Transforming ANOVA and Regression statistics for Meta-analysis

David LeBauer

1 Introduction

When conducting a meta-analysis that includes previously published data, differences between treatments reported with P-values, least significant differences (LSD), and other statistics provide no direct estimate of the variance.

In the context of the statistical meta-analysis models that we use, overestimates of variance are okay, because this effectively reduces the weight of a study in the overall analysis relative to an exact estimate, but provides more information than either excluding the study or excluding any estimate of uncertainty (though there are limits to this assumption such as ...).

Where available, direct estimates of variance are preferred, including Standard Error (SE), sample Standard Deviation (SD), or Mean Squared Error (MSE). SE is usually presented in the format of mean (±SE). MSE is usually presented in a table. When extracting SE or SD from a figure, measure from the mean to the upper or lower bound. This is different than confidence intervals and range statistics (described below), for which the entire range is collected.

If MSE, SD, or SE are not provided, it is possible that LSD, MSD, HSD, or CI will be provided. These are range statistics and the most frequently found range statistics include a Confidence Interval (95%CI), Fisher's Least Significant Difference (LSD), Tukey's Honestly Significant Difference (HSD), and Minimum Significant Difference (MSD). Fundamentally, these methods calculate a range that indicates whether two means are different or not, and this range uses different approaches to penalize multiple comparisons. The important point is that these are ranges and that we record the entire range.

Another type of statistic is a "test statistic"; most frequently there will be an F-value that can be useful, but this should not be recorded if MSE is available. Only if there is no other information available should you record the P-value.

2 Solutions

Below is a list of transformations to SE where $SE = \sqrt{MSE/n}$ (Saville 2003) that I am considering, feedback appreciated; below, I assume that $\alpha = 0.05$ so $1 - {}^{\alpha}/_{2} = 0.975$and variables are normally distributed unless otherwise stated:

- given P, n, and treatment means $\bar{X}_1$ and $\bar{X}_2$

$$SE = \frac{\bar{X}_1 - \bar{X}_2}{t_{(1-\frac{P}{2},2n-2)}\sqrt{2/n}}$$

- given LSD (Rosenberg 2004), α, n, b where b is number of blocks, and $n = b$ by default for RCBD

$$SE = \frac{LSD}{t_{(0.975,n)}\sqrt{2bn}}$$

- given MSD (minimum significant difference) (Wang 2000), n, α, df = $2n - 2$

$$SE = \frac{MSD}{t_{(0.975,2n-2}\sqrt{2}}$$

- given a 95% Confidence Interval (Saville 2003) (measured from mean to upper or lower confidence limit), α, and n

$$SE = \frac{CI}{t_{(\alpha/2,n)}}$$

- given Tukey's HSD, n, where q is the 'studentized range statistic',

$$SE = \frac{HSD}{q_{(0.975,n)}}$$

2.1 A summary of equations that can be to convert statistics from P, LSD, or MSD to SE

These are used in the Predictive Ecosystem Analyzer (PEcAn) ecosystem modeling workflow software (LeBauer 2013). Many statistical transformations are implemented in the transformstats (https://github.com/PecanProject/pecan/blob/master/utils/R/transformstats.R) function within the PEcAn.utils package. This function is described in more detail below. However, transformations that require detailed knowledge of the experimental design have not yet been automated within PEcAn (and the BETYdb is not designed to handle all of the information required to get the most precise estimate of uncertainty).

Table 1:

From	To	Conversion	R code	Notes
P	SE	$SE = \frac{\bar{X}_1 - \bar{X}_2}{t_{1-P/2,2n-2}\sqrt{2/n}}$	`(x1-x2)/(qt(1-P/2,2*n-2)*sqrt(2/n))`	$\bar{X}_{1,2}$ are two means being

LSD	SE	$SE = \frac{LSD}{t_{1-\alpha/2,n} * \sqrt{2b}}$	`LSD/(qt(1-P/2,n)*sqrt(2*b))`	compared. where b is the number of blocks, n is the number of replicates, and $n = b$ in a Randomized Complete Block Design
MSD	SE	$SE = \frac{MSD * n}{t_{1-\alpha,2n-2} * \sqrt{2}}$	`msd*n/(qt(1-P/2,2*n-2)*sqrt(2))`	

See related questions on Stats.SE 1 (http://stats.stackexchange.com/q/2917/1381) and 2 (http://stats.stackexchange.com/q/4485/1381)

3 Examples

3.1 Calculating MSE given F, df_{group}, and SS

Given:

F = MS*g*/MS*e*)

Where g indicates the group, or treatment. Rearranging this equation gives: $MS_e = MS_g / F$

Given

$$MS_x = SS_x / df_x$$

Substitute MS_e / df_e for SS_e in the first equation

$$F = \frac{SS_g / df_g}{MS_e}$$

Then solve for MS_e

$$MS_e = \frac{SS_g}{df_g \times F}$$

$$df_{\text{total}} = (df_a + 1) \times (df_b + 1) \ldots \times (n) - 1$$

Which depends on the experimental design:

For factors a, b... (usually 1 or 2, sometimes 3) where n is the number of replicates within each treatment combination.

