The Ancova model for comparing two groups: a tutorial emphasizing statistical distribution theory

Wolf Schwarz*

Department of Psychology, University of Potsdam, Potsdam, Germany

The analysis of covariance (Ancova) is a widely used statistical technique for the comparison of groups with respect to a quantitative dependent variable in such a way that the comparison takes into account concomitant differences in a quantitative covariate. Despite its widespread use, some of the main features of this technique have remained elusive, contentious, or misconceived in applied settings. For example, some authors claim that the validity of an Ancova depends on the assumption that the expected value of the covariate is the same for all participants, or that the adjusted mean difference evaluated in an Ancova has a useful interpretation as the difference between the mean change scores of each group, whereas these claims are disputed by other authors. I suggest that these issues are best addressed and settled in the context of the underlying exact sampling distribution theory since significance statements, effect size estimates, and statistical power all derive directly from the statistical sampling distribution theory implied by the Ancova model. The distributional approach also clarifies the central distinction between conditional and marginal means, and the way in which various study designs (controlled, randomized, observational) affect and modify conclusions derived from an Ancova. The tutorial provides an explicit distributional account of the standard Ancova model to compare two independent groups; it clarifies the assumptions underlying the Ancova model, the nature and limitations of the conclusions it provides, and corrects some common misconceptions associated with its applications.

KEYWORDS

analysis of covariance, conditional vs. marginal means, gain scores, noncentral *F*-distribution, randomized design, observational study

1 Introduction

Empirical studies often aim to compare two or more treatments applied to separate groups. A standard technique for this purpose is the analysis of variance (Anova), which in the case of two independent groups is equivalent to an independent samples $t-$test. In these techniques the mean variability within each group defines and delimits the resolution with which differences between the groups can be identified. It is therefore natural to consider ways to refine this resolution so that the comparison becomes more sensitive, and fewer units are required to detect a given difference. Effective techniques to achieve this aim are, for example, blocking, or the use of repeated measures from the same unit (for general background, see, e.g., Maxwell et al., 2018, ch. 9; Kutner et al., 2004, ch.s 26–29).

In the case of two groups, blocking typically uses a *covariate* to form pairs (blocks) of units that are as similar as possible with respect to the covariate. Treatments are then randomly assigned to one or the other member of any given pair; this procedure serves to make the two groups more similar on average. Blocking leads to more sensitive comparisons of the treatments, as potentially relevant differences in the covariate between the units are balanced more systematically than under unconstrained randomization. However, the technique of blocking requires that the covariate is available before the units are assigned to treatment groups, and with quantitative covariates it ignores the residual variability of the covariate within blocks (for a discussion of further limitations of blocking, see Maxwell et al., 2018, ch. 9; Schwarz and Reike, 2018; Schwarz, 2008). A widely used technique that makes use of all the quantitative information in the covariate is the analysis of covariance (Ancova; Huitema, 2011; Kutner et al., 2004, ch. 22; Schneider et al., 2015; Senn, 2006; Vickers and Altman, 2001). I illustrate typical Ancova applications by two examples, Study A and B.

1.1 Study A

To compare how much time it takes consumers to empty a straight (cylindrical) vs. curved (truncated cone) glass of beer (volume in both cases 12 fl oz), naive subjects were randomly assigned to two treatment groups defined by glass shape (for a detailed study, see Attwood et al., 2012). As a potentially relevant covariate, a questionnaire score was obtained at the end of the study from all participants which assessed the urge to consume alcoholic beverages so as to control the comparison related to glass shape for potential base differences of drinking habits. Here, the treatment variable defining the groups is the shape of the glass, the dependent variable is the drinking time, and the covariate is the questionnaire score.

1.2 Study B

The high jump performance of randomly selected male and female high-school graduates were compared. As a covariate, the body height of each graduate was recorded, so as to control the comparison related to gender for potential differences on this variable. Here, the variable defining the groups is gender, the dependent variable is the performance (maximum height jumped over) of the graduate, and the covariate is his/her body height.

Note that in neither Study A nor B did the researcher exert any *direct* control over the covariate. Rather, the covariate values arise with the specific participants who are randomly sampled, and these values would clearly change in any replication of either study. Study A uses a randomized design; consequently, in many hypothetical replications of that study the long-run average of the covariate would not be expected to differ between the two groups, although of course in each individual sample, the means would typically be different. In contrast, even though the male and female graduates were randomly selected in Study B, the assignment of a particular participant to the two gender groups is

not under the control of the researcher. Therefore, any covariate that differs between these two groups could be at least partially responsible for the differences in the dependent variable. When using body height as a covariate, the investigator asks whether there would be a systematic difference in the high jump performance of male vs. female graduates if these graduates were the same height. The question is clearly counterfactual, because it is well-established that male and female graduates differ systematically in height. Addressing this counterfactual aspect clearly requires some form of model-related assumptions in order to arrive at valid interpretations of an Ancova in observational designs (e.g., Huitema, 2011, ch. 8; Lord, 1967).

Expositions of the Ancova typically represent its central assumptions indirectly, for example in graphical form with Venn or flow diagrams, by focusing on detailed numerical calculations within a single specific sample, or by analogy to regression modeling. It is plausible that this indirect way of presenting the Ancova model has contributed to various ambiguities and controversies regarding, for example, the assertion that for an Ancova to be valid all participants must have the same expected value of the covariate (e.g., Schneider et al., 2015), or that (if the dependent variable and the covariate are commensurate) the adjusted mean difference evaluated in an Ancova represents the difference between the mean change scores of each group (e.g., Vickers and Altman, 2001). Surprisingly few accounts in the applied Ancova literature are based on the actual sampling distributions associated with the Ancova (for some exceptions at a technically advanced level, see Schneider et al., 2015; Shieh, 2017, 2021). This is unfortunate, because the central topics associated with an Ancova are best understood directly in terms of the underlying exact sampling distribution theory, which greatly helps to clarify the assumptions underlying the Ancova model, to understand the nature and limitations of its conclusions, and to correct various misconceptions related to its applications. More specifically, significance statements, effect size estimates, and statistical power are all ultimately derived from the statistical distribution theory underlying the Ancova model. In addition, the statistical sampling theory also helps to understand the influence on statistical power of the correlation between the dependent variable and the covariate, and the profound effect of various sampling designs (controlled, randomized, observational) on the conclusions that can be derived from an Ancova.

The aim of the present tutorial is therefore to provide an accessible and succinct account of the sampling theory that underlies the standard statistical Ancova model for comparing two independent groups. Specifically, I explain the application of these sampling distribution results to power computations, to the relation to gain score analyses, and to the effect of the study design. The general discussion will also explain how the contentious issues referred to above are clarified within the more general framework of exact sampling distribution theory.

