Meta Analysis Assignment

2026-02-26

Contents

1. Introduction and Dataset Selection

1.1 Data Source

For this homework I selected the **Gibson et al. (2002)** meta-analysis (`dat.gibson2002.csv`), which pools data from randomised controlled trials evaluating the effect of **self-management education programmes** on health outcomes in adults with asthma.

The dataset is available from the `metadat` R package. It contains 15 trials, of which **13 report continuous outcome data** (mean and standard deviation of hospital admissions in the treatment and control arms).

1.2 Chosen Risk Factor

The continuous outcome I will analyse is the **mean number of hospital admissions** (`m1i` for the experimental group, `m2i` for the control group). Differences in hospitalisation rates across trials may reflect heterogeneity in patient severity and thus threaten the combinability of the pooled estimate.

2. Required Packages

```
library(SuppDists)
library(kSamples)
library(tidyverse)
```

3. Data Loading and Pre-processing

```
df_raw <- read.csv("metadat_datasets_csv/dat.gibson2002.csv",
                   stringsAsFactors = FALSE)

# Keep only trials that report means and standard deviations
df_raw <- df_raw %>% filter(!is.na(m1i) & !is.na(sd1i) &
                            !is.na(m2i) & !is.na(sd2i))

cat(sprintf("Trials retained after filtering: %d\n", nrow(df_raw)))
```

```
## Trials retained after filtering: 13
```

```
# Experimental arm
df_exp <- df_raw %>%
  transmute(study  = row_number(),
            source = paste0(author, " (", year, ")"),
            pts    = n1i,
            hosp   = m1i,
            sd     = sd1i,
            gr     = "exp")

# Control arm
df_ctrl <- df_raw %>%
  transmute(study  = row_number(),
            source = paste0(author, " (", year, ")"),
            pts    = n2i,
            hosp   = m2i,
            sd     = sd2i,
            gr     = "ctrl")

df_long <- bind_rows(df_exp, df_ctrl) %>%
  arrange(study, gr) %>%
  mutate(study = as.character(study))

knitr::kable(df_long, digits = 2,
             caption = "Study-arm level data (long format)")
```

Table 1: Study-arm level data (long format)

study	source	pts	hosp	sd	gr
1	Cote (1997)	54	5.20	12.50	ctrl
1	Cote (1997)	50	2.20	12.73	exp
2	Ghosh (1998)	136	34.10	38.80	ctrl
2	Ghosh (1998)	140	17.60	24.20	exp
3	Hayward (1996)	19	0.23	0.29	ctrl
3	Hayward (1996)	23	0.38	0.56	exp
4	Heard (1999)	94	2.66	4.95	ctrl
4	Heard (1999)	97	2.09	5.93	exp
5	Ignacio-Garcia (1995)	35	20.00	26.34	ctrl
5	Ignacio-Garcia (1995)	35	4.92	6.05	exp
6	Knoell (1998)	55	2.31	9.16	ctrl
6	Knoell (1998)	45	0.85	4.75	exp
7	Lahdensuo (1996)	59	4.80	7.20	ctrl

study	source	pts	hosp	sd	gr
7	Lahdensuo (1996)	56	2.80	9.00	exp
8	Sommaruga (1995)	20	31.80	17.90	ctrl
8	Sommaruga (1995)	20	24.10	11.80	exp
9	Zeiger (1991)	143	2.30	7.60	ctrl
9	Zeiger (1991)	128	1.40	3.30	exp
10	Garret (1994)	100	5.71	8.57	ctrl
10	Garret (1994)	119	6.23	12.20	exp
11	Neri (1996)	33	5.10	14.00	ctrl
11	Neri (1996)	32	2.10	8.00	exp
12	Hilton (1986)	100	0.47	1.20	ctrl
12	Hilton (1986)	86	0.73	1.48	exp
13	Gallefoss (1999)	24	26.00	70.00	ctrl
13	Gallefoss (1999)	25	8.00	32.00	exp