- One-way anova $df_{\text{total}} = an - 1$, where a is the number of treatments
- Two-way anova without replication $df_{\text{total}} = (a+1)(b+1) - 1$, also known as ''randomized complete block design'' (RCBD)
- Two-way anova with n replicates $df_{\text{total}} = (a+1)(b+1)(n) - 1$, aka ''RCBD with replication''

3.1.1 Application

As an example, we will use the first ANOVA from Table 3 in (Starr 2008). The results are from one (two?) factor ANOVA with repeated measures, with treatment and week as the factors and no replication.

We will calculate MSE from the $SS_{\text{treatment}}$ $df_{\text{treatment}}$, and F-value given in the table; these are 109.58, 2, and 0.570, respectively; df_{weeks} is given as 10.

For the 1997 *Eriphorium vaginatum*, the mean A_{max} in table 4 is 13.49.

Calculate MS_e:

$$MS_e = \frac{109.58}{0.57 \times 2} = 96.12$$

related question on stats.se (http://stats.stackexchange.com/q/4485/1381)

	1997			
	d.f.	Sum of Squares	*F*-value	*P*-value
Eriophorum vaginatum				
Treatment	2	109.58	0.570	0.5812
Weeks	10	2151.52	5.095	0.0001**
treatment*weeks	20	1482.43	1.755	0.0349*

Table used to calculate SE from F in the example; excerpt from (Starr 2008) Table 3.

4 Application

This is an R function to encapsulate these equations, `transformstats` automates transformations of SD, MSE, LSD, 95\%CI, HSD, and MSD to conservative estimates of SE, and is used by PEcAn (LeBauer 2013). The `se.est` function estimates SE through simulation, but is not currently implemented within PEcAn, but has not been implemented.

4.1 Arithmetic Transformations

The `transformstats` function:

```
transformstats <- function(data) {
  if(!"SE" %in% levels(data$statname)){
```

```
    data$statname <- factor(data$statname, levels = c(levels(data$statname), "SE"))
  }
  ## Transformation of stats to SE
  ## transform SD to SE
  if (max(c("SD","sd") %in% data$statname)) {
    sdi <- which(data$statname %in% c("SD","sd"))
    data$stat[sdi] <- data$stat[sdi] / sqrt(data$n[sdi])
    data$statname[sdi] <- "SE"
  }
  ## transform MSE to SE
  if ("MSE" %in% data$statname) {
    msei <- which(data$statname == "MSE")
    data$stat[msei] <- sqrt (data$stat[msei]/data$n[msei])
    data$statname[msei] <- "SE"
  }
  ## 95%CI measured from mean to upper or lower CI
  ## SE = CI/t
  if ("95%CI" %in% data$statname) {
    cii <- which(data$statname == '95%CI')
    data$stat[cii] <- data$stat[cii]/qt(0.975,data$n[cii])
    data$statname[cii] <- "SE"
  }
  ## Fisher's Least Significant Difference (LSD)
  ## conservatively assume no within block replication
  if ("LSD" %in% data$statname) {
    lsdi <- which(data$statname == "LSD")
    data$stat[lsdi] <- data$stat[lsdi] / (qt(0.975,data$n[lsdi]) * sqrt( (2 * data$n[l
sdi])))
    data$statname[lsdi] <- "SE"
  }
  ## Tukey's Honestly Significant Difference (HSD),
  ## conservatively assuming 3 groups being tested so df =2
  if ("HSD" %in% data$statname) {
    hsdi <- which(data$statname == "HSD")
    n = data$n[hsdi]
    n[is.na(n)] = 2 ## minimum n that can be used if NA
    data$stat[hsdi] <- data$stat[hsdi] / (qtukey(0.975, n, df = 2))
    data$statname[hsdi] <- "SE"
    data$n[hsdi] <- n
  }
  ## MSD Minimum Squared Difference
  ## MSD = t_{\alpha/2, 2n-2}*SD*sqrt(2/n)
  ## SE  = MSD*n/(t*sqrt(2))
  if ("MSD" %in% data$statname) {
```

```
    msdi <- which(data$statname == "MSD")
    data$stat[msdi] <- data$stat[msdi] * data$n[msdi] / ( qt(0.975,2*data$n[msdi]-2)*s
qrt(2))
    data$statname[msdi] <- "SE"
  }
  if (FALSE %in% c('SE','none') %in% data$statname) {
    print(paste(trait, ': ERROR!!! data contains untransformed statistics'))
  }
  return(data)
}
```