2 The standard Ancova model for comparing two groups

Denote as y_{ij} the sample values of the dependent variable, and as x_{ij} the corresponding values of the covariate; the index $i = 1, 2$
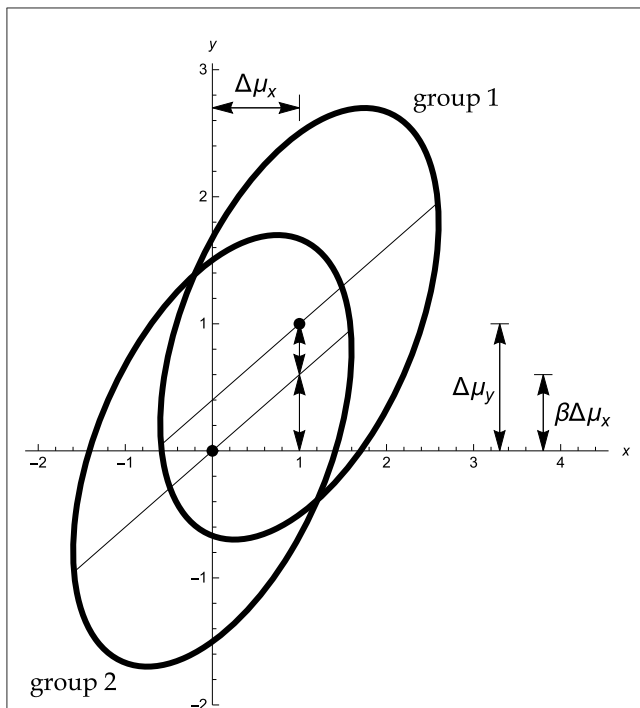
FIGURE 1
Standard Ancova model for random covariates with standardized parameters $\Delta\mu_y = 1$, $\Delta\mu_x = 1$, and $\varrho = 0.6$. Abscissa: covariate x, ordinate: dependent variable y. Part of the net mean group difference $\Delta\mu_y = 1$ in the dependent variable y can be explained in terms of the mean group difference in the covariate x, namely, $\beta\,\Delta\mu_x = 0.6$, so that there remains an adjusted mean difference equal to $\Delta\mu_y - \beta\,\Delta\mu_x = 0.4$. The H_0 is that in the population the mean difference in y is as predicted, via the regression coefficient β, by the difference in the covariate means, that is, $\Delta\mu_y = \beta\,\Delta\mu_x$. The Ancova group test evaluates whether the sample $\Delta\bar{y}$ differs from $b\,\Delta\bar{x}$ by more than chance.

denotes the group, and $j = 1, \ldots, n$ the unit within each group. As in the case of a simple $t-$test, a principal aim then is to test for a difference in the population means of the dependent variable. The specific feature of an Ancova is that it takes the covariate values x_{ij} into account in order to carry out that test.

To this end, a number of assumptions are required, as follows. In the standard Ancova model (Figure 1) the dependent variable y and the covariate x follow, separately in each group, a bivariate normal distribution. The variance–covariance matrix of these distributions is arbitrary but it must be identical in both groups. Let the population variances of the covariate x and the dependent variable y be denoted as σ_x^2, σ_y^2, and their within-group correlation as ϱ. These assumptions imply that in both populations the slope of the regression of y on x is equal to $\beta = \varrho\frac{\sigma_y}{\sigma_x}$. The population means of x, y in group 1 are μ_{x_1}, μ_{y_1}, respectively, and they are μ_{x_2}, μ_{y_2} in group 2. It is convenient to denote as $\Delta\mu_y = \mu_{y_1} - \mu_{y_2}$, and correspondingly $\Delta\mu_x = \mu_{x_1} - \mu_{x_2}$. In the following this set of standard assumptions is referred to as the *bivariate normal Ancova model* (cf., Schneider et al., 2015; Winer et al., 1991, p. 770ff).

In an alternative Ancova model variant the covariates x_{ij} are considered to be fixed (e.g., Cohen, 2009, ch. 9; Maxwell et al., 2018, ch. 9). Significance statements and parameter estimation then refer to hypothetical replications always using the same, given values x_{ij}

of the covariate. This fixed-covariate model may be interpreted as a special case of the model in Figure 1, such that it corresponds to the bivariate normal model but conditional on the realization of a specific set of values x_{ij}. While the fixed and random covariate frameworks test the same hypotheses, the difference between them is crucial for power analysis and sample-size planning (cf., Shieh, 2017, 2021). The fixed-covariate model is plausible, for example, when the covariate is limited to a small set of fixed values which are under the control of the investigator (e.g., when the participants of a study comparing two drugs are paid 100, 125, 150, 175, or 200 USD for their participation), and the intended inference is limited to these specific values. In more typical Ancova applications such as studies A and B, though, the investigator has no direct control over the covariate values, and then the appropriate assumption clearly is that hypothetical replications will generate a different set of covariate values, as assumed under the bivariate normal model. In many Ancova applications, the dependent variable and the covariate actually measure the same variable before and after a treatment; it is then inconsistent to treat the first measurement of that variable as fixed but the second as random.

The basic conceptual logic underlying an Ancova group comparison is simple, even if this simplicity tends to be clouded by technical details. Suppose the dependent variable is related to the covariate in the manner shown qualitatively in Figure 1, and that the treatment exerts no genuine effect of its own. Then we would expect that the group means of y differ by an amount that is proportional to the group difference in the covariate, as indicated by the within-group regression slope. To the degree that the actually observed difference in the group means of y differs from that prediction, there is evidence that the treatment adds a separate effect that can not simply be ascribed to differences in the covariate.

A central consideration of the Ancova is that even when the treatments are allocated at random to the units (so that $\mu_{x_1} = \mu_{x_2}$, as would be expected in Study A) in any particular sample the means of the covariate will practically never be exactly the same in the two groups, just as even a perfectly symmetric coin will in 100 tosses rarely land on head and tail exactly 50 times. There will then nearly always be some imbalance in the sample with respect to the covariate, and if differences in the covariate are associated with differences in the dependent variable, the comparison is biased. This point applies with even more force when there are systematic differences with respect to the covariate between the two groups (i.e., when $\mu_{x_1} \neq \mu_{x_2}$), as would be expected in non-randomized studies using intact (pre-existing) groups, such as Study B above. An Ancova seeks to compare the means of the dependent variable based on a statistical adjustment of these potential imbalances.

In the notation above, the population regressions (i.e., the conditional expectation of the dependent variable y, given a specific covariate value of x) for the groups are

$$\widehat{y}_1 = (\mu_{y_1} - \beta\mu_{x_1}) + \beta x \qquad (1.a)$$
$$\widehat{y}_2 = (\mu_{y_2} - \beta\mu_{x_2}) + \beta x \qquad (1.b)$$

The assumption of a common variance-covariance matrix implies that the regression lines in the two groups are parallel, as shown in Figure 1. Thus, for any specific value x of the covariate the expected difference in the dependent variable, that is, the vertical separation

of the regression lines $\hat{y}_1$ and $\hat{y}_2$ in Figure 1, is independent of x, and equals

$$
\begin{aligned}
\hat{y}_1 - \hat{y}_2 &= (\mu_{y_1} - \beta \mu_{x_1}) - (\mu_{y_2} - \beta \mu_{x_2}) \\
&= (\mu_{y_1} - \mu_{y_2}) - \beta(\mu_{x_1} - \mu_{x_2}) = \Delta \mu_y - \beta \Delta \mu_x \quad (2)
\end{aligned}
$$

If the treatment has no effect in the population then we expect no difference in the dependent variable for any given value x of the covariate. In view of Equation 2, the H_0 of the Ancova group comparison thus states that in the population

$$
\Delta \mu_y = \beta \Delta \mu_x \quad (3)
$$

The Ancova between-group test evaluates if the corresponding sample estimates $\Delta \bar{y}$ and $b\,\Delta \bar{x}$ differ by more than chance.

