Supporting information for 'A classical hypothesis test for assessing the homogeneity of disease transmission in stochastic epidemic models' by Aristotelous, Kypraios and O'Neill

Appendix S1

```
# function that calculates T #
#______________________________#

calculate.T.label <- function(label.group.order,C.group) {

  # argument description #
  # label.group.order: a vector of time-ordered group labels
     corresponding to time-ordered event times
  # C.group: a vector giving the number of individuals in each group

  # set some variables #
  n <- length(label.group.order) # total number of events
  group.number <- length(C.group) # total number of groups

  # calculate the contribution of each group #
  T.label.group <- rep(0,times=group.number) # if a group does not
     appear or has C_m=1 it has contribution 0
  for (k in 1:group.number) {
    times.group.appears <- sum(label.group.order==k)
    if (times.group.appears == 1 & C.group[k] > 1) {
      T.label.group[k] <- n - which(label.group.order==k)
    }
    if (times.group.appears >= 2) {
      T.label.group[k] <- sum(diff(which(label.group.order==k))-1)
    }
  }

  # calculate T.label #
  T.label <- sum(T.label.group)

  # output #
  output.list <- list(T.label=T.label,T.label.group=T.label.group)
  return(output.list)
}
```

```
# Illustration: Abakaliki dataset #
#_________________________________#

# calculate observed value #
C.group <- c(33,15,10,33,22,43,20,42,33) # Eichner and Dietz (2003,
   table 2)
label.group.order.obs <- c(rep(x=1,times=7)
   ,2,2,1,4,5,1,1,1,1,5,2,1,2,6,5,2,7,4,2,2,8,3,9,5,2) # Thompson
   and Foege (1968, table 1))
T.label.obs <- calculate.T.label(label.group.order = label.group.
   order.obs, C.group = C.group)$T.label

# sample from H_0 #
S <- 10000
T.label.rep <- rep(NA,times = S)
n <- length(label.group.order.obs)
group.number <- length(C.group)
for (s in 1:S) {
  label.group.rep <- sample(x=rep(1:group.number,times=C.group),size
     =n)
  T.label.rep[s] <- calculate.T.label(label.group.order = label.
     group.rep, C.group = C.group)$T.label
}

# plot histogram #
hist(T.label.rep,xlab=expression(bold(T^sam)),prob=T,breaks="fd")
abline(v=T.label.obs,col="red",lty=2,lwd=2)

# calculate p-value #
mean(T.label.rep <= T.label.obs)
```

Appendix S2

The random vector $\boldsymbol{g}^{\boldsymbol{e}^{sam}} = (g_1^{\boldsymbol{e}^{sam}}, g_2^{\boldsymbol{e}^{sam}}, ..., g_n^{\boldsymbol{e}^{sam}}) \sim H_0$ has support $g_k^{\boldsymbol{e}^{sam}} \in \{1, 2, ..., l\}$, $k = 1, 2, ..., n$, and for household label vector $\boldsymbol{g}^{\boldsymbol{e}} = (g_1^{\boldsymbol{e}}, g_2^{\boldsymbol{e}}, ..., g_n^{\boldsymbol{e}})$, corresponding to time-ordered event times $\boldsymbol{e} = (e_1, e_2, ..., e_n)$, its joint probability mass function (p.m.f.) $f_{\boldsymbol{g}^{\boldsymbol{e}^{sam}}}(\boldsymbol{g}^{\boldsymbol{e}})$ is given by

$$\begin{aligned} f_{\boldsymbol{g}^{\boldsymbol{e}^{sam}}}(\boldsymbol{g}^{\boldsymbol{e}}) = P(\boldsymbol{g}^{\boldsymbol{e}^{sam}} = \boldsymbol{g}^{\boldsymbol{e}}) &= P(g_1^{\boldsymbol{e}^{sam}} = g_1^{\boldsymbol{e}}, g_2^{\boldsymbol{e}^{sam}} = g_2^{\boldsymbol{e}}, ..., g_n^{\boldsymbol{e}^{sam}} = g_n^{\boldsymbol{e}}) \\ &= P(g_1^{\boldsymbol{e}^{sam}} = g_1^{\boldsymbol{e}}) P(g_2^{\boldsymbol{e}^{sam}} = g_2^{\boldsymbol{e}} \mid \mathcal{H}_{e_2^-}) \dots P(g_n^{\boldsymbol{e}^{sam}} = g_n^{\boldsymbol{e}} \mid \mathcal{H}_{e_n^-}) \end{aligned} \tag{1}$$

where the first of the above terms (the marginal p.m.f. of $g_1^{\boldsymbol{e}^{sam}}$) is calculated as $P(g_1^{\boldsymbol{e}^{sam}} = m) = \frac{C_m}{C}$, for $m = 1, 2, ..., l$, and the remaining terms (the marginal p.m.f. of $g_k^{\boldsymbol{e}^{sam}}$, conditioned on $\mathcal{H}_{e_k^-}$, $k = 2, 3, ..., n$) as $P(g_k^{\boldsymbol{e}^{sam}} = m \mid \mathcal{H}_{e_k^-}) = \frac{C_m - \nu^{(m)}_{\mathcal{H}_{e_k^-}}}{C-(k-1)}$, for $m = 1, 2, ..., l$, $k = 2, 3, ..., n$, where $\mathcal{H}_{t^-}$ denotes the history of the process up to time t^- (where t^- is the time just before time t) and $\nu^{(m)}_{\mathcal{H}_{t^-}}$ denotes the number of times that the label of household m appears in $\boldsymbol{g}^{\boldsymbol{e}}$, up to time t^-.

