Heuristic Bias Correction - Conditional Poisson Distribution

April 6, 2019

Contents

1 Description

The MLE is only meaningful for non-degenerate data. However, the bias correction does not take this fact properly into account.

The probability of obtaining data with $\sum_{k=1}^{n} N_k = N$ is given by

$$q_1 := \Big(\sum_{j=1}^{n} Q_{\boldsymbol{e}_j}\Big)^N, \tag{1}$$

where

$$Q_{\boldsymbol{e}_j} = \frac{e^{\lambda p_j} - 1}{e^{\lambda} - 1}. \tag{2}$$

The probability that only one lineage is observed in the data is given by

$$\sum_{j=1}^{n} Q_{\boldsymbol{e}_j}^N.$$

In this case there are no signs of super-infections. Hence, the MLE is not uniquely defined, rather all admissible values of λ are equally likely given that $p_j = 1$ if only lineage A_j is observed. The probability of obtaining data with no sings of super-infection but at least two different lineages present, $\sum_{k=1}^{n} N_k = N$ and $N_k \neq N$ for all $k = 1, 2, \ldots, n$, is given by

$$q_2 := \Big(\sum_{j=1}^{n} Q_{\boldsymbol{e}_j}\Big)^N - \sum_{j=1}^{n} Q_{\boldsymbol{e}_j}^N. \tag{3}$$

In this case the MLE would formally be $\hat{\lambda} = 0$. The probability of obtaining data with $N_k = N$ for at least one k is given by

$$q_3 := \frac{1}{(1-e^{-\lambda})^N}\left(1 - \prod_{j=1}^{n}\left(1 - (1-e^{-\lambda p_j})^N\right)\right) \tag{4}$$

in which case no MLE exists. In summary the probability of having degenerate data becomes

$$q := q_3 + q_2 \tag{5}$$

These facts are not properly addressed by the bias correction. A regular dataset $\boldsymbol{X}$, where $\boldsymbol{X}$ is a 0-1 matrix of size $N \times n$, is a dataset where $N_k = \sum_{j=1}^{N} \boldsymbol{X}_{kj} \neq N$ for $k = 1, 2, \ldots n$ and $\sum_{k=1}^{n} N_k = \sum_{k=1}^{n}\sum_{j=1}^{N} \boldsymbol{X}_{kj} > N$. Let $\mathcal{X}$ be the set of all regular datasets. By iterated expectation we have

$$\begin{aligned}
\mathbb{E}\{\hat{\lambda}^{(\mathrm{bc})}\} &= \mathbb{E}\{\mathbb{E}[\hat{\lambda}^{(\mathrm{bc})}|\mathbb{1}_{\mathcal{X}}(\boldsymbol{X})\}\} && \text{(6a)}\\
&= \Pr\{\mathbb{1}_{\mathcal{X}}(\boldsymbol{X}) = 1\}\,\mathbb{E}\{\hat{\lambda}^{(\mathrm{bc})}|\mathbb{1}_{\mathcal{X}}(\boldsymbol{X}) = 1\} + \Pr\{\mathbb{1}_{\mathcal{X}}(\boldsymbol{X}) = 0\}\,\mathbb{E}\{\hat{\lambda}^{(\mathrm{bc})}|\mathbb{1}_{\mathcal{X}}(\boldsymbol{X}) = 0\} && \text{(6b)}\\
&\approx \Pr\{\mathbb{1}_{\mathcal{X}}(\boldsymbol{X}) = 1\}\,\mathbb{E}\{\hat{\lambda}^{(\mathrm{bc})}|\mathbb{1}_{\mathcal{X}}(\boldsymbol{X}) = 1\} && \text{(6c)}
\end{aligned}$$

where

$$\Pr\{\mathbb{1}_{\mathcal{X}}(\boldsymbol{X}) = 1\} = 1 - q \tag{7}$$

is the probability of regular data. Asymptotically equality holds in 6c. However, if N is small and λ moderate and the lineage frequencies are highly skewed, second term in 6c contributes substantially to the sum. In moderate settings, a better bias-corrected estimate would be

$$\tilde{\lambda}^{(\mathrm{bc})} = (1-q)\hat{\lambda}^{(\mathrm{bc})} \tag{8}$$

where q_3 and q_2 are evaluated at MLE. If the bias function is non-constant, preferably the analytic bias expression should be evaluated at BCMLE (MacKinnon and Smith-1998). Consequently a better estimate for q is derived at the bias-corrected estimate $\hat{\lambda}^{(bc)}$.