Note especially that the H_0 tested by an Ancova is *not* that the marginal means of the dependent variable are equal, $\mu_{y_1} = \mu_{y_2}$ (i.e., $\Delta \mu_y = 0$), as would be the case in a standard t−test ignoring the covariate x. Rather, the H_0 in an Ancova group comparison is that the means of the dependent variable are equal after they have been adjusted for imbalances of the covariate. Specifically, in Study B the H_0 tested by an Ancova asserts that male and female graduates of the same height would on average show the same performance.

Two special cases illustrate the nature of the logic of adjustment underlying an Ancova particularly well. In a randomized design we must have $\Delta \mu_x = 0$, as each unit is equally likely to receive the one or other treatment. Thus, differences in the covariate between the two groups are unsystematic, and will arise only on a random basis, as in Study A. Even if random, in any specific sample there will still arise a nonzero difference $\Delta \bar{x}$ of the covariate sample means, and the Ancova then adjusts the comparison of the means of the dependent variable for this random component.

Another special case arises when the covariate is unrelated to the dependent variable, $\varrho = 0$, which implies that the population regression slope $\beta = 0$. There may then well be (random or systematic) differences between the two groups regarding the covariate but these differences would not systematically influence the comparison of the means of the dependent variable. Again, even if x and y are unrelated in the population, in any specific sample the slope estimate b would typically not be equal to zero, and an adjustment would be applied to compare the means. As explained below, if $\varrho = 0$ this adjustment adds a noise component that reduces the sensitivity of the comparison, relative to an analysis igoring the covariate altogether.

3 Sample statistics and main distributional results

Let the sample means of the covariate and the dependent variable in group $i = 1, 2$ be denoted as $\bar{x}_i$ and $\bar{y}_i$, respectively. Their differences in the sample are

$$
\Delta \bar{x} = \bar{x}_1 - \bar{x}_2 \quad (4.a)
$$
$$
\Delta \bar{y} = \bar{y}_1 - \bar{y}_2 \quad (4.b)
$$

In the usual summation notation the sample sums of squares and products are

$$
s_{xx} = \sum_{i=1}^{2} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2 \quad (5.a)
$$

$$
s_{yy} = \sum_{i=1}^{2} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 \quad (5.b)
$$

$$
s_{xy} = \sum_{i=1}^{2} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i) \quad (5.c)
$$

Note that all sums are taken relative to the means of their respective group, i. The sample estimate of the (common) slope β, and the sample estimate of the (common) squared correlation ϱ^2 are then

$$
b = \frac{s_{xy}}{s_{xx}} \quad \text{and} \quad r^2 = \frac{s_{xy}^2}{s_{xx} \cdot s_{yy}} \quad (6)
$$

The mean squared error of the bivariate normal Ancova model is estimated in the sample as

$$
\text{MSE} = (1 - r^2) \cdot \frac{s_{yy}}{2n - 3} \quad (7)
$$

which is nearly always, and often considerably, smaller than the MSE of $\frac{s_{yy}}{2n-2}$ in an Anova of the dependent variable.

The sample estimate of the mean squared treatment effect, adjusted for the covariate is

$$
\text{MST} = \frac{(\Delta \bar{y} - b \Delta \bar{x})^2}{\frac{2}{n}\left[1 + \frac{n}{2}\frac{(\Delta \bar{x})^2}{s_{xx}}\right]} = n(n-1) \cdot \frac{(\Delta \bar{y} - b \Delta \bar{x})^2}{2(n-1) + f} \quad (8)
$$

Note that if the covariate means do not differ in the two samples ($\Delta \bar{x} = 0$) then MST reduces to the standard expression $\frac{n}{2} \cdot (\Delta \bar{y})^2$ used in the analysis of variance. The conceptually important quantity f in Equation 8 is defined as

$$
f = n(n-1) \cdot \frac{(\Delta \bar{x})^2}{s_{xx}} = \frac{n}{2} \cdot \left(\frac{\Delta \bar{x}}{s_x}\right)^2 \quad (9)
$$

where $s_x^2 = s_{xx}/[2(n-1)]$ is the usual unbiased sample estimate of the variance σ_x^2; note that $\frac{\Delta \bar{x}}{s_x}$ is Cohen's standardized mean difference d_p (Goulet-Pelletier and Cousineau, 2018). According to Equation 9, the sample statistic f is equal to the value of the standard statistic $t_{2(n-1)}^2 = F_{1,2(n-1)}$ used to compare the covariate means $\bar{x}_1$ and $\bar{x}_2$ in the two groups[1]. It thus measures the imbalance of the sample covariate means between the two groups; according to Equation 8, large values of f will generally reduce the value of MST. As a rough rule of thumb, for $n > 5$ a value of about $f > 5$ suggests that the covariate means differ by more than chance.

The F−statistic for the comparison of the adjusted group means is, as usual, computed in the sample as the ratio of the treatment and error mean squares

$$
F_{1,2n-3} = \frac{\text{MST}}{\text{MSE}} \quad (10)
$$

---

1 The notation t_m denotes a t−distributed variate with m degrees of freedom. Similarly, $F_{m,n}$ denotes an F−distributed variate with m numerator and n denominator degrees of freedom.

If the H_0 described in Equation 3 is true, that is, if $\Delta\mu_y = \beta\Delta\mu_x$, then the sample statistic defined in Equation 10 has a central $F_{1,2n-3}$—distribution. This result ensures that we can compare the sample statistic $F_{1,2n-3}$ to the critical F—value at the desired level of significance. As explained below, this remains true even when the covariates x_{ij} are considered as fixed values.

Alternatively, the t—statistic for the adjusted mean difference takes the standard form

$$t_{2n-3} = \frac{\Delta\bar{y} - b\,\Delta\bar{x}}{\text{s.e.}\,(\Delta\bar{y} - b\,\Delta\bar{x})} \qquad (11)$$

and the standard error of the adjusted mean difference is estimated in the sample as

$$\text{s.e.}\,(\Delta\bar{y} - b\,\Delta\bar{x}) = \sqrt{\text{MSE} \cdot \frac{2}{n} \cdot \left[1 + \frac{n}{2} \cdot \frac{(\Delta\bar{x})^2}{s_{xx}}\right]} \qquad (12)$$

(for related results, see Groß and Möller, 2024). Again, if $\Delta\mu_y = \beta\Delta\mu_x$, then the sample statistic defined in Equation 11 has a central t_{2n-3}—distribution. From Equations 7, 8, the statistic $F_{1,2n-3}$ in Equation 10 for the Ancova group comparison is equal to the square of the statistic t_{2n-3} in Equation 11, and in this sense the two approaches based on t or F are equivalent.

From Equation 11, the point estimate of the adjusted mean difference is $\Delta\bar{y} - b\,\Delta\bar{x}$, and its two-sided $1-\alpha$ confidence interval is computed as

$$\text{C.I.}_{1-\alpha} = (\Delta\bar{y} - b\,\Delta\bar{x}) \pm t_{2n-3}(\alpha/2) \cdot \text{s.e.}\,(\Delta\bar{y} - b\,\Delta\bar{x}) \qquad (13)$$

The $H_0 : \Delta\mu_y = \beta\,\Delta\mu_x$ is rejected by the t— or the F—test at the significance level α if and only if the confidence interval C.I.$_{1-\alpha}$ does not contain the value of zero. From Equation 12, the standard error of the adjusted mean difference increases with the sample imbalance $\Delta\bar{x}$ of the covariate between the two groups, and when $\Delta\bar{x} = 0$ it reduces to the standard form for the difference of two means.