Appendix S3

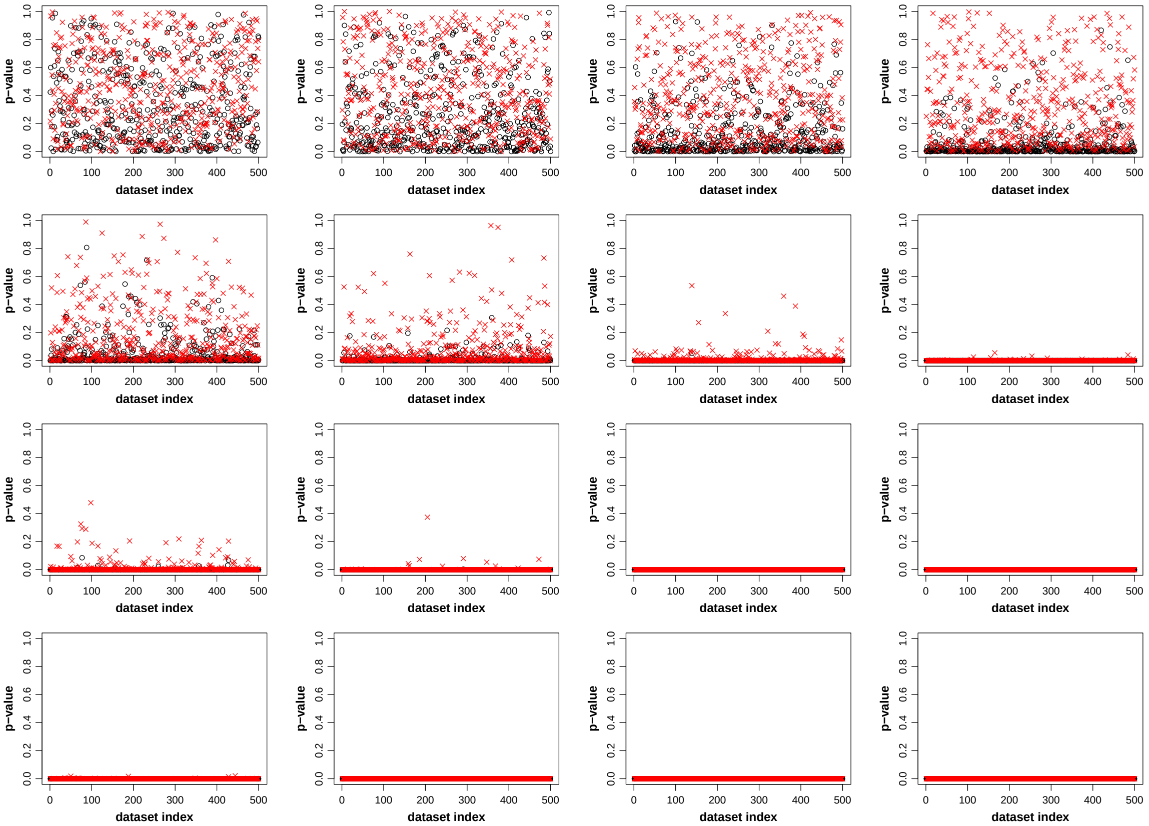

Figure 1: p-value from the group label test based on observing infection times, p-value$_{\boldsymbol{i}}$ (black circles), and based on observing removal times, p-value$_{\boldsymbol{r}}$ (red crosses), against dataset index, from the simulation study. Rows (top to bottom) correspond to scenarios 1-4 (μ =0.61, 1.32, 2.44, 3.88 and $\bar{p_L}$ = 0.09, 0.27, 0.51, 0.70, respectively). Columns (left to right) correspond to rounds (N = 99, 199, 499, 999, respectively).

Appendix S4

Comparison Against Other Tests

We consider the synthetic data that were generated from the Exp-2L model under the four simulation scenarios presented in Section 4 of the main manuscript and also presented in Table 1 below. For each of the 500 datasets that has been generated under the four scenarios below and for a given significance level α, we implement our test using infection times and estimate its power by the proportion of datasets in which the null hypothesis (of homogeneity of disease transmission) is rejected out of 500.

Table 1: Simulation conditions for the simulation study. Each simulation scenario consists of 4 rounds, where the number of initial susceptibles N is set at 99, 199, 499 and 999, respectively. For each round 500 datasets are generated. The number of individuals in each group is set at $C_H = 5$, in all instances.

	Data generating process	**Parameter values**	$\bar{p_L}$
Scenario 1	Exp-2L	$R_* = 2.5, \gamma = 0.1, \mu = 0.61$	0.09
Scenario 2	Exp-2L	$R_* = 2.5, \gamma = 0.1, \mu = 1.32$	0.27
Scenario 3	Exp-2L	$R_* = 2.5, \gamma = 0.1, \mu = 2.44$	0.51
Scenario 4	Exp-2L	$R_* = 2.5, \gamma = 0.1, \mu = 3.88$	0.70

We considered three significance levels $\alpha = 0.01, 0.05$ and 0.10 and the estimated power is given in Tables 2, 3, 4 respectively. The results are in line with those presented in Table 1 in the main manuscript, in the sense that even in the presence of a very mild within-group-transmission effect (scenario 1) the test has adequate power if the population size N is large. For mild within-group-transmission (scenario 2) the power of the test is somewhat evident even for small N and significance level α.

Although there are tests in the literature for assessing the transmission homogeneity assumption, they are fundamentally different to ours in many aspects as discussed in Section 7.2 of the main manuscript. To the best of our knowledge there is no other non-asymptotic test in the literature which can be applied when only removal times are observed. Hence any fair and like-for-like comparison between existing tests and ours is very difficult, if not impossible.

Nevertheless, we consider the case where complete temporal information is available, so that infection and recovery times are known for each individual, and we implement the test of Britton (1997). We estimate its power and compare it against that of our test. We note that Britton's test makes use of two sources of information in the data (both infection and recovery times) whereas our is implemented using infection times only. The results are shown in Tables 2, 3, 4 respectively, and reveal that under scenario 1 (i.e. very mild within-household transmission) our test outperform Britton's test especially for relatively small population size across all three significance levels. This is not surprising given that the test of Britton (1997) is based on asymptotic results and hence more accurate for large populations under which (scenarios 2, 3, and 4) both tests perform equally well and have similarly high power.

Table 2: Power of our proposed test at significance level $\alpha = 0.01$ for datasets generated under different scenarios. In brackets, the corresponding power at the same significance level for Britton's test.

	$N = 99$	$N = 199$	$N = 499$	$N = 999$
Scenario 1	0.056 (0.000)	0.094 (0.008)	0.232 (0.066)	0.492 (0.306)
Scenario 2	0.534 (0.350)	0.846 (0.862)	0.998 (0.996)	1.000 (1.000)
Scenario 3	0.986 (0.998)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
Scenario 4	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)

Table 3: Power of our proposed test at significance level $\alpha = 0.05$ for datasets generated under different scenarios. In brackets, the corresponding power at the same significance level for Britton's test.

	$N = 99$	$N = 199$	$N = 499$	$N = 999$
Scenario 1	0.172 (0.040)	0.262 (0.142)	0.494 (0.410)	0.742 (0.744)
Scenario 2	0.746 (0.814)	0.940 (0.992)	0.998 (1.000)	1.000 (1.000)
Scenario 3	0.996 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
Scenario 4	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)

Table 4: Power of our proposed test at significance level $\alpha = 0.10$ for datasets generated under different scenarios. In brackets, the corresponding power at the same significance level for Britton's test.

	$N = 99$	$N = 199$	$N = 499$	$N = 999$
Scenario 1	0.274 (0.138)	0.372 (0.288)	0.600 (0.684)	0.830 (0.908)
Scenario 2	0.832 (0.938)	0.968 (0.998)	1.000 (1.000)	1.000 (1.000)
Scenario 3	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
Scenario 4	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)

Appendix S5

Introduction

This appendix contains the proofs of Lemmas 6.1 and 6.2 and Theorem 6.3, plus some numerical illustrations. Note that all equations references are to equations found in this appendix. We refer to Barbour and Eagleson (1986) as simply 'Barbour and Eagleson' throughout.