Remark 1. *A better bias-corrected estimate of MOI parameter is*

$$\tilde{\lambda}^{(bc)} = (1-q)\hat{\lambda}^{(bc)}, \tag{9}$$

where q is the probability of having pathological data. Similarly, we can obtain better estimations for lineage frequencies

$$\hat{p}_k^{bc} = (1-q)\hat{p}_k^{bc}. \tag{10}$$

Let $r = \frac{1-q_3}{1-q_1}$ an alternative estimate is introduced in the next remark.

Remark 2. *An alternative bias-corrected estimate is*

$$\tilde{\lambda}^{(bc)} = \hat{\lambda} - \text{bias}(\hat{\lambda})r, \tag{11}$$

where q_1, q_3 and $\text{bias}\,\hat{\lambda}$ *are evaluated at* $\hat{\boldsymbol{\theta}}^{(bc)}$.

2 Bias Function

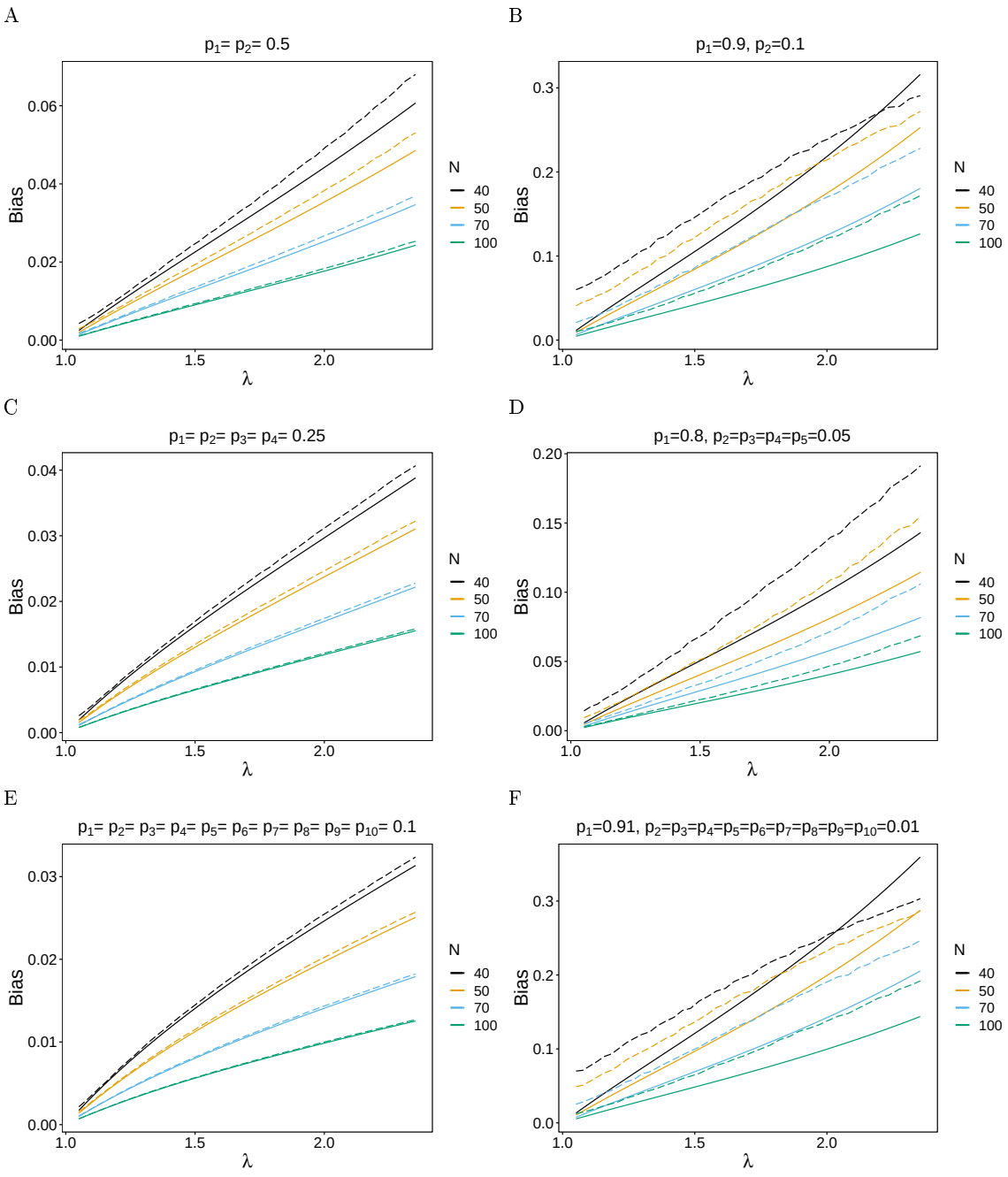


Figure 1: **Bias function**. The figure shows the behavior of average approximation of theoretical bias from $S = 10,000$ simulated datasets (solid lines) versus the theoretical bias at the true parameter (long-dashed lines). Different sample sizes are specified with different colors.