We next turn to the expectation and distribution of these sample statistics under the bivariate normal model in the general case, that is, without assuming the H_0 to hold; as always, distributional results for the non-null case are needed in order to compute statistical power.

The sample estimate MSE of the mean squared error in Equation 7 is unbiased; its expectation thus equals that part, σ_e^2 say, of the variance in the dependent variable not accounted for by the covariate

$$\mathsf{E}[\text{MSE}] = (1 - \varrho^2)\sigma_y^2 = \sigma_e^2 \qquad (14)$$

The sample estimate MST of the adjusted treatment mean square in Equation 8 depends on the individual covariate values x_{ij} only through the mean difference $\Delta\bar{x}$ and the sample statistic f as defined in Equation 9. Given f, the conditional expectation of MST is (cf., Sprott, 1970, Equation 5; Schneider et al., 2015, Equation B11)

$$\mathsf{E}[\text{MST}|f] = (\Delta\mu_y - \beta\Delta\mu_x)^2 \cdot \frac{n(n-1)}{2(n-1)+f} + \sigma_e^2 \qquad (15)$$

As expected, if the $H_0 : \Delta\mu_y = \beta\Delta\mu_x$ tested by the Ancova holds, then $\mathsf{E}[\text{MST}|f]$ reduces to σ_e^2, the expected MSE. Note that if $f = 0$, then $\mathsf{E}[\text{MST}|f]$ will exceed σ_e^2 by the amount $\frac{n}{2}(\Delta\mu_y - \beta\Delta\mu_x)^2$. Recall that the statistic f evaluates the sample imbalance of the covariate between the two groups. To the degree that f is greater than zero it will reduce the amount by which $\mathsf{E}[\text{MST}|f]$ exceeds σ_e^2, that is, it will reduce power.

In the bivariate normal Ancova model the covariate values x_{ij} have a normal marginal distribution within each group, with possibly different means but equal variance. Therefore, the sample statistic f defined in Equation 9 is distributed as a noncentral F—variate

$$f \sim F_{1,2(n-1),\lambda} \qquad (16)$$

where the noncentrality parameter λ has the standard form

$$\lambda = \frac{n}{2} \cdot \left(\frac{\Delta\mu_x}{\sigma_x}\right)^2 \qquad (17)$$

If the expected values of the covariate do not differ between the groups—for example, because a randomized design, as in Study A above, is used—then $\Delta\mu_x = 0$, and in this case f has a *central* F—distribution, that is, $\lambda = 0$.

Integrating Equation 15 across the density of f, the unconditional expected mean square of the treatment is (cf., Schneider et al., 2015, Equation B13)

$$\mathsf{E}[\text{MST}] = (\Delta\mu_y - \beta\Delta\mu_x)^2 \cdot \int_0^\infty \frac{n(n-1)}{2(n-1)+f} \cdot q_{1,2(n-1),\lambda}(f)\,df + \sigma_e^2 \qquad (18)$$

where $q_{1,2(n-1),\lambda}$ is the density of f, that is, of a noncentral $F_{1,2(n-1),\lambda}$, with noncentrality parameter λ. Note that f depends only on the covariates x_{ij}, and so the integral in Equation 18 depends only on the marginal distribution of the covariates x_{ij}, and is independent of the parameters μ_{y_i}, σ_y^2, and ϱ of the basic model in Figure 1. In essence, it is a multiplier of $(\Delta\mu_y - \beta\Delta\mu_x)^2$, defined by the features, as summarized by λ, of the (normal) marginal distribution of the covariate. It may be shown that if $\Delta\mu_x = 0$, as in randomized designs, then $\mathsf{E}[\text{MST}]$ reduces to $\sigma_e^2 + \frac{n(n-1)}{2n-1} \cdot (\Delta\mu_y)^2$.

The conditional distribution function of the Ancova sample F—value, given the statistic f, may be written as

$$\mathsf{P}\left(F = \frac{\text{MST}}{\text{MSE}} \le t \,\middle|\, f\right) = \mathsf{P}(F_{1,2n-3,k} \le t|f) \qquad (19)$$

where $F_{1,2n-3,k}$ follows a noncentral F—distribution with df equal to 1 and $2n - 3$, and noncentrality parameter k, defined by

$$k = k(f) = n(n-1)\frac{\left(\frac{\Delta\mu_y - \beta\Delta\mu_x}{\sigma_e}\right)^2}{2(n-1)+f} \qquad (20)$$

If the sample means of the covariate are equal ($\Delta\bar{x} = 0$) then $f = 0$, and the noncentrality parameter k reduces to the standard form $\frac{n}{2} \cdot \left(\frac{\Delta\mu_y - \beta\Delta\mu_x}{\sigma_e}\right)^2$, whereas greater values of f will reduce the value of k.

The result Equation 19 is conditional on the x_{ij}, as summarized by f. As the noncentrality parameter $k = k(f)$ in Equation 19

depends on the statistic f, we get the unconditional distribution of the sample F statistic by integrating the conditional distribution function across the density $q_{1,2(n-1),\lambda}$ of f, which leads to a mixture of the family $F_{1,2n-3,k}$ across its noncentrality parameter k. The unconditional distribution function of the sample F statistic may thus be represented as

$$\mathsf{P}(F = \frac{\mathrm{MST}}{\mathrm{MSE}} \le t) = \int_0^\infty \mathsf{P}(F_{1,2n-3,k(f)} \le t|f) \cdot q_{1,2(n-1),\lambda}(f)\, df \quad (21)$$

where $q_{1,2(n-1),\lambda}$ is as above. In essence, the sample $F-$value follows a mixture of noncentral $F-$distributions, with a mixing distribution that is itself a noncentral $F-$distribution, namely, $q_{1,2(n-1),\lambda}$.

Equation 21 is a central result that may be used to compute statistical power in the general case under the bivariate normal model (cf., Shieh, 2017, 2021). An important general implication of Equation 21 is that in an Ancova with random covariates the sample F statistic does *not* follow a standard noncentral $F-$distribution but has the more complex mixture structure indicated in Equation 21. If, in contrast, the covariates x_{ij} are considered as fixed (e.g., Cohen, 2009, ch. 8.3.5), then one would use the conditional distribution $\mathsf{P}\left(F_{1,2n-3,k} \le t \middle| f\right)$ of Equation 19, where the noncentrality parameter $k = k(f)$ defined by Equation 20 is determined by the value of f as given by the fixed x_{ij} in the way prescribed by Equation 9.