Proof of Lemma 6.1

Proof. For $\varepsilon > 0$,

$$\begin{aligned} P\left(|2T_n/n(n-1) - 1| > \varepsilon\right) &= P(|T_n - n(n-1)/2| > \varepsilon n(n-1)/2) \\ &= 1 - P(|T_n - n(n-1)/2| \leq \varepsilon n(n-1)/2) \\ &\leq 1 - P(T_n = n(n-1)/2) = 1 - P(\mathcal{E}_n) \to 0 \end{aligned}$$

as $n \to \infty$ and the result follows. □

Proof of Lemma 6.2

Proof. First note that

$$P(\mathcal{E}_n) = \begin{cases} \frac{\binom{l}{n} m^n}{\binom{ml}{n}} & \text{if } n \leq l, \\ 0 & \text{if } n > l, \end{cases} \tag{2}$$

since for $n \leq l$ (i) under the null hypothesis there are $\binom{C}{n} = \binom{ml}{n}$ equally likely ways to select the n individuals in the outbreak; (ii) the number of ways of choosing n different groups is $\binom{l}{n}$; (iii) in each group selected there are m ways of choosing a single individual. Conversely if $n > l$ then it is impossible for all n individuals to belong to different groups.

Recall that $C = ml$ and $C \sim \theta n^\beta$, so $l \sim (\theta/m)n^\beta = \alpha n^\beta$, say, where $\alpha > 0$ since $\theta > 0$.

Suppose first that $\beta = 1$, so that $l \sim \alpha n$. Since $P(\mathcal{E}_n) = 0$ for $n > l$ we need now only consider the subsequence of $P(\mathcal{E}_n)$ values for which $n \leq l$.

Let $0 < \gamma < \min(\alpha, 1)$, let $\lfloor x \rfloor$ denote the integer part of x, and let n be large enough that $\lfloor \gamma n \rfloor \geq 1$. Since $n \leq l$ it follows from (2) that

$$P(\mathcal{E}_n) = \prod_{j=1}^{n-1} \left(\frac{l-j}{l-(j/m)} \right) \leq \prod_{j=\lfloor \gamma n \rfloor}^{n-1} \left(\frac{l-j}{l-(j/m)} \right) \leq \left(\frac{l - \lfloor \gamma n \rfloor}{l - (\lfloor \gamma n \rfloor)/m} \right)^{n - \lfloor \gamma n \rfloor}$$

since for $0 < x < l$, $f(x) = (l-x)/(l-(x/m))$ is decreasing in x for $m > 1$. Now

$$\frac{l - \lfloor \gamma n \rfloor}{l - (\lfloor \gamma n \rfloor)/m} = \frac{(l/n) - (\lfloor \gamma n \rfloor)/n}{(l/n) - (\lfloor \gamma n \rfloor)/mn} \to \frac{\alpha - \gamma}{\alpha - (\gamma/m)}$$

as $n \to \infty$, and since $n - \lfloor \gamma n \rfloor \to \infty$ it follows that $P(\mathcal{E}_n) \to 0$ along the subsequence for which $n \leq l$.

Suppose now that $\beta > 1$. Since $l \sim \alpha n^\beta$ then $n \leq l$ for all sufficiently large n, which we assume henceforth to be the case. We proceed by applying Stirling's approximation (namely that

$k! \sim \sqrt{2\pi k}\,(k/\mathrm{e})^k$) to (2).

$$\begin{aligned} P(\mathcal{E}_n) = \frac{\binom{l}{n} m^n}{\binom{ml}{n}} &= m^n \frac{l!(ml-n)!}{(ml)!(l-n)!} \\ &\sim m^n \frac{\sqrt{2\pi l}\,(l/\mathrm{e})^l \sqrt{2\pi(ml-n)}\,((ml-n)/\mathrm{e})^{ml-n}}{\sqrt{2\pi ml}\,(ml/\mathrm{e})^{ml}\sqrt{2\pi(l-n)}\,((l-n)/\mathrm{e})^{l-n}} \\ &= \sqrt{\frac{ml-n}{m(l-n)}}\,\frac{(ml-n)^{ml-n}}{(ml)^{(m-1)l}(ml-mn)^{l-n}} \\ &= \sqrt{\frac{ml-n}{m(l-n)}}\,\frac{(1-(n/ml))^{ml-n}}{(1-(n/l))^{l-n}}. \end{aligned} \tag{3}$$

We now consider the two terms in (3) as follows. Recall that that $l \sim \alpha n^\beta$. Since $\beta > 1$ then for the first term in (3) we have

$$\lim_{n\to\infty} \sqrt{\frac{ml-n}{m(l-n)}} = 1.$$

For the second term, first observe that for sufficiently large n we have $0 < (n/ml) < (n/l) < 1$. Recall that, for $|x| < 1$, $\log(1-x) = -\sum_{k=1}^{\infty}(x^k)/k$. Then for all sufficiently large n, the logarithm of the second term in (3) is

$$\begin{aligned} \log\left(\frac{(1-(n/ml))^{ml-n}}{(1-(n/l))^{l-n}}\right) &= -(ml-n)\sum_{k=1}^{\infty}\left(\frac{n}{ml}\right)^k \frac{1}{k} + (l-n)\sum_{k=1}^{\infty}\left(\frac{n}{l}\right)^k\frac{1}{k} \\ &= -\sum_{k=1}^{\infty}\frac{n^k}{(ml)^{k-1}}\frac{1}{k} + \sum_{k=1}^{\infty}\frac{n^{k+1}}{(ml)^k}\frac{1}{k} \\ &\quad + \sum_{k=1}^{\infty}\frac{n^k}{l^{k-1}}\frac{1}{k} - \sum_{k=1}^{\infty}\frac{n^{k+1}}{l^k}\frac{1}{k} \\ &= -n - \sum_{k=1}^{\infty}\frac{n^{k+1}}{(ml)^k}\frac{1}{k+1} + \sum_{k=1}^{\infty}\frac{n^{k+1}}{(ml)^k}\frac{1}{k} \\ &\quad + n + \sum_{k=1}^{\infty}\frac{n^{k+1}}{l^k}\frac{1}{k+1} - \sum_{k=1}^{\infty}\frac{n^{k+1}}{l^k}\frac{1}{k} \\ &= -\sum_{k=1}^{\infty}\frac{n^{k+1}}{l^k}\left(\frac{m^k-1}{m^k}\right)\frac{1}{k(k+1)}. \end{aligned} \tag{4}$$

Suppose that $1 < \beta < 2$. Then

$$-\sum_{k=1}^{\infty}\frac{n^{k+1}}{l^k}\left(\frac{m^k-1}{m^k}\right)\frac{1}{k(k+1)} < -\frac{n^2}{l}\left(\frac{m-1}{m}\right)\frac{1}{2} \to -\infty$$

as $n \to \infty$ since $n^2/l \to \infty$, whence $P(\mathcal{E}_n) \to 0$.