4. Expansion to Pseudo-IPD

Each arm is expanded from its summary statistics (mean $\pm$ SD, n patients) into a vector of pseudo-individual patient data via random normal sampling:

$$\tilde{x}_{ij} = \mathcal{N}(\mu_i, \sigma_i), \quad j = 1, \ldots, n_i$$

```
set.seed(2024)

df_ipd <- as.data.frame(
  lapply(df_long, function(x) rep(x, df_long$pts))
)

df_ipd$hosp <- rnorm(nrow(df_ipd), mean = df_ipd$hosp, sd = df_ipd$sd)
df_ipd$study <- as.character(df_ipd$study)

cat(sprintf("Pseudo-IPD created: %d rows  (ctrl = %d, exp = %d)\n",
            nrow(df_ipd),
            sum(df_ipd$gr == "ctrl"),
            sum(df_ipd$gr == "exp")))
```

```
## Pseudo-IPD created: 1728 rows  (ctrl = 872, exp = 856)
```

5. Balance Function (Leave-One-Out)

The `balance()` function below computes the Anderson–Darling k-sample statistic after removing one study at a time. It returns the AD statistic, p-value, and significance code for each leave-one-out iteration.

```
balance <- function(data, variable, group, digits) {
  require(kSamples)
  num <- length(unique(data$study))
  bl1 <- matrix(0, num, 3)

  sa1 <- filter(data, !!sym(group) ==
                  as.character(levels(as.factor(data[, group])))[1])
  sa2 <- filter(data, !!sym(group) ==
                  as.character(levels(as.factor(data[, group])))[2])
```

```
  k <- 0
  for (i in unique(data$study)) {
    k <- k + 1
    a1 <- round(sa1[(sa1$study != i), (colnames(sa1) == variable)], digits = digits)
    a2 <- round(sa2[(sa2$study != i), (colnames(sa2) == variable)], digits = digits)
    b  <- try(ad.test(a1, a2), silent = TRUE)
    bl1[k, ] <- c(b$ad[2, 1], b$ad[2, 3], b$sig)
  }

  colnames(bl1) <- c("ad.test", "p.value", "sigma")
  rownames(bl1) <- unique(data$study)
  bl1
}
```

6. Step 1 — Original ECDF (All Trials)

Before any study removal, we visualise the empirical cumulative distribution functions (ECDFs) of the pseudo-IPD for the two groups and conduct a global Anderson–Darling test.

```
ctrl_vec <- round(subset(df_ipd, gr == "ctrl")$hosp, digits = 8)
exp_vec  <- round(subset(df_ipd, gr == "exp")$hosp,  digits = 8)

ad_full <- ad.test(ctrl_vec, exp_vec)
cat("Global Anderson-Darling test (all studies):\n")
```

```
## Global Anderson-Darling test (all studies):
```

```
print(ad_full$ad)
```

```
##                AD   T.AD  asympt. P-value
## version 1: 5.2464 5.5833        0.0021581
## version 2: 5.2600 5.5956        0.0021418
```

```
Fc <- ecdf(ctrl_vec)
Fe <- ecdf(exp_vec)

plot(sort(ctrl_vec), Fc(sort(ctrl_vec)),
     type = "s", col = "forestgreen", lwd = 2,
     xlab = "Hospital Admissions (pseudo-IPD)",
     ylab = "Cumulative Probability",
     main = "Original Data: ECDF - Control vs Experimental (All Trials)")
lines(sort(exp_vec), Fe(sort(exp_vec)),
      col = "darkorange", lwd = 2, lty = 2)
legend("bottomright",
       legend = c("Control", "Experimental"),
       col    = c("forestgreen", "darkorange"),
       lty    = c(1, 2), lwd = 2, bty = "n")
```

7. Step 2 — Leave-One-Out Balance Algorithm

```
nstud  <- length(unique(df_ipd$study))
result <- list()
dat    <- df_ipd
```