Note, however, that the noncentrality parameter $k = k(f)$ given in Equation 20 is generally zero if the null hypothesis $\Delta\mu_y = \beta\Delta\mu_x$ holds, independent of the sample statistic f, which depends only on the covariates x_{ij}. In this case, in Equation 21 the first factor of the integrand (which then becomes a central $F-$distribution function, $k = 0$, with df of 1 and $2n - 3$) may be taken before the integral sign, and the remaining integral across the density $q_{1,2(n-1),\lambda}$ necessarily evaluates to 1. Accordingly, in the null case of $\mathrm{H}_0 : \Delta\mu_y = \beta\Delta\mu_x$ the empirical Ancova F will always follow the *central* $F_{1,2n-3}-$distribution, both with fixed and with random covariates. However, in the non-null case $\Delta\mu_y \neq \beta\Delta\mu_x$ the computation of statistical power or required sample sizes must be based on Equation 21 when the covariates vary randomly, or on Equation 19 when the covariates are considered as fixed (cf., Shieh, 2017, 2021).
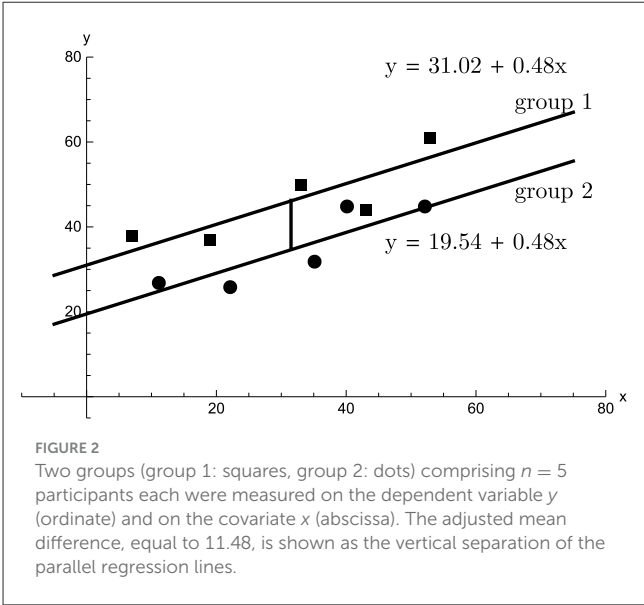
4 A numerical example

These results can be illustrated using the fictitious data set shown in Table 1, Figure 2.

Two groups comprising $n = 5$ participants each were measured on the dependent variable y and on the covariate x. The sample covariate means are $\bar{x}_1 = 31$ and $\bar{x}_2 = 32$, a difference of $\Delta\bar{x} = -1$. The sample means of the dependent variable are $\bar{y}_1 = 46$ and $\bar{y}_2 = 35$, so that $\Delta\bar{y} = 11$. The sample sums of squares and products are $s_{xx} = 2366, s_{yy} = 744$, and $s_{xy} = 1143$. From these values by Equation 6 we have $b = 0.483$ and $r^2 = 0.742$. Thus, there is a positive relation between x and y, and a slight covariate imbalance favoring group 2. The estimate of the adjusted

TABLE 1 Two groups comprising $n = 5$ participants each were measured on the dependent variable, y, and on the covariate, x.

Participant within group	Group 1		Group 2	
	covariate x	dep. variable y	covariate x	dep. variable y
1	43	44	52	45
2	33	50	11	27
3	19	37	22	26
4	7	38	40	45
5	53	61	35	32
means	31	46	32	35

The data set is shown in Figure 2.

FIGURE 2
Two groups (group 1: squares, group 2: dots) comprising $n = 5$ participants each were measured on the dependent variable y (ordinate) and on the covariate x (abscissa). The adjusted mean difference, equal to 11.48, is shown as the vertical separation of the parallel regression lines.

mean difference, shown as a vertical line in Figure 2, equals $\Delta\bar{y} - b \cdot \Delta\bar{x} = 11.48$, and by Equations 12, 13 the corresponding 95%−confidence interval is [3.66, 19.30]. Using Equations 7, 8, 10 the Ancova group comparison sample value $F_{1,7} = 12.02$, giving $p = .011$. By comparison, for the same data set an Anova of the change scores produces $F_{1,8} = 3.49, p = .099$, and an Anova of the dependent variable alone, disregarding the covariate altogether, gives $F_{1,8} = 3.25, p = .109$.

5 General discussion

What exactly can be concluded under the bivariate normal Ancova model shown in Figure 1 from a significant $F-$value computed from the sample data in the way indicated by Equations 4–10? Regardless of the design used, a significant sample F indicates, at the chosen level of significance, that the mean group difference in the dependent variable, adjusted for (i.e., conditional on) the covariate, is greater than would be expected by chance. A corresponding confidence interval for the adjusted mean difference can be computed as indicated in Equation 13,

which will, in a proportion of $1 - \alpha$ of all samples, cover the true adjusted mean difference for any set of parameters of the model in Figure 1, including, specifically, the case in which $\Delta\mu_x \neq 0$ (cf., Senn, 2006). This most basic form of an Ancova-related conclusion addresses the purely *statistical* question of whether the observed adjusted mean difference is greater than the error margin, and not the question of how to *interpret* any such difference. It essentially states *that* the adjusted means differ systematically, but—up to ruling out the effect of the covariate—it leaves open *why* they differ.

As in other research contexts, more specific causal interpretations of significant results depend critically on the design of the study (e.g., Huitema, 2011, ch. 8). Specifically, under a nonrandomized design, even a significant mean difference, adjusted for the chosen covariate, can still be related to other, uncontrolled covariates rather than to the factor defining the two groups. In the case of Study B, even if male and female graduates of the same height differ in the maximum height jumped over, this adjusted effect might still be unrelated to gender per se, but may reflect, for example, that better training opportunities were provided for male graduates. In a nonrandomized design, an Ancova clearly provides no firm basis to attribute a significant adjusted mean difference in high jump performance, specifically, to the group variable gender, even though it does provide a valid evaluation of the statistical significance of the adjusted mean group difference (for a recent account of how to interpret effects in observational designs in terms of acyclic directed graphs, see Cinelli et al., 2024). On the other hand, in the context of a randomized study a significant difference in the adjusted means can be attributed specifically to the factor that defines group membership. For example, in the randomized Study A, a significant mean difference in drinking time, adjusted for the urge to consume alcoholic beverages, leads to the more specific interpretation that the glass shape was the cause of the adjusted differences observed.

These main forms of potential conclusions that can be drawn from an Ancova are complemented by various aspects related to the application and interpretation of an Ancova, which I discuss next in turn.

6 The standardized Ancova model

It is readily seen from the basic distributional features, such as Equations 9, 17, 20, that as far as significance and power are concerned the bivariate normal Ancova model may be standardized with no change of probability statements. Specifically, we may set $\mu_{x_2} = \mu_{y_2} = 0$ and $\sigma_x = \sigma_y = 1$ in which case only three effective parameters of the standardized population model remain. These three parameters are the standardized distance $\Delta\mu_y$ of the marginal group means in the dependent variable y, the standardized distance $\Delta\mu_x$ of the marginal group means in the covariate x, and the correlation ϱ of x and y which in the standardized model is equal to the within-group regression slope β. Therefore, any parameter combination within the general bivariate normal population model may be reduced to the more manageable standardized Ancova model without affecting power and significance.