If $\beta = 2$ then as $n \to \infty$ the first term ($k = 1$) in the sum in (4) tends to $-(m-1)/(2m\alpha)$ while the remaining terms ($k \geq 2$) all tend to zero. We now show that the sum of the remaining terms also tends to zero. Recall that $n \leq l$, and observe that since $\beta = 2$, $n^3/l^2 < 1$ for all sufficiently large n; for such n and $k \geq 2$,

$$\frac{n^{k+1}}{l^k}\left(\frac{m^k-1}{m^k}\right)\frac{1}{k(k+1)} \leq \frac{n^3}{l^2}\left(\frac{n}{l}\right)^{k-2}\frac{1}{k(k+1)} \leq \frac{1}{k^2}.$$

Recall Tannery's Theorem (a special case of the Lebesque's dominated convergence theorem), which states that if $S_k = \sum_{k=1}^{\infty} a_k(n)$, $\lim_{n\to\infty} a_k(n) = b_k$, $|a_k(n)| \leq c_k$ and $\sum_{k=1}^{\infty} c_k < \infty$ then $\lim_{n\to\infty} S_n = \sum_{k=1}^{\infty} b_k$. Applying this with $a_k(n)$ equal to the kth summand in (4) for $k \geq 2$ and $c_k = 1/k^2$ yields that $\sum_{k=2}^{\infty} a_k(n) \to 0$ and hence $P(\mathcal{E}_n) \to \exp(-(m-1)/(2m\alpha))$ as $n \to \infty$.

Finally, let $\beta > 2$. In this case all the terms in the sum in (4) converge to zero, and an identical argument to the $1 < \beta < 2$ case shows that the sum itself converges to zero, so that $P(\mathcal{E}_n) \to 1$ as $n \to \infty$.

□

Although we have assumed that the population consists of groups of equal size, we conjecture that corresponding results apply if this assumption is relaxed. Such results appear harder to prove. For instance, deriving the corresponding equation for (2) requires summation over all possible ways to select individuals in each possible group size, resulting in an unwieldy expression. It might be possible to proceed using an approach in which $P(\mathcal{E}_n)$ is bounded by comparison with populations of equal-sized groups, but it is not immediately obvious how to do this.

Proof of Theorem 6.3

Proof. We adopt the following notation from Barbour and Eagleson. First, let $\sum_{i,j}$ ($\sum_{i,j,k}$, $\sum_{i,j,k,m}$) denote the sum over all ordered pairs (i, j) (triples (i, j, k), quadruples (i, j, k, m)) of distinct integers from $\{1, \ldots, C\}$. For $k = 2, 3, \ldots$ let $(C)_k$ denote $C(C-1)\ldots(C-k+1)$. Define

$$\begin{aligned} D &= \frac{1}{(C)_2} \sum_{i,j} d_{ij}, \\ d_i^* &= \frac{1}{C-2} \sum_{\substack{j=1 \\ j \neq i}}^{C} (d_{ij} - D), \\ D_1 &= \frac{1}{C} \sum_{i=1}^{C} (d_i^*)^2, \\ D_2 &= \frac{1}{(C)_2} \sum_{i,j} (d_{ij} - d_i^* - d_j^* - D)^2, \end{aligned}$$

and define Y, y_i^*, Y_1 and Y_2 similarly using y_{ij} in place of d_{ij}. Next, define

$$\begin{aligned} d_1 &= \frac{1}{(C)_2} \sum_{i,j} d_{ij}, \quad d_2 = \frac{1}{(C)_3} \sum_{i,j,k} d_{ij} d_{ik}, \\ d_3 &= \frac{1}{(C)_4} \sum_{i,j,k,m} d_{ij} d_{ik} d_{im}, \quad d_4 = \frac{1}{(C)_4} \sum_{i,j,k,m} d_{ij} d_{ik} d_{jm}, \\ d_5 &= \frac{1}{(C)_3} \sum_{i,j,k} d_{ij}^2 d_{ik}, \quad d_6 = \frac{1}{(C)_2} \sum_{i,j} d_{ij}^3, \\ d_7 &= \frac{1}{(C)_3} \sum_{i,j,k} d_{ij} d_{ik} d_{jk}, \quad d_8 = \frac{1}{(C)_2} \sum_{i,j} d_{ij}^2, \end{aligned}$$

and define $y_1, \ldots, y_8$ similarly using y_{ij}. (Aside: in Barbour and Eagleson all of these quantities are defined in terms of absolute values, e.g. $|d_{ij} d_{ik}|$ for d_2, but our definitions are identical for our setting since $d_{ij}, y_{ij} \geq 0$. We also have $D = d_1$ and $Y = y_1$ but for ease of comparison with Barbour and Eagleson we adopt the same notation.) Then from Barbour and Eagleson,

$$\begin{aligned} E[T_n] &= C(C-1)DY, && (5) \\ s_n^2 &= \operatorname{var}(T_n) = \frac{4C^2(C-2)^2}{C-1} D_1 Y_1 + \frac{2C(C-1)^2}{C-3} D_2 Y_2. && (6) \end{aligned}$$

Define

$$\begin{aligned} \varepsilon_n &= s_n^{-3} \left\{ C^4(d_1^3 + d_1 d_2 + d_3 + d_4)(y_1^3 + y_1 y_2 + y_3 + y_4) \right. \\ &\quad \left. + C^3(d_5 + d_1 d_8)(y_5 + y_1 y_8) + C^2 d_6 y_6 \right\}. && (7) \end{aligned}$$

Corollary 2.1 of Barbour and Eagleson states that if $\epsilon_n \to 0$ as $n \to \infty$ then T_n is asymptotically Gaussian in the sense given in Theorem 6.3. We therefore proceed by obtaining suitable upper bounds for the d_i and y_i terms and a lower bound for the variance s_n^2, starting with terms involving d_{ij}.

Observe that the $C \times C$ matrix of d_{ij} values, M_d say, takes the form

$$M_d = (d_{ij}) = \begin{pmatrix} 0 & 1/2 & 0 & 0 & \dots & 0 & 0 \\ 1/2 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 1/2 & \dots & 0 & 0 \\ 0 & 0 & 1/2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 1/2 \\ 0 & 0 & 0 & 0 & \dots & 1/2 & 0 \end{pmatrix}. \qquad (8)$$

It follows that $\sum_{i,j} d_{ij} = C/2$ since the sum contains precisely C non-zero entries, all of which equal 1/2. Then

$$D = d_1 = \frac{1}{C(C-1)} \sum_{i,j} d_{ij} = \frac{1}{2(C-1)}.$$

Next,

$$d_i^* = \frac{1}{C-2} \sum_{\substack{j=1 \\ j \neq i}}^{C} (d_{ij} - D) = \frac{1}{C-2}(1/2) - \frac{1}{C-2}(C-1)\frac{1}{2(C-1)} = 0.$$