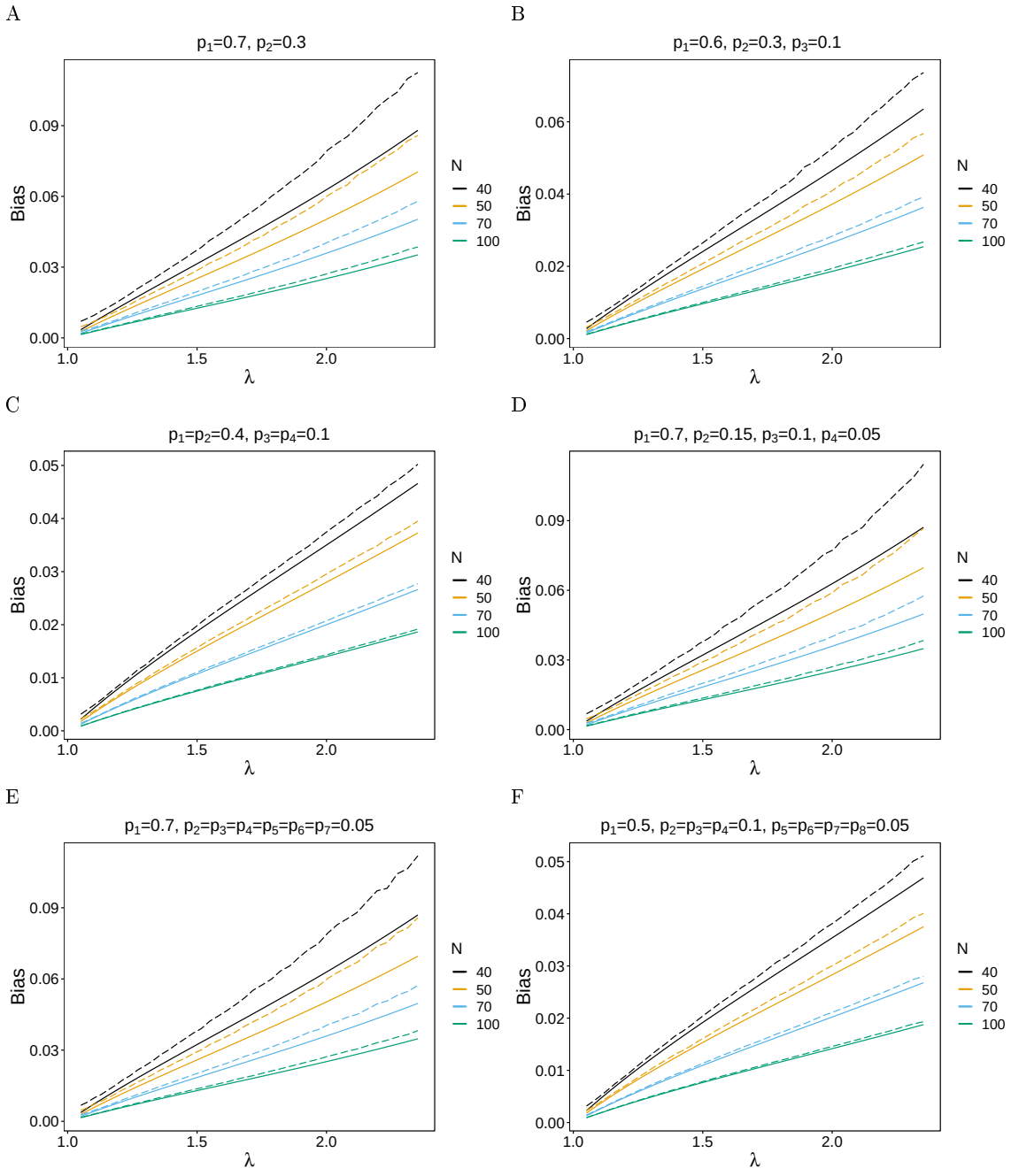


Figure 2: **Bias function**. Same as Figure 1

3 Alternative bias-corrected estimators

3.1 Bias in %

3.1.1 $\hat{\psi}$

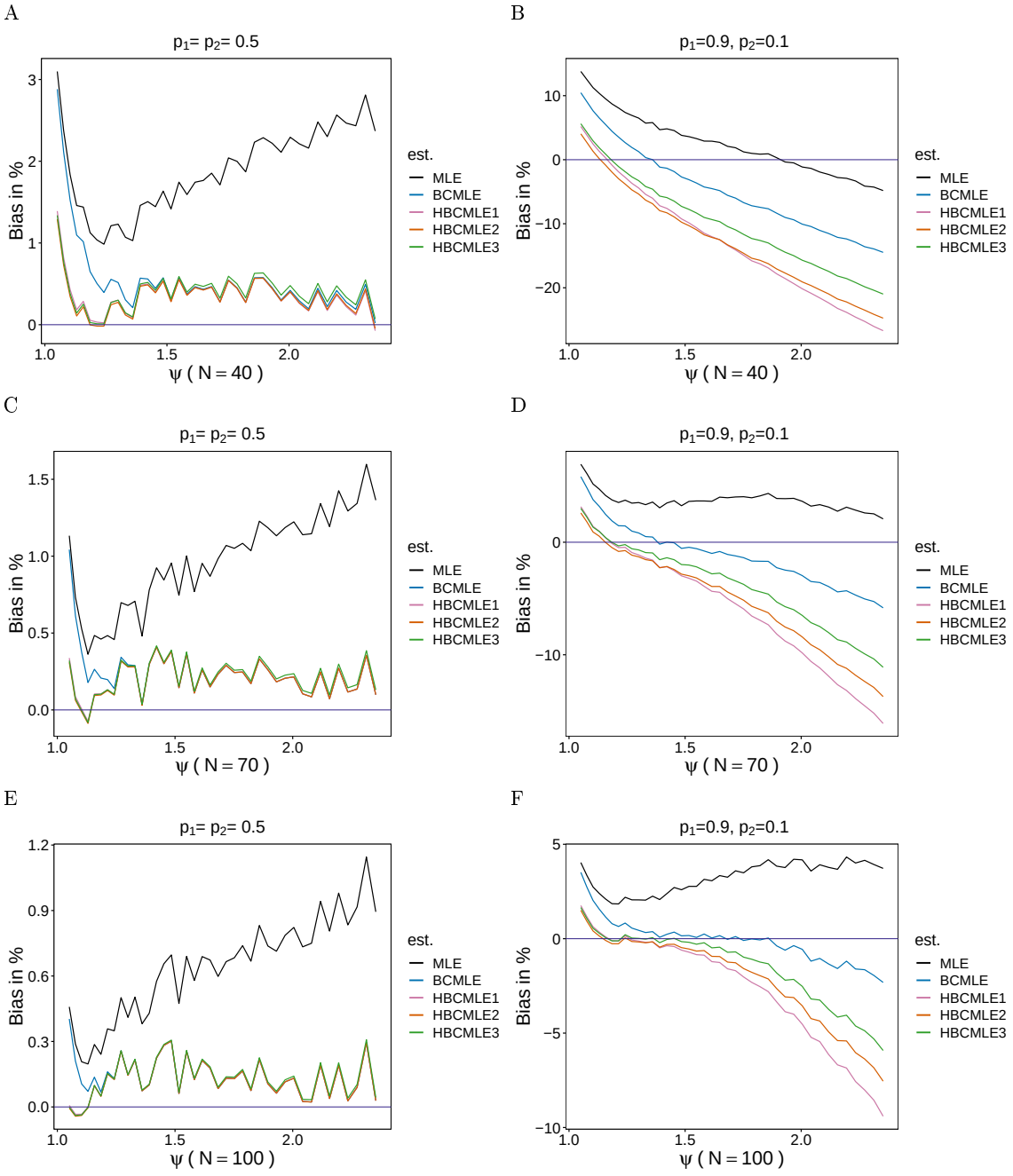


Figure 3: **Bias of alternative estimators in%**. The figure shows the bias in % of different estimators of $\psi = \frac{\lambda}{1-e^{-\lambda}}$ (average MOI). We have HBCMLE1 $= (\hat{\lambda} - \text{bias}(\hat{\lambda})|_{\lambda=\hat{\lambda}})(1 - q_1 - q_3 + q_2)$ where q_1, q_2 and q_3 are evaluated at MLE. Further, HBCMLE2 $= \hat{\lambda} - \text{bias}(\hat{\lambda})|_{\lambda=\hat{\lambda}} \frac{1-q_3}{1-q_1}$ and HBCMLE3 $= (\hat{\lambda} - \text{bias}(\hat{\lambda})|_{\lambda=\hat{\lambda}})(1 - q_1 - q_3 + q_2)$ and HBCMLE4 $= \hat{\lambda} * (1 - q_3 - q_1 + q_2) - \text{bias}(\hat{\lambda})|_{\lambda=\hat{\lambda}}$ and HBCMLE5 $= (\hat{\lambda} - \text{bias}(\hat{\lambda})|_{\lambda=\hat{\lambda}^{(\text{bc})}}) * (1 - q_3 - q_1 + q_2)$ where q_1, q_2 and q_3 are evaluated at BCMLE. The transformation $\frac{\lambda}{1-e^{-\lambda}}$ is applied afterwards.