7 Statistical power as a function of the correlation between the covariate and the dependent variable
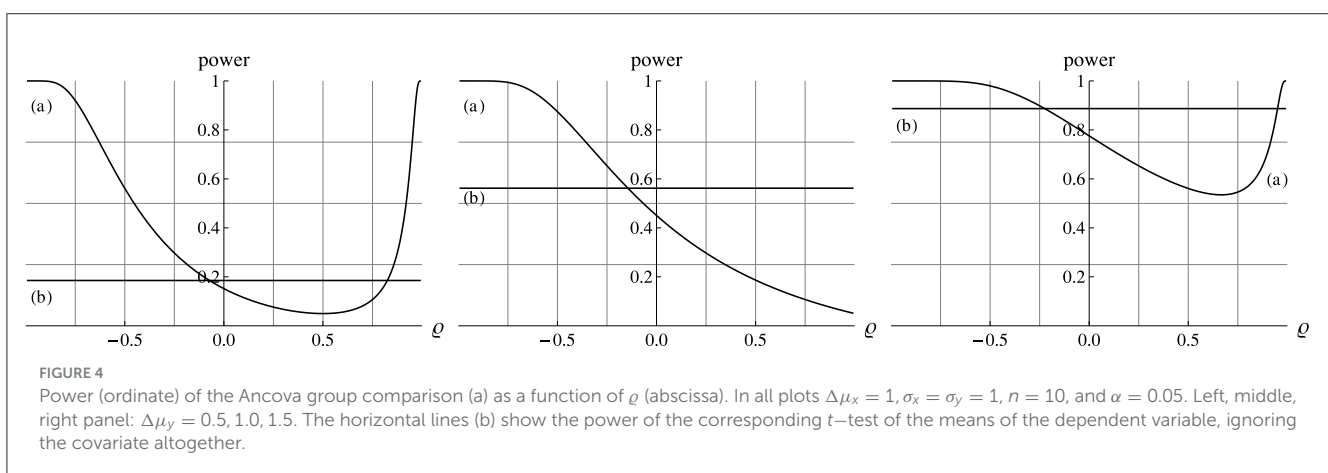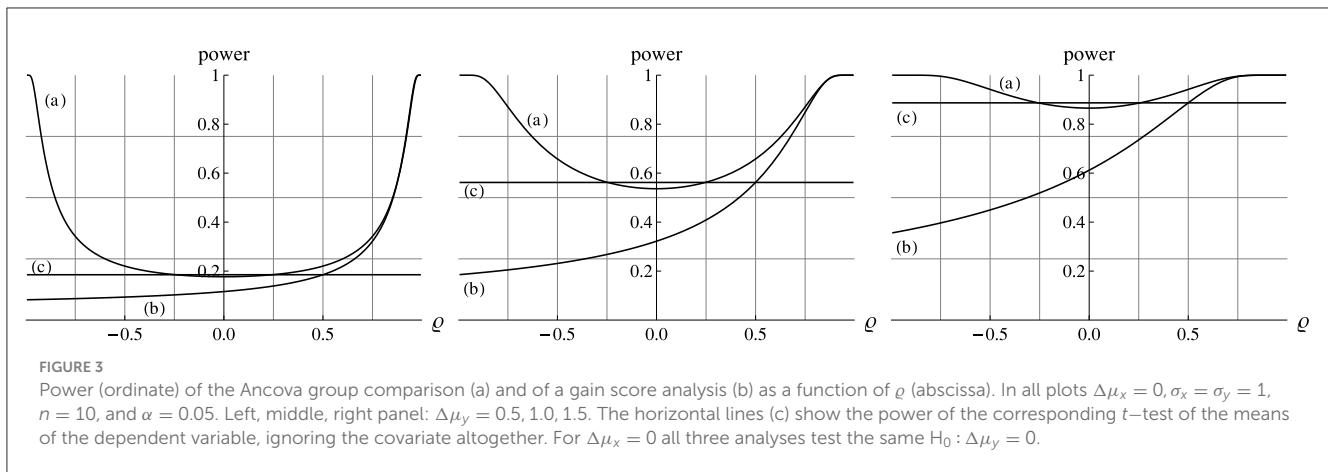
Equation 21 may be used to address the practically relevant question: Which correlation ϱ should the covariate and the dependent variable ideally have in order to maximize power? To address this point succinctly, I will use the standardized formulation of the bivariate normal Ancova model. As we will see, the answer depends on whether the study is randomized (so that $\Delta\mu_x = 0$, as in Study A) or not ($\Delta\mu_x \neq 0$, as in Study B).

In principle, increasing the correlation ϱ has two quite separate effects. First, by Equations 7, 14 an increase in ϱ reduces the error variance $\sigma_e^2 = \sigma_y^2(1 - \varrho^2)$, which by Equation 10 in turn increases the empirical $F_{1,2n-3}$−values for the adjusted group comparison, and thus power. Second, all other aspects equal, an increase of ϱ also increases the regression slope $\beta = \varrho\frac{\sigma_y}{\sigma_x}$, and thus typically reduces the size $\Delta\mu_y - \beta \cdot \Delta\mu_x$ of the adjusted mean difference (cf., Figure 1). Put simple, more of the total increase $\Delta\mu_y$ in mean y may then be explained on the basis of the increase $\Delta\mu_x$ in mean x. The first of these counteracting effects will tend to be dominant if $\Delta\mu_x$ is small or zero, but if $\Delta\mu_x$ is medium or large then increasing ϱ may entail a considerable loss of power. I discuss these two scenarios separately.

If $\Delta\mu_x = 0$ then for any ϱ (and thus any β) Equation 3 for the Ancova group comparison reduces to testing $H_0 : \Delta\mu_y = 0$. Figure 3 indicates that for $\Delta\mu_x = 0$ statistical power then generally increases as ϱ^2 increases. For $\varrho = 0$ power is slightly lower than that of an independent t−test that ignores the covariate altogether. The reason is that for $\varrho = 0$ the regression-based Ancova adjustment of the sample means leads to a loss of one degree of freedom, and adds a noise component that reduces the sensitivity of the comparison. This effect is typically small unless n is very small, and it is soon compensated and then superseded as ϱ^2 increases. Thus, in a randomized study a covariate that correlates strongly with the dependent variable within each group will be useful in increasing power.

These simple relations become considerably more complex if $\Delta\mu_x \neq 0$, as illustrated in Figure 4 for the case of $\Delta\mu_x = 1$. By Equation 3, the Ancova group comparison tests if $\Delta\mu_y - \beta\Delta\mu_x = 0$, and in the standardized model version we have $\beta = \varrho$. Therefore, the H_0 holds if $\varrho = \frac{\Delta\mu_y}{\Delta\mu_x}$, at which point the power function reaches its minimum, the level α.