Thus

$$D_1 = \frac{1}{C} \sum_{i=1}^{C} (d_i^*)^2 = 0$$

and

$$\begin{aligned} D_2 &= \frac{1}{C(C-1)} \sum_{i,j} (d_{ij} - d_i^* - d_j^* - D)^2 \\ &= \frac{1}{C(C-1)} \sum_{i,j} \left(d_{ij} - \frac{1}{2(C-1)} \right)^2 \\ &= \frac{1}{C(C-1)} \sum_{i,j} \left(d_{ij}^2 - \frac{d_{ij}}{C-1} + \frac{1}{4(C-1)^2} \right) \\ &= \frac{1}{C(C-1)} \left(\frac{C}{4} - \frac{C}{2(C-1)} + \frac{(C)(C-1)}{4(C-1)^2} \right) \\ &= \frac{C-2}{4(C-1)^2}. \end{aligned}$$

Next, from (8) each row and column of M_d contains only one non-zero entry and it follows immediately that

$$d_2 = d_3 = d_4 = d_5 = d_7 = 0. \tag{9}$$

Finally we have

$$d_6 = \frac{1}{C(C-1)} \sum_{i,j} d_{ij}^3 = \frac{1}{8(C-1)}$$

and

$$d_8 = \frac{1}{C(C-1)} \sum_{i,j} d_{ij}^2 = \frac{1}{4(C-1)}.$$

We now consider the terms in (7) that involve y_{ij} values. It is helpful to tabulate these as follows.

y_{ij}	1	2	3	4	5	$\ldots$	$n-1$	n	$n+1$	$\ldots$	C
1	0	0	1	2	3	$\ldots$	$n-3$	$n-2$	$n-1$	$\ldots$	$n-1$
2	0	0	0	1	2	$\ldots$	$n-4$	$n-3$	$n-2$	$\ldots$	$n-2$
3	1	0	0	0	1	$\ldots$	$n-5$	$n-4$	$n-3$	$\ldots$	$n-3$
4	2	1	0	0	0	$\ldots$	$n-6$	$n-5$	$n-4$	$\ldots$	$n-4$
5	3	2	1	0	0	$\ldots$	$n-7$	$n-6$	$n-5$	$\ldots$	$n-5$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-1$	$n-3$	$n-4$	$n-5$	$n-6$	$n-7$	$\ldots$	0	0	1	$\ldots$	1
n	$n-2$	$n-3$	$n-4$	$n-5$	$n-6$	$\ldots$	0	0	0	$\ldots$	0
$n+1$	$n-1$	$n-2$	$n-3$	$n-4$	$n-5$	$\ldots$	1	0	0	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C	$n-1$	$n-2$	$n-3$	$n-4$	$n-5$	$\ldots$	1	0	0	$\ldots$	0

Writing M_y for the matrix of (y_{ij}) values we see that M_y has the form

$$M_y = \begin{pmatrix} U & V \\ V^T & Z \end{pmatrix}$$

where
(i) U is an $n \times n$ symmetric matrix whose entries $u_{ij} = |i-j| - 1$ $(i \neq j)$ are the contribution to T_n of two individuals from the same group who appear at locations i and j in the sample of n individuals;
(ii) V is an $n \times (C-n)$ matrix with identical columns and with rows whose entries $v_{ij} = n - i$ are the contributions to T_n of two individuals from the same group, exactly one of whom appears in the sample, at location i; and
(iii) Z is a $(C-n) \times (C-n)$ matrix of zeroes, these being the contribution to T_n from two individuals from the same group who do not appear in the sample.

Observe that $y_{ij} < n$ for all $1 \leq i, j \leq C$. Then

$$y_2 = \frac{1}{C(C-1)(C-2)} \sum_{i,j,k} y_{ij} y_{ik} < \frac{1}{C(C-1)(C-2)} \sum_{i,j,k} n^2 = n^2,$$

and identical reasoning yields that

$$y_3 < n^3,\ y_4 < n^3,\ y_5 < n^3,\ y_6 < n^3,\ y_7 < n^3,\ y_8 < n^2. \tag{10}$$

We thus have expressions or bounds for all terms in (7) apart from the variance s_n^2. However, since $D_1 = 0$ it follows from (6) that we can ignore Y_1 and so it only remains to consider Y_2. To do so requires a number of preliminary calculations. First, note that

$$\sum_{i,j} y_{ij} = \left(\sum_{i,j} u_{ij} + 2 \sum_{i,j} v_{ij} \right),$$

where on the right-hand side the summations are over all pairs (i,j) of distinct indices in the U and V matrices. Since U is symmetric and $u_{ii} = 0$, $\sum_{i,j} u_{ij} = 2 \sum_{i<j} u_{ij}$. Furthermore, for $k = 1, \ldots, (n-1)$ each k-diagonal of U (i.e. the diagonal offset by k places above the leading diagonal) consists of $(n-k)$ entries all equal to $k-1$. Thus

$$\sum_{i<j} u_{ij} = \sum_{k=2}^{n-1} (k-1)(n-k) = (n-1) \sum_{k=1}^{n-2} k - \sum_{k=1}^{n-2} k^2,$$

and after a few lines of algebra we obtain

$$\sum_{i,j} u_{ij} = 2 \sum_{i<j} u_{ij} = \frac{n(n-1)(n-2)}{3}.$$

Since V has $C-n$ identical columns consisting of the entries $1, 2, \ldots, n-1$,

$$\sum_{i,j} v_{ij} = (C-n) \sum_{k=1}^{n-1} k = \frac{(C-n)n(n-1)}{2}$$

and thus

$$\begin{aligned} \sum_{i,j} y_{ij} &= \frac{n(n-1)(n-2)}{3} + (C-n)n(n-1) \\ &= \frac{n(n-1)(3C-2n-2)}{3}. \end{aligned} \tag{11}$$

In a similar fashion we have

$$\sum_{i,j} y_{ij}^2 = \left(\sum_{i,j} u_{ij}^2 + 2 \sum_{i,j} v_{ij}^2 \right),$$

where

$$\begin{aligned} \sum_{i<j} u_{ij}^2 = \sum_{k=2}^{n-1} (k-1)^2 (n-k) &= (n-1) \sum_{k=1}^{n-2} k^2 - \sum_{k=1}^{n-2} k^3, \\ &= \frac{n(n-1)^2(n-2)}{12} \end{aligned}$$

and

$$\sum_{i,j} v_{ij}^2 = (C-n) \sum_{k=1}^{n-1} k^2 = \frac{(C-n)n(n-1)(2n-1)}{6}$$

yielding

$$\begin{aligned}\sum_{i,j} y_{ij}^2 &= \frac{n(n-1)^2(n-2)}{6} + \frac{(C-n)n(n-1)(2n-1)}{3} \\ &= \frac{n(n-1)[2C(2n-1)-n(3n+1)+2]}{6}. \qquad (12)\end{aligned}$$