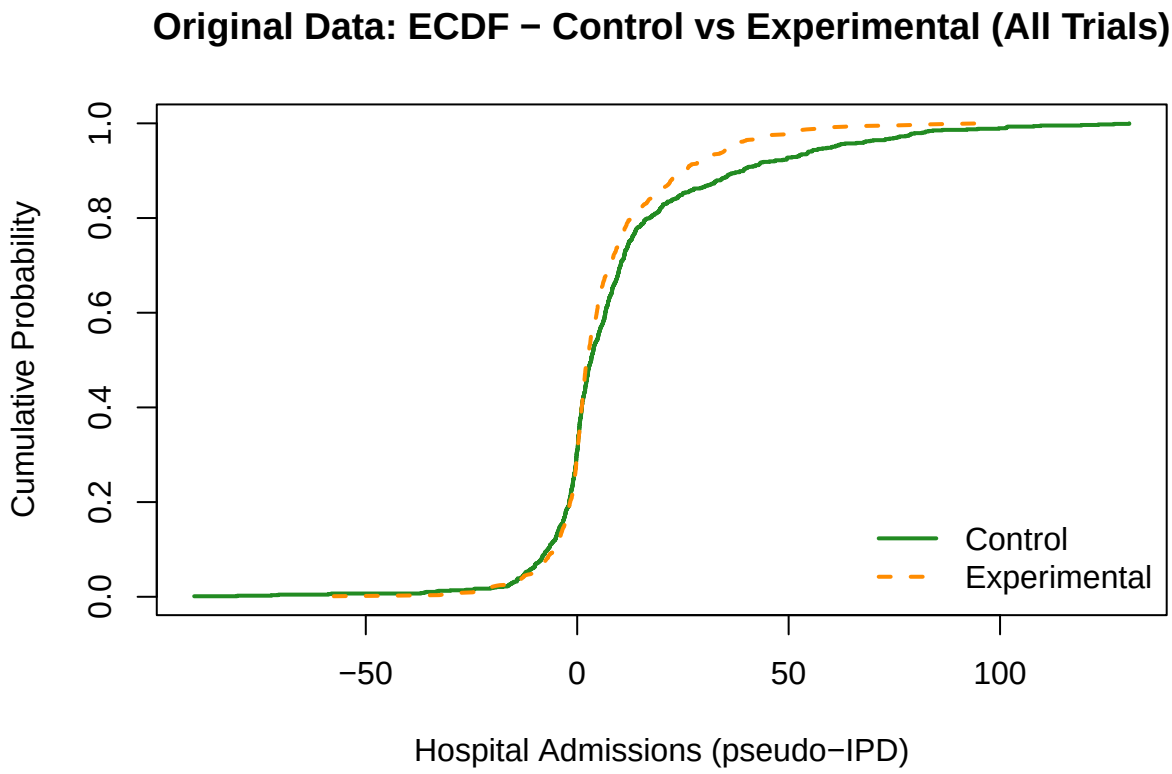
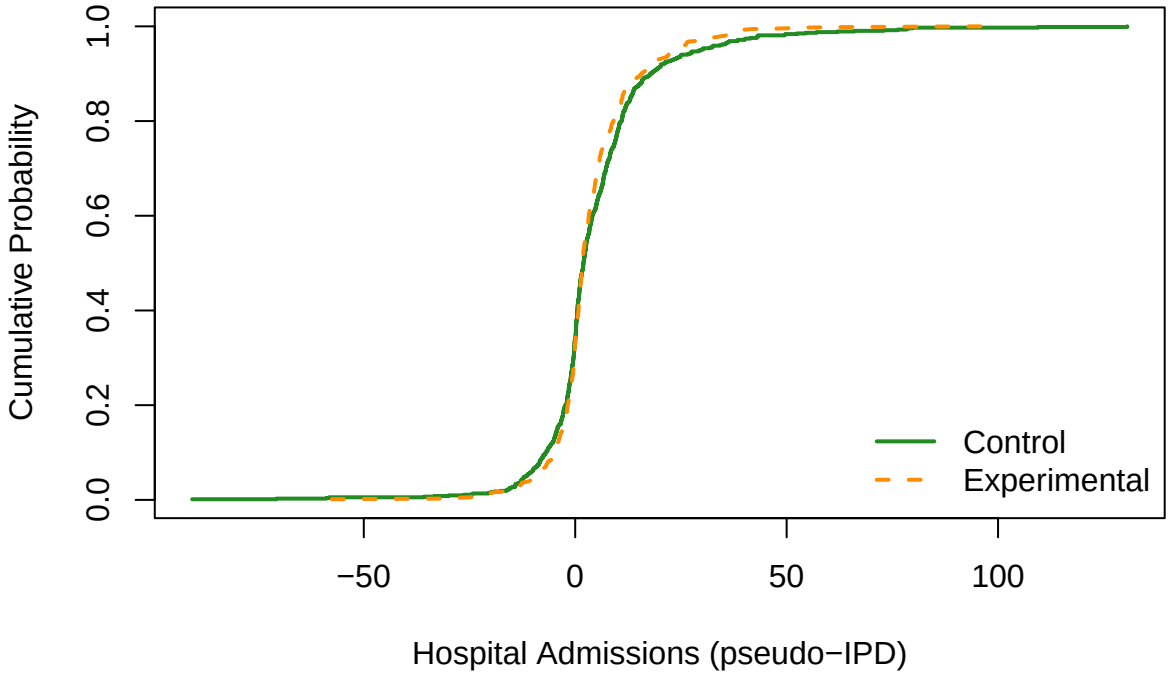