For example, in the left panel of Figure 4 ($\Delta\mu_y = 0.5$) the adjusted mean difference $\Delta\mu_y - \beta\Delta\mu_x$ falls to the H_0 value of zero at $\varrho = \beta = 0.5$, at which point the power curve reaches the minimum value of $\alpha = .05$. In the middle panel ($\Delta\mu_y = 1.0$) that minimum is just reached at the maximum correlation, $\varrho = 1.0$, when $\Delta\mu_y - \beta\Delta\mu_x = 0$, so that in this scenario power generally decreases as the correlation ϱ increases. In the right panel of Figure 4 ($\Delta\mu_y = 1.5$) the adjusted mean difference $\Delta\mu_y - \beta\Delta\mu_x$ remains positive even for the strongest correlation but it declines as $\varrho = \beta$ increases. However, beyond $\varrho = 0.67$ this decline is more than compensated by the simultaneous reduction of the error variance σ_e^2, so that power increases again beyond that point. In summary, to maximize power in a nonrandomized study design ($\Delta\mu_x \neq 0$) it is critical to select a covariate whose correlation with

FIGURE 3

Power (ordinate) of the Ancova group comparison (a) and of a gain score analysis (b) as a function of ϱ (abscissa). In all plots $\Delta\mu_x = 0$, $\sigma_x = \sigma_y = 1$, $n = 10$, and $\alpha = 0.05$. Left, middle, right panel: $\Delta\mu_y = 0.5, 1.0, 1.5$. The horizontal lines (c) show the power of the corresponding $t-$test of the means of the dependent variable, ignoring the covariate altogether. For $\Delta\mu_x = 0$ all three analyses test the same $H_0 : \Delta\mu_y = 0$.

FIGURE 4

Power (ordinate) of the Ancova group comparison (a) as a function of ϱ (abscissa). In all plots $\Delta\mu_x = 1$, $\sigma_x = \sigma_y = 1$, $n = 10$, and $\alpha = 0.05$. Left, middle, right panel: $\Delta\mu_y = 0.5, 1.0, 1.5$. The horizontal lines (b) show the power of the corresponding $t-$test of the means of the dependent variable, ignoring the covariate altogether.

the dependent variable is not close to $\varrho = \frac{\Delta\mu_y}{\Delta\mu_x}$ in the neighborhood of which the power curve drops to a minimum of α.

8 Ancova and gain score analysis

In some study designs to which an Ancova is applied, the covariate and the dependent variable are commensurate—typically measuring the same variable before and after a treatment. In such cases, a popular alternative to an Ancova is the comparison between groups of the mean difference, or *gain*, scores $\bar{y}_i - \bar{x}_i$, and the hypothesis being tested is that in the population the mean gain scores are the same in both groups. In the present notation, an equivalent formulation of this hypothesis is $H_0 : \Delta\mu_y = \Delta\mu_x$. A gain score analysis tests this H_0 by a $t-$test or an Anova comparing the mean gain scores between the two groups. In the latter case, the sample $F-$value has 1 and $2(n-1)$ degrees of freedom, and the associated noncentrality parameter is readily shown to be

$$\theta = \frac{n}{2} \cdot \left(\frac{\Delta\mu_y - \Delta\mu_x}{\sigma_{y-x}} \right)^2 \qquad (22)$$

where $\sigma_{y-x}^2 = \sigma_x^2 + \sigma_y^2 - 2\varrho\sigma_x\sigma_y$. Note that as ϱ decreases the noncentrality parameter θ – and thus the statistical power of a gain score analysis – generally decreases.

It is essential to appreciate the relationship between an Ancova and a gain score analysis (e.g., van Breukelen, 2006; Samuels, 1986; Senn, 2006). The formulation of the $H_0 : \Delta\mu_y = \Delta\mu_x$ underlines the central point, first emphasized by Lord (1967), that in general a gain score analysis and an Ancova test *different* hypotheses, each of which may or may not be rejected for any given data set. Specifically, an Ancova focuses on (the differences in) *conditional* means—namely, of y, given x—whereas a gain score analysis focuses on (the differences in) *marginal* means.

Comparing the $H_0 : \Delta\mu_y = \Delta\mu_x$ of a gain score analysis to the corresponding Ancova formulation $H_0 : \Delta\mu_y = \beta\Delta\mu_x$ in Equation 3, we see that under the general bivariate normal model in Figure 1 the two hypotheses tested will coincide if either $\Delta\mu_x = 0$ (i.e., under a randomized design) or $\beta = 1$. Note that although the *hypothesis* tested under those two scenarios is the same, the *test statistic* computed from the sample in a gain score analysis and in an Ancova is not, as can be seen from the fact that the sample $F-$values have $(1, 2n-2)$ degrees of freedom for a gain score analysis, and $(1, 2n-3)$ for an Ancova. Also note that under the model in Figure 1, the gain scores in both groups are in any case—not just for $\beta = 1$ or $\Delta\mu_x = 0$ – normally distributed with equal variance, and thus satisfy the formal requirements of a valid $t-$ or $F-$test. Figure 3 illustrates that when the variances of the covariate and the dependent variable are equal ($\sigma_x^2 = \sigma_y^2$) then for all correlation levels and for all effect sizes, an Ancova is more powerful than a

gain score analysis under a randomized design (i.e., $\Delta\mu_x = 0$). When the variances are unequal ($\sigma_x^2 \neq \sigma_y^2$), which is necessarily the case when $\beta = \varrho\frac{\sigma_y}{\sigma_x} = 1$, then an Ancova is not uniformly more powerful than a gain score analysis.

9 Power as a function of the study design

Equations 20, 21 also explain why the statistical power of an Ancova depends critically on the study design. Specifically, the study design determines the noncentrality parameters k in Equation 20 and λ in Equation 17, and according to Equation 21 power will increase with k and decrease with λ.

In a *controlled* design, the researcher choses deliberately all values x_{ij} of the covariate, typically such that $\bar{x}_1 = \bar{x}_2$. Therefore, in Equation 9 the quantity $f = 0$ *in each sample*. As indicated by Equation 19 for $f = 0$, the Ancova sample statistic $F_{1,2n-3}$ as defined in Equation 10 then follows a standard noncentral F−distribution, with a noncentrality parameter k that takes, by Equation 20, the simplified standard form $k = \frac{n}{2} \cdot \left(\frac{\Delta\mu_y}{\sigma_e}\right)^2$.

In a *randomized* design, the units are assigned at random to their treatment group (cf., Study A). In this case $\Delta\mu_x = 0$; the sample means $\bar{x}_1$ and $\bar{x}_2$ will then be equal, not in each individual sample, but *in expectation*, which implies $\lambda = 0$. In a randomized design the Ancova sample statistic $F_{1,2n-3}$ follows no longer a standard noncentral F−distribution (as it does under a controlled design), but rather a mixture of noncentral F−distributions, and the noncentrality parameter k in Equation 20 then takes the form $k = \frac{n(n-1)}{2(n-1)+f} \cdot \left(\frac{\Delta\mu_y}{\sigma_e}\right)^2$ which is smaller than in a controlled design, indicating some loss of power.

In an *observational* design, the researcher has no control over the covariates, and compares pre-existing ("intact") groups, in which typically $\Delta\mu_x \neq 0$, which implies $\lambda > 0$. In this case, the full form given in Equation 20 applies, and the noncentrality parameter k will in typical applications be smaller than in controlled or randomized designs. This is because the sample statistic f measuring the covariate imbalance tends to be larger in an observational design, which by Equation 20 reduces k. Thus, other things equal the power of the Ancova group comparison is largest in controlled designs with $\bar{x}_1 = \bar{x}_2$ (and so $f = 0$), it is intermediate for randomized designs ($\mu_{x_1} = \mu_{x_2}$), and it is lowest with observational designs ($\mu_{x_1} \neq \mu_{x_2}$).