Example Use:

```
statdf <- data.frame(Y=rep(1,5),
    stat=rep(1,5),
    n=rep(4,5),
    statname=c('SD', 'MSE', 'LSD', 'HSD', 'MSD'))
transformstats(statdf)
```

4.2 Transformation via simulation

As an alternative to solving for SE given n, $\bar{X}_i$ and summary statistics from post-hoc multiple comparisons such as Fisher's LSD and Tukey's HSD, I propose a general approach that can be applied to other post-hoc tests as well as other statistics. Ultimately, such an approach would be less prone to (my own) arithmetic error and easier to implement than trying to analytically solve each new statistic. In practice, this has not been implemented because the analytical solutions above cover the majority of cases. However, this could be a good way to test some of the analytical solutions.

The goal here is to generalize the estimation of SE from other statistics by simulating the assumed data generating models.

Is there a general simulation / optimization approach to this problem?

- $\bar{X}_1$, $\bar{X}_2$, some statistic, and n are given
- I am looking for $SE \pm 10\%$
- As in the earlier post, over estimates of SE are o.k.

The following example implementations would be a solution for SE when the LSD is given:

```
se.est <- function(x1bar, x2bar, lsd_obs, n, se) {
       y <- c(rnorm(n, x1bar, se), rnorm(n, x2bar, se)).
       x <- c(rep(1,n), rep(2,n))
       mse <- sum(lm(y~factor(x))$residuals^2)/(2*n-2)
       lsd_est <- qt(0.975,n)*sqrt(2*mse/2)
       ans <- (lsd_obs - lsd_est)^2
       }

SE <- optimize(par = c(x1bar, x2bar, lsd_obs,n), fn = se.est)
```

alternatively

```
for (i in 1:10000){
     se[i] <- rgamma(1, 1, 0.01)
     y[i] <- c(rnorm(n, x1bar, se), rnorm(n, x2bar, se)).
     x[i] <- c(rep(1,n), rep(2,n))
     mse[i] <- sum(lm(y~factor(x))$residuals^2)/(2*n-2)
     lsd_est[i] <- qt(0.975,n)*sqrt(2*mse/2)
     }

 se <- se[which.min((lsd_est - lsd_obs)^2)]
```

4.3 Related Questions on Stats.SE

- Given sample size, group means, and misc. post-hoc range statistics, can you suggest a good way to estimate variance through simulation? (http://stats.stackexchange.com/questions/3207/given-sample-size-group-means-and-misc-post-hoc-range-statistics-can-you-sug?lq=1)
- Are these formulas for transforming P, LSD, MSD, HSD, CI, to SE as an exact or inflated/conservative estimate of $\hat{\sigma}$ correct? (http://stats.stackexchange.com/questions/2917/are-these-formulas-for-transforming-p-lsd-msd-hsd-ci-to-se-as-an-exact-or-i?lq=1)
- Is this the correct way to calculate MSE from $SS_{\text{treatment(s)}}$, $df_{\text{treatment(s)}}$, F? (http://stats.stackexchange.com/questions/4485/is-this-the-correct-way-to-calculate-mse-from-ss-texttreatments-df?lq=1)

References

David J. Saville. Basic statistics and the inconsistency of multiple comparison procedures.. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* **57**, 167–175 (2003). Link (http://dx.doi.org/10.1037/h0087423)

M. S. Rosenberg, K. A. Garrett, Z. Su, R. L. Bowden. Meta-Analysis in Plant Pathology: Synthesizing Research Results. *Phytopathology* **94**, 1013–1017 (2004). Link (http://dx.doi.org/10.1094/phyto.2004.94.9.1013)

Qin Wang, Debra L. Denton, Rakesh Shukla. Applications and statistical properties of minimum significant difference-based criterion testing in a toxicity testing program. *Environmental Toxicology and Chemistry* **19**, 113–117 (2000). Link (http://dx.doi.org/10.1002/etc.5620190113)

David S. LeBauer, Dan Wang, Katherine T. Richter, Carl C. Davidson, Michael C. Dietze. Facilitating feedbacks between field measurements and ecosystem models. *Ecological Monographs* **83**, 133–154 (2013). Link (http://dx.doi.org/10.1890/12-0137.1)

Gregory Starr, Steven F Oberbauer, Lorraine E Ahlquist. The photosynthetic response of Alaskan tundra plants to increased season length and soil warming. *Arctic, Antarctic, and Alpine Research* **40**, 181–191 (2008).