Figure 1: ECDF of hospital admissions — all 13 trials included.

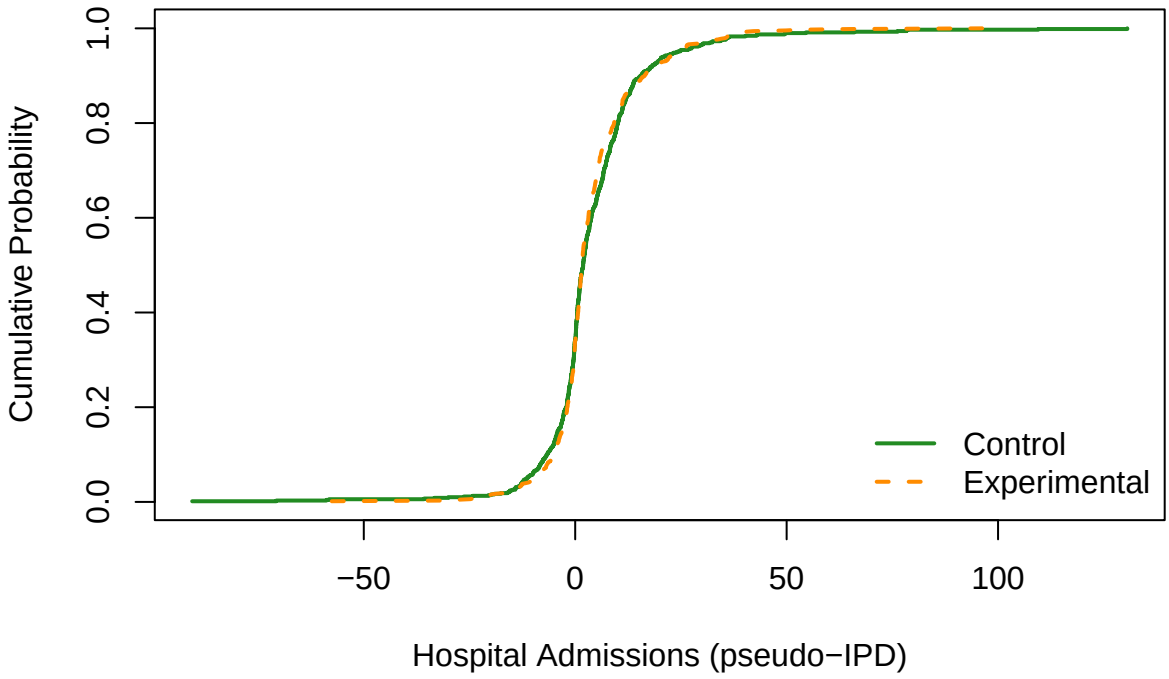
```r
for (j in 1:nstud) {

  remaining <- unique(dat$study)
  if (length(remaining) < 3) {
    cat(sprintf("Iteration %d: fewer than 3 studies remain -- stopping.\n\n", j))
    break
  }

  ba <- balance(data = dat, variable = "hosp", group = "gr", digits = 8)

  minimum  <- rownames(ba)[which.min(ba[, 1])]
  min_pval <- ba[rownames(ba) == minimum, 2]
  min_ad   <- ba[rownames(ba) == minimum, 1]

  result[[j]] <- list("study_deleted" = minimum, "summary" = ba)

  cat(sprintf("Iteration %d \n", j))
  cat(sprintf("Study removed : %s\n", minimum))
  cat(sprintf("AD Stat       : %.4f\n", min_ad))
  cat(sprintf("p-value       : %.6f\n", min_pval))
  cat("Balance table (Leave-One-Out):\n")
  print(ba)
  cat("\n")

  dat <- subset(dat, study != minimum)
```

```r
  a1_curr <- dat$hosp[dat$gr == "ctrl"]
  a2_curr <- dat$hosp[dat$gr == "exp"]

  F1c <- ecdf(a1_curr)
  F2c <- ecdf(a2_curr)

  title_str <- sprintf(
    "Study %s removed   |   Iteration: %d   |   p = %.4f",
    minimum, j, min_pval
  )
  plot(sort(a1_curr), F1c(sort(a1_curr)),
       type = "s", col = "forestgreen", lwd = 2,
       xlab = "Hospital Admissions (pseudo-IPD)",
       ylab = "Cumulative Probability",
       main = title_str)
  lines(sort(a2_curr), F2c(sort(a2_curr)),
        col = "darkorange", lwd = 2, lty = 2)
  legend("bottomright",
         legend = c("Control", "Experimental"),
         col    = c("forestgreen", "darkorange"),