10 Conditional vs. marginal means

As indicated by Equations 1, 2, the Ancova model compares *conditional* group means, namely, the conditional means of the dependent variable y given some specific value x of the covariate. In contrast, standard Anova or t−tests and gain score analyses compare *marginal* means. This important conceptual distinction is sometimes misunderstood. For example, in their otherwise eminent treatment Schneider et al. (2015, p. 2) state that "it is not so widely-known that the validity of an Ancova also depends on [the] assumption ... that the expected value of the covariate is the same

for all of the participants in the experiment". Accordingly, they conclude (p. 3) that "the statistical [Ancova] test for the between-subject main effect is not valid unless $\mu_d = 0$" [i.e., $\Delta\mu_x = 0$, in the present notation]. The distributional results summarized by Equations 19–Equation 21 clearly indicate that this conclusion is incorrect. Even if $\Delta\mu_x \neq 0$, the statistical Ancova model as shown in Figure 1 correctly tests the H_0 of equal conditional means at the specified level of significance, and Equation 13 provides a valid confidence interval for the difference in the conditional means.

The argument of Schneider et al. (for a similar view, see, for example, Miller and Chapman, 2001) is based on the erroneous notion that an Ancova compares the *marginal* means of the dependent variable, and not the *conditional* means, given a specific value of the covariate. It is, of course, true that an Ancova does not usually provide a valid statistical test for the equality of the marginal means of the dependent variable, that is, of the $H_0 : \Delta\mu_y = 0$. However, as indicated by Equation 3, this hypothesis is in fact only tested by an Ancova if $\Delta\mu_x = 0$ (i.e., under a randomized design), or if $\beta = 0$, that is, if the dependent variable and the covariate are uncorrelated. In the general case, the equality of the *marginal* means of y is simply not the statistical question that an Ancova addresses, and it is misleading to fault Ancova techniques for providing its consistent and valid answers (given the model assumptions are met) aimed at comparing *conditional* means. For example, in Study B an Ancova tests if the mean difference in high jump performance between male and female graduates is larger than expected on the basis of the height differences alone. Even if an Ancova indicates, for example, that male graduates jump on average higher by just the amount predicted from the between-group difference in the covariate height, that would not represent a claim that the marginal mean performance in both groups is equal.

The confusion between marginal and conditional means in Ancova techniques is widespread, even in renowned accounts. For example, Vickers and Altman (2001, p. 1123) state that, when the variables are commensurate, the difference in the sample regression intercepts in the two groups (i.e., the adjusted mean difference) "has a useful interpretation: it is the difference between the mean change scores of each group", which is incorrect. In the present notation, "the difference between the mean change scores of each group" in the sample is equal to $\Delta\bar{y} - \Delta\bar{x}$. In contrast, the difference of the sample regression intercepts is given by $\Delta\bar{y} - b \cdot \Delta\bar{x}$, which obviously differs from the marginal differences referred to in the statement of Vickers and Altman—even in their own Table 1, where these quantities are equal to 12.7 and 10.8, respectively.

A main limitation of the present tutorial is its restriction to two groups and a single covariate. This deliberate choice is motivated by two considerations: i.) to present the basic distributional features as clearly as possible, and ii.) by the fact that in many areas, Ancova applications typically refer to this most prominent case. The general case involving more groups, unequal group sizes, or more covariates, follows similar distributional principles as outlined above but requires a more elaborate technical apparatus that tends to cloud the intrinsic simplicity of basic Ancova principles. Excellent recent contributions, such as Huitema (2011, ch. 6), Kutner et al. (2004, ch. 22), Maxwell et al. (2018, ch. 9), Schneider et al. (2015), or Shieh (2017, 2021) offer a broader coverage of Ancova techniques.

Data availability statement

Author contributions

Funding

Acknowledgments

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Attwood, A. S., Scott-Samuel, N. E., Stothart, G., and Munafò, M. R. (2012). Glass shape influences consumption rate for alcoholic beverages. *PLoS ONE*. 7:e43007. doi: 10.1371/journal.pone.0043007

Cinelli, C., Forney, A., and Pearl, J. (2024). A crash course in good and bad models. *Sociol. Methods Res*. 53, 1071–1104. doi: 10.1177/00491241221099552

Cohen, J. (2009). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Psychology Press.

Goulet-Pelletier, J.-C., and Cousineau, D. (2018). A review of effect sizes and their confidence intervals, part I: the Cohen's *d* family. *Quant. Methods Psychol*. 14, 242–265. doi: 10.20982/tqmp.14.4.p242

Groß, J., and Möller, A. (2024). Some additional remarks on statistical properties of Cohen's *d* in the presence of covariates. *Statist. Papers* 65, 3971–3979. doi: 10.1007/s00362-023-01527-9

Huitema, B. E. (2011). *The Analysis of Covariance and Alternatives* (2nd ed.). New York: Wiley.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004). *Applied Linear Statistical Models* (5th ed.). Chicago: McGraw-Hill.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychol. Bullet*. 68, 304–305. doi: 10.1037/h0025105

Maxwell, S. E., Delaney, H. D., and Kelley, K. (2018). *Designing Experiments and Analyzing Data: A Model Comparison Perspective* (3rd ed.). New York, London: Routledge.

Miller, G. A., and Chapman, J. P. (2001). Misunderstanding analysis of covariance. *J. Abnormal Psychol*. 110, 40–48. doi: 10.1037/0021-843X.110.1.40

Samuels, M. L. (1986). Use of analysis of covariance in clinical trials: a clarification. *Control. Clini. Trials* 7, 325–329. doi: 10.1016/0197-2456(86)90039-5

Schneider, B. A., Avivi-Reich, M., and Mozuraitis, M. (2015). A cautionary note on the use of the analysis of covariance (Ancova) in classification designs with and without within-subject factors. *Front. Psychol*. 8:474. doi: 10.3389/fpsyg.2015.00474

Schwarz, W. (2008). *40 Puzzles and Problems in Probability and Mathematical Statistics*. New York: Springer-Verlag.

Schwarz, W., and Reike, D. (2018). Regression away from the mean: theory and examples. *Br. J. Mathem. Statist. Psychol*. 71, 186–203. doi: 10.1111/bmsp.12106

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statist. Med*. 25, 4334–4344. doi: 10.1002/sim.2682

Shieh, G. (2017). Power and sample size calculations for contrast analysis in Ancova. *Multivariate Behav. Res*. 52, 1–11. doi: 10.1080/00273171.2016.1219841

Shieh, G. (2021). Appraising minimum effect of standardized contrasts in Ancova: statistical power, sample size, and covariate imbalance considerations. *Statist. Biopharmaceut. Res*. 13, 468–475. doi: 10.1080/19466315.2020.1788982

Sprott, D. A. (1970). Note on Evans and Anastasio on the analysis of covariance. *Psychol. Bullet*. 73, 303–306. doi: 10.1037/h0028923

van Breukelen, G. J. P. (2006). Ancova vs. change from baseline: more power in randomized studies, more bias in nonrandomized studies. *J. Clin. Epidemiol*. (2006) 59:1334. doi: 10.1016/j.jclinepi.2006.10.002

Vickers, A. J., and Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. *Br. Med. J*. 323, 1123–1124. doi: 10.1136/bmj.323.7321.1123

Winer, B. J., Brown, D. R., and Michels, K. M. (1991). *Statistical Principles in Experimental Design* (3rd ed.). New York: McGraw-Hill.