It follows from (11) that

$$\begin{aligned}Y = y_1 &= \frac{1}{C(C-1)} \sum_{i,j} y_{ij} \\ &= \frac{n(n-1)}{C(C-1)} \frac{(3C-2n-2)}{3},\end{aligned}$$

and from (5) we obtain

$$E[T_n] = \frac{n(n-1)(3C-2n-2)}{6(C-1)}.$$

We next evaluate the row sums of M_y. For $i = 1, \ldots, C$ we define

$$\begin{aligned}\psi_i = \sum_{\substack{j=1 \\ j \neq i}}^{C} y_{ij} &= \sum_{\substack{j=1 \\ j \neq i}}^{n} y_{ij} + \sum_{\substack{j=n+1 \\ j \neq i}}^{C} y_{ij} \\ &= \phi_i + \chi_i,\end{aligned}$$

say. The structure of M_y implies that

$$\chi_i = \begin{cases} (C-n)(n-i) & i = 1, \ldots, n, \\ 0 & i = n+1, \ldots, C, \end{cases}$$

and that $\phi_i = n(n-1)/2$ for $i = n+1, \ldots, C$. Inspection of the structure of U yields that, for $2 \leq i \leq n-1$,

$$\begin{aligned}\phi_i &= [1+2+\ldots+(n-i-1)] + [1+2+\ldots+(i-2)] \\ &= (n-i-1)(n-i)/2 + (i-2)(i-1)/2 \\ &= i^2 - (n+1)i + n(n-1)/2 + 1,\end{aligned}$$

and moreover this final expression also holds for $i = 1$ and $i = n$. Assembling these results gives

$$\psi_i = \begin{cases} i^2 - (n+1)i + n(n-1)/2 + 1 + (C-n)(n-i) & i = 1, \ldots, n, \\ n(n-1)/2 & i = n+1, \ldots, C. \end{cases}$$

Note that

$$\sum_{i=1}^{n} \phi_i = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} y_{ij} = \sum_{i,j} u_{ij} = \frac{n(n-1)(n-2)}{3} \qquad (13)$$

and, using (11),

$$\sum_{i=1}^{C} \psi_i = \sum_{i=1}^{C} \sum_{\substack{j=1 \\ j \neq i}}^{C} y_{ij} = \sum_{i,j} y_{ij} = \frac{n(n-1)(3C-2n-2)}{3}. \qquad (14)$$

Next,

$$\begin{aligned} y_i^* = \frac{1}{C-2}\sum_{\substack{j=1\\ j\neq i}}^{C}(y_{ij}-Y) &= \frac{1}{C-2}\psi_i - \frac{(C-1)}{(C-2)}\frac{n(n-1)}{C(C-1)}\frac{(3C-2n-2)}{3} \\ &= \frac{\psi_i}{C-2} - \frac{n(n-1)}{C(C-2)}\frac{(3C-2n-2)}{3} \end{aligned}$$

and thus

$$\begin{aligned} y_i^* + y_j^* + Y &= \frac{\psi_i+\psi_j}{C-2} - \frac{2n(n-1)}{C(C-2)}\frac{(3C-2n-2)}{3} + \frac{n(n-1)}{C(C-1)}\frac{(3C-2n-2)}{3} \\ &= \frac{\psi_i+\psi_j}{C-2} - \frac{n(n-1)(3C-2n-2)}{3(C-1)(C-2)} \\ &= A_{ij} - B, \end{aligned}$$

say. Then

$$\begin{aligned} \sum_{i,j}(y_{ij}-y_i^*-y_j^*-Y)^2 &= \sum_{i,j}(y_{ij}-(A_{ij}-B))^2 \\ &= \sum_{i,j} y_{ij}^2 + 2B\sum_{i,j} y_{ij} - 2\sum_{i,j} y_{ij}A_{ij} \\ &\quad + \sum_{i,j} A_{ij}^2 - 2B\sum_{i,j} A_{i,j} + \sum_{i,j} B^2. \end{aligned} \tag{15}$$

We now consider the three terms in (15) involving A_{ij}; the remaining terms can be evaluated using our existing results. First,

$$\begin{aligned} \sum_{i,j} y_{ij}A_{ij} &= \frac{1}{C-2}\left(\sum_{i,j} y_{ij}\psi_i + \sum_{i,j} y_{ij}\psi_j\right) \\ &= \frac{1}{C-2}\left(\sum_{i=1}^{C}\psi_i\sum_{\substack{j=1\\ j\neq i}}^{C} y_{ij} + \sum_{j=1}^{n}\psi_j\sum_{\substack{i=1\\ i\neq j}}^{C} y_{ij}\right) \\ &= \frac{1}{C-2}\left(\sum_{i=1}^{C}\psi_i^2 + \sum_{j=1}^{n}\psi_j^2\right) \\ &= \frac{2}{C-2}\sum_{i=1}^{C}\psi_i^2. \end{aligned} \tag{16}$$

Next,

$$\begin{aligned} \sum_{i,j} A_{ij}^2 &= \sum_{i,j}\left(\frac{\psi_i+\psi_j}{C-2}\right)^2 \\ &= \frac{1}{(C-2)^2}\left(\sum_{i,j}\psi_i^2 + \sum_{i,j}\psi_j^2 + 2\sum_{i,j}\psi_i\psi_j\right). \end{aligned}$$

However,

$$\sum_{i,j} \psi_i^2 = \sum_{i=1}^{C} \psi_i^2 \sum_{\substack{j=1 \\ j\neq i}}^{C} 1 = (C-1) \sum_{i=1}^{C} \psi_i^2,$$

and so

$$\begin{aligned}
\sum_{i,j} A_{ij}^2 &= \frac{1}{(C-2)^2} \left\{ 2(C-1) \sum_{i=1}^{C} \psi_i^2 + 2 \left(\sum_{i=1}^{C} \sum_{j=1}^{C} \psi_i \psi_j - \sum_{i=1}^{C} \psi_i^2 \right) \right\} \\
&= \frac{1}{(C-2)^2} \left\{ 2(C-1) \sum_{i=1}^{C} \psi_i^2 + 2 \left(\sum_{i=1}^{C} \psi_i \right) \left(\sum_{j=1}^{C} \psi_j \right) - 2 \sum_{i=1}^{C} \psi_i^2 \right\} \\
&= \frac{2C-4}{(C-2)^2} \sum_{i=1}^{C} \psi_i^2 + \frac{2}{(C-2)^2} \left(\sum_{i=1}^{C} \psi_i \right)^2 . \\
&= \frac{2C-4}{(C-2)^2} \sum_{i=1}^{C} \psi_i^2 + \frac{2n^2(n-1)^2(3C-2n-2)^2}{9(C-2)^2}, \qquad (17)
\end{aligned}$$

using (14). Next,

$$\begin{aligned}
\sum_{i,j} A_{ij} &= \frac{2}{C-2} \sum_{i=1}^{C} \sum_{\substack{j=1 \\ j\neq i}}^{C} \psi_i \\
&= \frac{2(C-1)}{C-2} \sum_{i=1}^{C} \psi_i \\
&= \frac{2n(n-1)(C-1)(3C-2n-2)}{3(C-2)} \qquad (18)
\end{aligned}$$

using (14).