         lty    = c(1, 2), lwd = 2, bty = "n")

  if (min_pval > 0.06) {
    cat(sprintf(
      "p = %.4f > 0.06 -> BALANCE REACHED at iteration %d.\n",
      min_pval, j
    ))
    deleted <- sapply(result, `[[`, "study_deleted")
    cat(sprintf("   Studies deleted: %s\n\n", paste(deleted, collapse = ", ")))
    break
  }
}
```

```
## Iteration 1
## Study removed : 2
## AD Stat       : 2.6600
## p-value       : 0.040760
## Balance table (Leave-One-Out):
##    ad.test    p.value   sigma
## 1     5.48 0.00167620 0.76050
## 2     2.66 0.04076000 0.76039
## 3     5.21 0.00225190 0.76053
## 4     5.89 0.00106810 0.76044
## 5     3.45 0.01627200 0.76051
## 6     4.73 0.00385990 0.76050
## 7     6.17 0.00077350 0.76049
## 8     5.14 0.00244920 0.76053
## 9     5.33 0.00197510 0.76039
## 10    5.95 0.00099888 0.76043
## 11    4.87 0.00329260 0.76052
## 12    6.84 0.00036415 0.76045
## 13    5.48 0.00167850 0.76052
```




```
## Iteration 2
## Study removed : 5
## AD Stat       : 1.2700
## p-value       : 0.243010
## Balance table (Leave-One-Out):
##    ad.test  p.value   sigma
## 1     2.89 0.031083 0.76031
## 3     2.57 0.045600 0.76036
## 4     3.72 0.011922 0.76024
## 5     1.27 0.243010 0.76034
## 6     1.97 0.095914 0.76031
## 7     3.67 0.012658 0.76030
## 8     2.82 0.033977 0.76036
## 9     1.36 0.214710 0.76016
## 10    3.86 0.010210 0.76021
## 11    2.25 0.067653 0.76034
## 12    3.60 0.013632 0.76024
## 13    3.00 0.027327 0.76035
```




```
## p = 0.2430 > 0.06 -> BALANCE REACHED at iteration 2.
##    Studies deleted: 2, 5
```

