc1 | 0.7425 | 10.55 | 3.870 | 85 | loc2 | 0.8466 | 2.63 |
| 3.952 | 36 | loc2 | 0.7425 | 20.98 | 3.870 | 50 | loc4 | 0.8493 | 5.27 |
| 3.744 | 88 | loc4 | 0.7452 | 1.52 | 1.768 | 50 | loc4 | 0.8685 | 1.58 |
| 3.194 | 50 | loc1 | 0.7479 | 7.07 | 2.496 | 51 | loc4 | 0.0000 | 3.29 |
| 3.744 | 38 | loc2 | 0.7507 | 14.51 | 3.640 | 52 | loc4 | 0.0192 | 4.67 |
| 3.305 | 79 | loc4 | 0.7616 | 1.43 | | | | | |

An extensive table of results is shown in Table 7. The `time` does not show any significance and therefore no relevance either. The relevances of the coefficient and drop effects are related by (14). Thus, their ratio equals $\sqrt{1 - R_j^2}$ and is a useful measure of collinearity.

For the shortest description, the coefficient of `log10(charge)` would be indicated as $0.752^{+++}$.

**Table 7.** Extensive results for the blasting expample `location` is a factor with 3 levels. Relevance thresholds are $10\,\%$ for standardized coefficients and $5\,\%\ell$ for the prediction effect. The columns shown in bold face should be routinely shown.

| | | | | | | stand. | coef. effect | | drop effect | | prediction eff. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| term | **coef.** | **df** | se | Sig0 | p.value | coef. | **Rlp** | **Rls** | Rlp | **Rls** | Rlp | Rls |
| location | | **2** | | 1.56 | 1.27e-03** | | | | 8.70 | **2.33**$^{++}$ | 5.58 | 0.12 . |
| log10(distance) | **-2.022** | **1** | 0.198 | -5.06 | 4.68e-13*** | -1.666 | **19.87** | **13.5**$^{+++}$ | 17.82 | **11.94**$^{+++}$ | 14.75 | 9.16$^{+++}$ |
| log10(charge) | **0.752** | **1** | 0.130 | 2.86 | 7.94e-07*** | 0.959 | **12.85** | **6.3**$^{+++}$ | 11.74 | **5.01**$^{+++}$ | 8.96 | 2.20$^{++}$ |
| time | **0.062** | **1** | 0.138 | 0.22 | 0.66 | 0.069 | **3.68** | **-2.3**$^{-}$ | 3.45 | **0.00** | 1.00 | 0.000 |

The results for these examples have been otained by the R package `relevance`, available from `https://r-forge.r-project.org`.

# 6 Relevance instead of p-values

The deficiencies of the common use of p-values has lead to a fierce debate and a flood of papers, often resulting in the vague conclusion that the accused concept should be used with caution. Some alternatives have nevertheless been given, such as the "Second Generation P-Value" by Blume *et al.* [1], which I have discussed above. Bayesian methods have also been advertized as a methodology that gives a more differentiated picture. None of these proposals have yet been widely applied.

Here, I have argued that the origin of the crisis roots deeper: The misuse of the p-value reflects a way to avoid the effort of asking relevant scientific questions to begin with. Typical problems in empirical research often concern a quantity like the effect of a treatment on a specified target variable. These problems are only well-posed if there is a threshold of relevance. I am not the first to advocate this requirement, I emphasize its importance again and develop it further into the novel measure of relevance. It is essential to keep in mind that the threshold should be determined only by the scientific problem and therefore must not depend on the design of the study that estimates the effect.

The paradigm of null hypothesis significance testing that is so well established asks for the choice of a threshold: the significance level $\alpha$ of the test, or the confidence level $1-\alpha$. In principle, $\alpha$ could be arbitrarily chosen, but tradition has fixed it at 5 % for most scientific fields. The relevance threshold introduces yet another choice to be made. A careful selection should be sought in each scientific study. Since this is a cumbersome requirement, conventions have been proposed in this paper for the most common situations.