To fully evaluate (16) and (17) we need to find $\sum \psi_i^2$. We have

$$\begin{aligned}
\sum_{i=1}^{C} \psi_i^2 &= \sum_{i=1}^{n} (\phi_i + (C-n)(n-i))^2 + \sum_{i=n+1}^{C} (n(n-1)/2)^2 \\
&= \sum_{i=1}^{n} \phi_i^2 + 2(C-n)n \sum_{i=1}^{n} \phi_i - 2(C-n) \sum_{i=1}^{n} i\phi_i \\
&\quad + (C-n)^2 \sum_{i=1}^{n} (n-i)^2 + \frac{n^2(n-1)^2(C-n)}{4} \\
&= \sum_{i=1}^{n} \phi_i^2 + \frac{2(C-n)n^2(n-1)(n-2)}{3} \\
&\quad -2(C-n) \sum_{i=1}^{n} \left\{ i^3 - (n+1)i^2 + (n(n-1)/2+1)i \right\} \\
&\quad + (C-n)^2 \sum_{i=0}^{n-1} i^2 + \frac{n^2(n-1)^2(C-n)}{4} \\
&= \sum_{i=1}^{n} \phi_i^2 + \frac{2(C-n)n^2(n-1)(n-2)}{3} \\
&\quad - \frac{2(C-n)n^2(n+1)^2}{4} + \frac{2(C-n)n(n+1)^2(2n+1)}{6} \\
&\quad -2(C-n)(n(n-1)/2+1)n(n+1)/2 \\
&\quad + (C-n)^2 \left(\frac{n(n-1)(2n-1)}{6} \right) \\
&\quad + \frac{n^2(n-1)^2(C-n)}{4}.
\end{aligned} \tag{19}$$

The only remaining term to evaluate in (19) is $\sum \phi_i^2$. Writing $m = n(n-1)/2+1$ we have

$$\begin{aligned}
\sum_{i=1}^{n} \phi_i^2 &= \sum_{i=1}^{n} (i^2 - (n+1)i + m)^2 \\
&= \sum_{i=1}^{n} \left\{ i^4 + (n+1)^2 i^2 + m^2 - 2(n+1)i^3 + 2mi^2 - 2m(n+1)i \right\} \\
&= \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30} + \frac{n(n+1)^3(2n+1)}{6} \\
&\quad + n(n(n-1)/2+1)^2 - \frac{2n^2(n+1)^3}{4} \\
&\quad + \frac{(n(n-1)+2)n(n+1)(2n+1)}{6} \\
&\quad - \frac{(n(n-1)+2)n(n+1)^2}{2}.
\end{aligned} \tag{20}$$

Since

$$Y_2 = \frac{1}{C(C-1)} \sum_{i,j} (y_{ij} - y_i^* - y_j^* - Y)^2,$$

equation (15) along with the calculations for the six right-hand side terms yields an explicit, albeit very unwieldy, expression for Y_2, and hence for the variance s_n^2 defined at (6). However, since our main objective is to show that $\epsilon_n \to 0$ as $n \to \infty$, it will be sufficient to find the order of magnitude of s_n^2 and combine this with the expressions or bounds for the other terms in (7). To begin with, observe from (13) that the order n^5 terms are dominant, and collecting them together yields

$$\sum_{i=1}^{n} \phi_i^2 = \left(\frac{6}{30} + \frac{2}{6} + \frac{1}{4} - \frac{2}{4} + \frac{2}{6} - \frac{1}{2}\right) n^5 + O(n^4) \sim \frac{7}{60} n^5.$$

Recalling that $C \sim \theta n$, a similar calculation applied to (19) yields

$$\begin{aligned}
\sum_{i=1}^{C} \psi_i^2 &= \sum_{i=1}^{n} \phi_i^2 + \left(\frac{2(\theta-1)}{3} - \frac{(\theta-1)}{2} + \frac{2(\theta-1)}{3} - \frac{(\theta-1)}{2} + \frac{(\theta-1)^2}{3} + \frac{(\theta-1)}{4}\right) n^5 \\
&\quad + O(n^4) \\
&= \left(\frac{7}{60} + \frac{7(\theta-1)}{12} + \frac{(\theta-1)^2}{3}\right) n^5 + O(n^4) \\
&\sim \left(\frac{7}{60} + \frac{7(\theta-1)}{12} + \frac{(\theta-1)^2}{3}\right) n^5. \qquad (21)
\end{aligned}$$

We now derive the orders of magnitude of the six terms on the right-hand side of (15). First, from (12) we have

$$\sum_{i,j} y_{ij}^2 \sim \frac{(4\theta-3)}{6} n^4.$$

Next, from (11),

$$2B \sum_{i,j} y_{ij} \sim \frac{2}{9} \left(\frac{3\theta-2}{\theta}\right)^2 n^4.$$

From (16), (17) and (21) we obtain

$$\begin{aligned}
-2\sum_{i,j} y_{ij} A_{ij} + \sum_{i,j} A_{ij}^2 &= \frac{4-2C}{(C-2)^2} \sum_{i=1}^{C} \psi_i^2 + \frac{2n^2(n-1)^2(3C-2n-2)^2}{9(C-2)^2} \\
&\sim -\frac{2n\theta}{(n\theta)^2} \left(\frac{7}{60} + \frac{7(\theta-1)}{12} + \frac{(\theta-1)^2}{3}\right) n^5 \\
&\quad + \frac{2}{9} \left(\frac{3\theta-2}{\theta}\right)^2 n^4 \\
&= \left(-\frac{2\theta}{3} + \frac{13}{6} - \frac{12}{5\theta} + \frac{8}{9\theta^2}\right) n^4.
\end{aligned}$$

From (18) we have

$$-2B \sum_{i,j} A_{ij} \sim -\frac{4}{9} \left(\frac{3\theta-2}{\theta}\right)^2 n^4,$$

and finally

$$\sum_{i,j} B^2 \sim \frac{1}{9} \left(\frac{3\theta-2}{\theta}\right)^2 n^4.$$

Applying these results to (15) yields

$$\begin{aligned}\sum_{i,j}(y_{ij}-y_i^*-y_j^*-Y)^2 &\sim \left(\frac{2\theta}{3}-\frac{1}{2}-\frac{2\theta}{3}+\frac{13}{6}-\frac{12}{5\theta}+\frac{8}{9\theta^2}-\frac{1}{9}\left(\frac{3\theta-2}{\theta}\right)^2\right)n^4\\ &= \left(\frac{2}{3}-\frac{16}{15\theta}+\frac{4}{9\theta^2}\right)n^4.\end{aligned}$$