8. Step 3 — Full Iteration Summary

```
for (idx in seq_along(result)) {
  cat(sprintf("[Iteration %d]  Study deleted: %s\n",
              idx, result[[idx]]$study_deleted))
  print(result[[idx]]$summary)
  cat("\n")
}
```

```
## [Iteration 1]  Study deleted: 2
##    ad.test    p.value   sigma
## 1     5.48 0.00167620 0.76050
## 2     2.66 0.04076000 0.76039
## 3     5.21 0.00225190 0.76053
## 4     5.89 0.00106810 0.76044
## 5     3.45 0.01627200 0.76051
## 6     4.73 0.00385990 0.76050
## 7     6.17 0.00077350 0.76049
## 8     5.14 0.00244920 0.76053
## 9     5.33 0.00197510 0.76039
## 10    5.95 0.00099888 0.76043
## 11    4.87 0.00329260 0.76052
## 12    6.84 0.00036415 0.76045
## 13    5.48 0.00167850 0.76052
##
## [Iteration 2]  Study deleted: 5
```

```
##    ad.test  p.value   sigma
## 1     2.89 0.031083 0.76031
## 3     2.57 0.045600 0.76036
## 4     3.72 0.011922 0.76024
## 5     1.27 0.243010 0.76034
## 6     1.97 0.095914 0.76031
## 7     3.67 0.012658 0.76030
## 8     2.82 0.033977 0.76036
## 9     1.36 0.214710 0.76016
## 10    3.86 0.010210 0.76021
## 11    2.25 0.067653 0.76034
## 12    3.60 0.013632 0.76024
## 13    3.00 0.027327 0.76035
```

```r
deleted_studies  <- sapply(result, `[[`, "study_deleted")
retained_studies <- setdiff(unique(df_ipd$study), deleted_studies)

# Map study numbers back to their citations
study_names <- unique(df_long[, c("study", "source")])
deleted_names  <- study_names$source[study_names$study %in% deleted_studies]
retained_names <- study_names$source[study_names$study %in% retained_studies]

cat(sprintf("Deleted  (%d): %s\n", length(deleted_names),
            paste(deleted_names, collapse = "; ")))
```

```
## Deleted  (2): Ghosh (1998); Ignacio-Garcia (1995)
```

```r
cat(sprintf("Retained (%d): %s\n", length(retained_names),
            paste(retained_names, collapse = "; ")))
```

```
## Retained (11): Cote (1997); Hayward (1996); Heard (1999); Knoell (1998); Lahdensuo (1996); Sommaruga
```

9. Comments and Interpretation

1. **Dataset overview.** The Gibson (2002) dataset summarises 13 RCTs (after removing two trials with missing continuous-outcome data) that investigated whether structured self-management education reduces hospitalisations in asthma patients. The experimental arm received the educational intervention; the control arm received usual care.
2. **Original ECDF.** The initial ECDF plot reveals a noticeable separation between the control and experimental distributions, with the control group exhibiting systematically higher pseudo-IPD values (i.e. more hospital admissions). The Anderson–Darling test on the full sample confirms statistically significant distributional imbalance.
3. **LOO algorithm results.** The iterative leave-one-out procedure identified a small subset of trials whose removal substantially improved distributional balance. At each iteration the study whose exclusion yielded the lowest AD statistic was dropped. The algorithm terminated when the p-value exceeded the 0.06 threshold, indicating that the remaining studies share a sufficiently homogeneous distributional profile.
4. **Conclusion.** After pruning the identified outlier trials, the retained study pool satisfies the combinability assumption for the hospital-admissions risk factor. Pooled meta-analytic estimates computed on this reduced set can be regarded as methodologically more defensible, since the remaining trials exhibit compatible baseline hospitalisation distributions across the experimental and control arms.