The traditional method to convey the assessment of an effect in a more informative way than the p-value is the confidence interval. Its downside is that it consists of two numbers that carry the measurement unit of the effect and are therefore not directly comparable between studies. The significance measure introduced here is a single, standardized number that conveys the essentials of the confidence interval. It depends, however, again on a given value of the effect. When this value is 0, the basic flaw of the p-value is inherited. Combining it with the relevance threshold is a necessary step to give an appropriate characterization of the relevance of a result.

The combination is best achieved by focussing on the confidence interval for the relevance measure, with boundaries called "secured" and "potential" relevance. The secured relevance Rls may even be used as a single number characterizing the knowledge gained about the effect of interest.

A conclusion from the p-value debate is that a simple yes-no decision about the result is misleading. Since our thinking likes categorization, I have introduced labels characterizing the comparison of the confidence interval with both the zero effect and the relevance threshold. It is defined on the basis of the two significance values $\text{Sig}_0$ and $\text{Sig}_\zeta$ or of the two relevance limits Rls and Rlp.

The significance and relevance measures and the classification are straightforward enhancements of concepts that are well established and ubiquously known. There is hope that they can form a new standard of presenting statistical results.

**Replicability.** The p-value debate is closely related to and often confounded with the reproducibility crisis. In fact, there is ample evidence that in several fields of science, when a statistical study is replicated, a significant effect found in the original study turns out to be non-significant in the replication, thereby formally failing the fundamental requirement of reproducibility of empirical science. While many causes are suggested and found for such failures, prominent ones are tied to the problems with statistical testing and the p-value discussed in the Introduction. Here is the argument:

The p-value was originally advocated as a filter against publication of results that may be due to pure randomness. It was soon converted in a tool to generate "significant" results regardless of their scientific relevance. This leads to so-called selection bias: When many studies examine small true effects with limited precision, some of them will turn out significant by chance, will thus pass the filter and be published, whereas the non-significant ones will go unnoticed. These studies will have a low probability of being successfully replicated.

Clearly, using the criterion of a *relevant* secured effect (case Rlv) as a filter would reduce the frequency of phony results drastically: A relevant result in this sense will usually have a high probability of showing at least a significant estimate (case Amb.Sig) upon replication—unless the precision is low or data snooping has been extensively applied to get it. The concepts introduced here can be profitably applied to assess replications of results also in more depth, as will be shown in a forthcoming paper.

# 7 Conclusion

The p-value has been (mis-) used to express the results of statistical data analyses for too long, in spite of the extensive discussions about the bad consequences of this practice for science.

It is time to introduce a new concept for the presentation of the statistical inference for an effect under study. The measure of relevance introduced here is suitable to achieve this goal. It needs the choice of a relevance threshold for the effect of interest, a requirement posed by the desire to ask a scientifically meaningful question to begin with.

The goal of a typical statistical enquiry is to prove that an effect is relevant. Based on the measures "secured relevance," Rls, and "potential relevance." Rlp, either this can be achieved, or a "negligible" effect can be found—or the answer may be "ambiguous."

Application of these concepts will enhance reproducibility: When relevant effects are examined rather than merely significant ones, the replication will much more often turn out to be at least significant in the replication.

**Acknowledgement.** My colleagues Martin Mächler and Matteo Tanadini as well as Samuel Pawel of the Biostatistics group at University of Zurich have provided valuable comments and suggestions on the writeup of this paper.

# References

1. Blume JD, McGowan LD, Greevy RA, Dupont WD. Second-Generation p-Values: Improved Rigor, Reproducibility, and Transparency in Statistical Analyses. PLOS ONE. 2018;13:e0188299.

2. Benjamin D, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, many more. Redefine statistical significance. Nature Human Behavior. 2018;2:6–10.

3. Lakens D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science. 2017;8:355–362.

4. Goodman WM, Spruill SE, Komaroff E. A Proposed Hybrid Effect Size Plus p-Value Criterion: Empirical Evidence Supporting its Use. The American Statistician. 2019;73(1, suppl.):168–185.

5. Student. The probable error of a mean. Biometrika. 1908;6:1–25.

6. Blume JD, Greevy RA, Welty VF, Smith JR, Dupont WD. An Introduction to Second-Generation p-Values. The American Statistician. 2019;73(1, suppl.):157–167.