It is straightforward to show that if $f(\theta) = 2/3 - (16/15\theta) + (4/9\theta^2)$ then f is strictly increasing on $[1,\infty)$ and thus for $\theta \geq 1$, $f(\theta) \geq f(1) = 2/45 > 0$. It follows that

$$Y_2 = \frac{1}{C(C-1)}\sum_{i,j}(y_{ij}-y_i^*-y_j^*-Y)^2 \sim \left(\frac{2}{3\theta^2}-\frac{16}{15\theta^3}+\frac{4}{9\theta^4}\right)n^2,$$

and thus from (6) and the facts that $D_1 = 0$ and $D_2 = (C-2)/(4(C-1)^2)$ we obtain

$$\begin{aligned}s_n^2 &= \frac{2C(C-1)^2}{C-3}D_2Y_2\\ &\sim \left(\frac{1}{3\theta}-\frac{8}{15\theta^2}+\frac{2}{9\theta^3}\right)n^3,\end{aligned}$$

whence

$$s_n^{-3} \sim \left(\frac{1}{3\theta}-\frac{8}{15\theta^2}+\frac{2}{9\theta^3}\right)^{-3/2} n^{-9/2}.$$

Noting that $d_1, d_6, d_8 = O(n^{-1})$ and recalling (9) and (10), from (7) we have

$$\begin{aligned}\varepsilon_n &= s_n^{-3}\left\{C^4(d_1^3)(y_1^3+y_1y_2+y_3+y_4)\right.\\ &\quad \left.+C^3(d_1d_8)(y_5+y_1y_8)+C^2d_6y_6\right\}.\\ &\leq s_n^{-3}\left\{C^4(d_1^3)(4n^3)+C^3(d_1d_8)(2n^3)+C^2d_6n^3\right\}\\ &= O(n^{-9/2})\left\{O(n^4n^{-3}n^3)+O(n^3n^{-2}n^3)+O(n^2n^{-1}n^3)\right\}\\ &= O(n^{-1/2})\end{aligned}$$

and thus $\varepsilon_n \to 0$ as $n \to \infty$ and the result follows. □

It is possible that a central limit theorem still holds if $C \sim \theta n^\beta$ for $1 < \beta < 2$, but the method of proof used for the $\beta = 1$ case appears to fail, even with tighter upper bounds on $y_2, \ldots, y_8$. It is however possible to follow the same arguments for the derivation of the asymptotic variance to show that, if $C \sim \theta n^\beta$ for $1 < \beta < 2$, then $s_n^2 \sim n^{4-\beta}/3\theta$.

Numerical illustrations

To illustrate Theorem 6.3 numerically for populations consisting of groups of size 2 we drew 10^5 samples from the sampling distribution of T for different choices of n and θ and then normalised them using (5) and (6). Figure 2 demonstrates that even for relatively small population sizes (e.g. 250 individuals) the Gaussian approximation to the sampling distribution of T is fairly good unless θ is large.

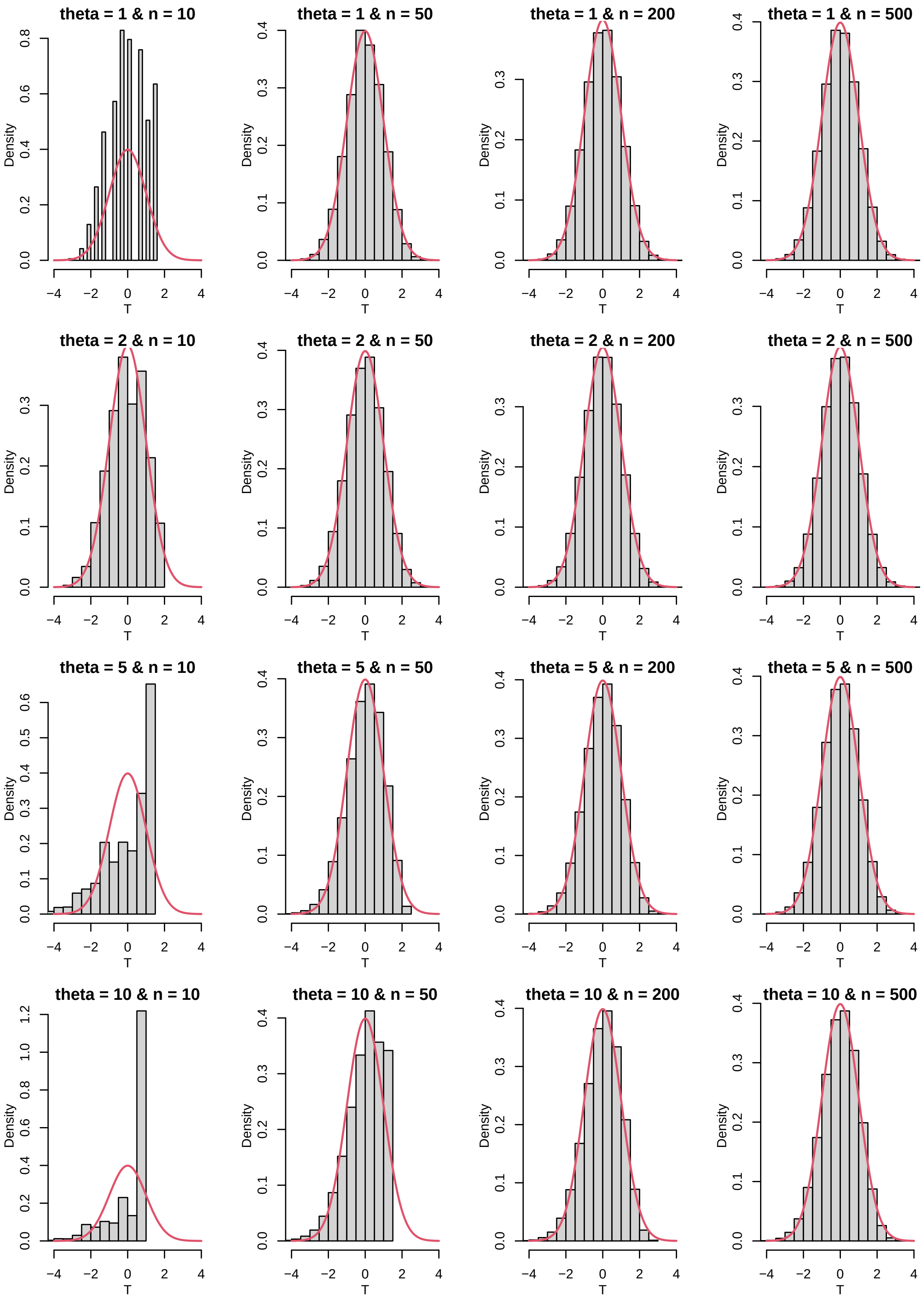

Figure 2: Histograms of 10^5 realizations from the sampling distribution under H_0 of $\frac{T-E[T]}{\sqrt{\text{var}(T)}}$ for a population of size $C = \theta n$, for different values of θ and n. The solid black lines show the probability density function of a $N(0, 1)$ distribution.

References

Barbour, A. and Eagleson, G. (1986). Random association of symmetric arrays. *Stochastic Analysis and Applications*, 4(3):239–281.

Britton, T. (1997). A test to detect within-family infectivity when the whole epidemic process is observed. *Scandinavian Journal of Statistics*, 24(3):315–330.