7. Lakens D, Delacre M. Equivalence Testing and the Second Generation P-Value; 2020.

8. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale, N.J./Academic Press; 1988/2013.

9. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology. 2013;4(Article 863):1–12.

10. Klein RA, Ratliff KA, Vianello, M et al . Investigating variation in replicability: A "many labs" replication project. Social Psychology. 2014;45(3):142–152.

11. Jacowitz KE, Kahneman D. Measures of anchoring in estimation tasks. Social Psychology Bulletin. 1995;21:1161–1166.

12. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science. 1981;211:453–458.

# Appendix

**Derivation of** (11), $\nu F = \underline{\widehat{\beta}}_a^\top \widehat{\mathrm{var}}\left(\underline{\widehat{\beta}}_a\right)^{-1} \underline{\widehat{\beta}}_a$.

The general formula for the inversion of a partitioned matrix reads

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix}.$$

Let $\mathbf{X} = [\mathbf{X}_r \;\; \mathbf{X}_a]$ and

$$\begin{aligned} \mathbf{K} &= \left(\mathbf{X}_r^\top \mathbf{X}_r\right)^{-1}, \qquad \mathbf{H} = \mathbf{X}_r \mathbf{K} \mathbf{X}_r^\top \\ \mathbf{M}^{-1} &= \mathbf{X}_a^\top (\mathbf{I}-\mathbf{H}) \mathbf{X}_a, \qquad \mathbf{G} = \mathbf{X}_a \mathbf{M} \mathbf{X}_a^\top. \end{aligned}$$

Inverting $\mathbf{X}^\top\mathbf{X}$ then leads to

$$\begin{aligned} \left(\mathbf{X}^\top\mathbf{X}\right)^{-1} &= \begin{bmatrix} \mathbf{K} + \mathbf{K}\mathbf{X}_r^\top \mathbf{X}_a \mathbf{M} \mathbf{X}_a^\top \mathbf{X}_r \mathbf{K} & -\mathbf{K}\mathbf{X}_r^\top \mathbf{X}_a \mathbf{M} \\ -\mathbf{M}\mathbf{X}_a^\top \mathbf{X}_r \mathbf{K} & \mathbf{M} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{K} + \mathbf{K}\mathbf{X}_r^\top \mathbf{G} \mathbf{X}_r \mathbf{K} & -\mathbf{K}\mathbf{X}_r^\top \mathbf{X}_a \mathbf{M} \\ -\mathbf{M}\mathbf{X}_a^\top \mathbf{X}_r \mathbf{K} & \mathbf{M} \end{bmatrix} \\ \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top &= \begin{bmatrix} \mathbf{K}\mathbf{X}_r^\top + \mathbf{K}\mathbf{X}_r^\top \mathbf{G} \mathbf{X}_r \mathbf{K}\mathbf{X}_r^\top - \mathbf{K}\mathbf{X}_r^\top \mathbf{X}_a \mathbf{M}\mathbf{X}_a^\top \\ -\mathbf{M}\mathbf{X}_a^\top \mathbf{X}_r \mathbf{K}\mathbf{X}_r^\top + \mathbf{M}\mathbf{X}_a^\top \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{K}\mathbf{X}_r^\top (\mathbf{I} + \mathbf{G}\mathbf{H} - \mathbf{G}) \\ \mathbf{M}\mathbf{X}_a^\top (\mathbf{I} - \mathbf{H}) \end{bmatrix} \\ \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top &= \mathbf{H}(\mathbf{I} - \mathbf{G} - \mathbf{G}\mathbf{H}) + \mathbf{G}(\mathbf{I} - \mathbf{H}) = \mathbf{H} + (\mathbf{I} - \mathbf{H})\mathbf{G}(\mathbf{I} - \mathbf{H}) \end{aligned}$$

Since $\underline{\widehat{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\underline{Y}$,

$$\begin{aligned}
\underline{\widehat{\beta}}_a &= \mathbf{M}\mathbf{X}_a^\top(\mathbf{I}-\mathbf{H})\underline{Y} \\
\underline{\widehat{\beta}}_a^\top\mathbf{M}^{-1}\underline{\widehat{\beta}}_a &= \underline{Y}^\top(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})\underline{Y} \\
\mathsf{SSRed} &= -(\mathsf{SSE}-\mathsf{SSR}) = \underline{Y}^\top\underline{Y} - \underline{Y}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\underline{Y} - \left(\underline{Y}^\top\underline{Y} - \underline{Y}^\top\mathbf{H}\underline{Y}\right) \\
&= \underline{Y}^\top\left(\mathbf{H} + (\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H}) - \mathbf{H}\right)\underline{Y} \;=\; \underline{\widehat{\beta}}_a^\top\mathbf{M}^{-1}\underline{\widehat{\beta}}_a \\
\mathsf{SSRed}/\widehat{\sigma}^2 &= \underline{\widehat{\beta}}_a^\top\widehat{\mathrm{var}}\Big(\underline{\widehat{\beta}}_a\Big)^{-1}\underline{\widehat{\beta}}_a\ .
\end{aligned}$$

If $\underline{\beta}_a$ is one-dimensional, $\mathbf{X}_a = \underline{X}^{(j)}$, then

$$\mathbf{M}^{-1} = \underline{X}^{(j)\top}(\mathbf{I}-\mathbf{H})\underline{X}^{(j)} = ||\underline{X}^{(j)} - \overline{X}^{(j)}||^2(1-R_j^2)$$

where $R_j$ is the multiple correlation between $\underline{X}^{(j)}$ and the other predictors, $\mathbf{X}_r$, and therefore,

$$\widehat{\vartheta}_j^{*2} = (n-1)\,\widehat{\beta}_j^2/\widehat{\mathrm{var}}\Big(\widehat{\beta}_j\Big) = (n-1)\,\widehat{\beta}_j^2\widehat{\mathrm{var}}\Big(X^{(j)}\Big)\,(1-R_j^2)/\widehat{\sigma}^2 = \widehat{\vartheta}_j^2(1-R_j^2)\ .$$

**Derivation of (16),** $\rho^2 = \nu_0\nu_1 d^2/(1+\nu_0\nu_1 d^2)$, assuming that $X = Y^{(1)}$ is binary, $P(X=k) = \nu_k$ for $k = 0, 1$, and $Y|X{=}k \sim \mathcal{N}\big(\mu_k, \sigma^2\big)$, $d = (\mu_1-\mu_0)/\sigma$. Let $\Delta = \mu_1 - \mu_0$, $Y = Y^{(2)}$.

$$\begin{aligned}
\mathcal{E}(X) &= \nu_1 \qquad \mathrm{var}(X) = \nu_0\nu_1 \\
\mathcal{E}(Y) &= \mu_0 + \nu_1\Delta \\
\mathcal{E}\big(Y^2\big) &= \sigma^2 + \nu_0\mu_0^2 + \nu_1(\mu_0+\Delta)^2 \\
\mathrm{var}(Y) &= \sigma^2 + \nu_0\mu_0^2 + \nu_1\mu_0^2 + 2\nu_1\mu_0\Delta + \nu_1\Delta^2 - \big(\mu_0^2 + \nu_1^2\Delta^2 + 2\nu_1\mu_0\Delta\big) \\
&= \sigma^2 + \nu_0\nu_1\Delta^2 \\
\mathcal{E}(XY) &= \nu_1\mu_1 \\
&\mathcal{E}((X-\mathcal{E}(X))(Y-\mathcal{E}(Y))) \\
&= \nu_1\mu_1 - \nu_1(\mu_0+\nu_1\Delta) = \nu_1(\mu_1-\mu_0) - \nu_1^2\Delta = \nu_0\nu_1\Delta \\
\rho^2 &= (\nu_0\nu_1\Delta)^2\,\big/\nu_0\nu_1\left(\sigma^2 + \nu_0\nu_1\Delta^2\right) = \nu_0\nu_1 d^2\,\big/\big(1+\nu_0\nu_1 d^2\big)
\end{aligned}$$