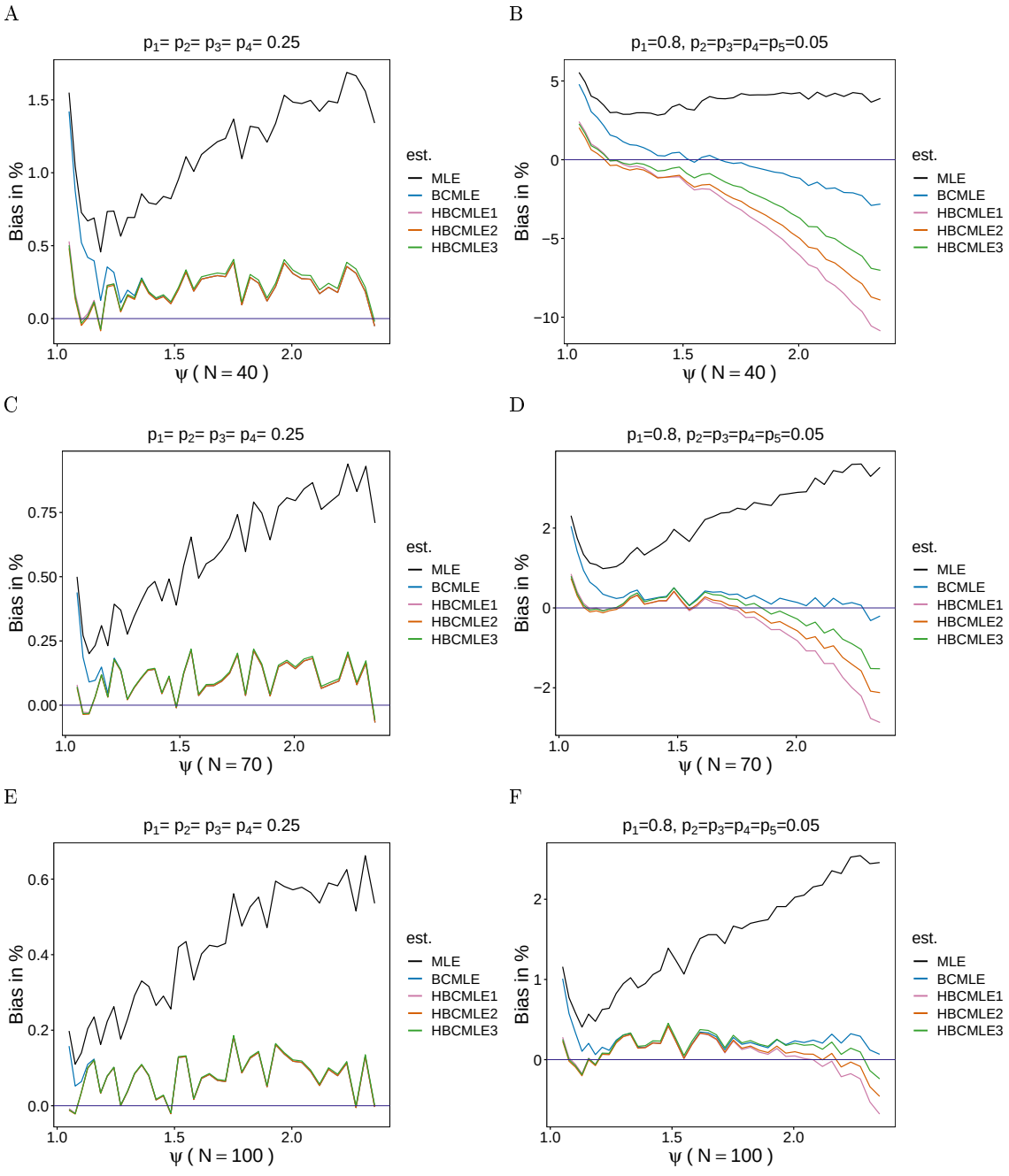


Figure 4: Same as Figure 3 but for different lineage-frequency distributions.

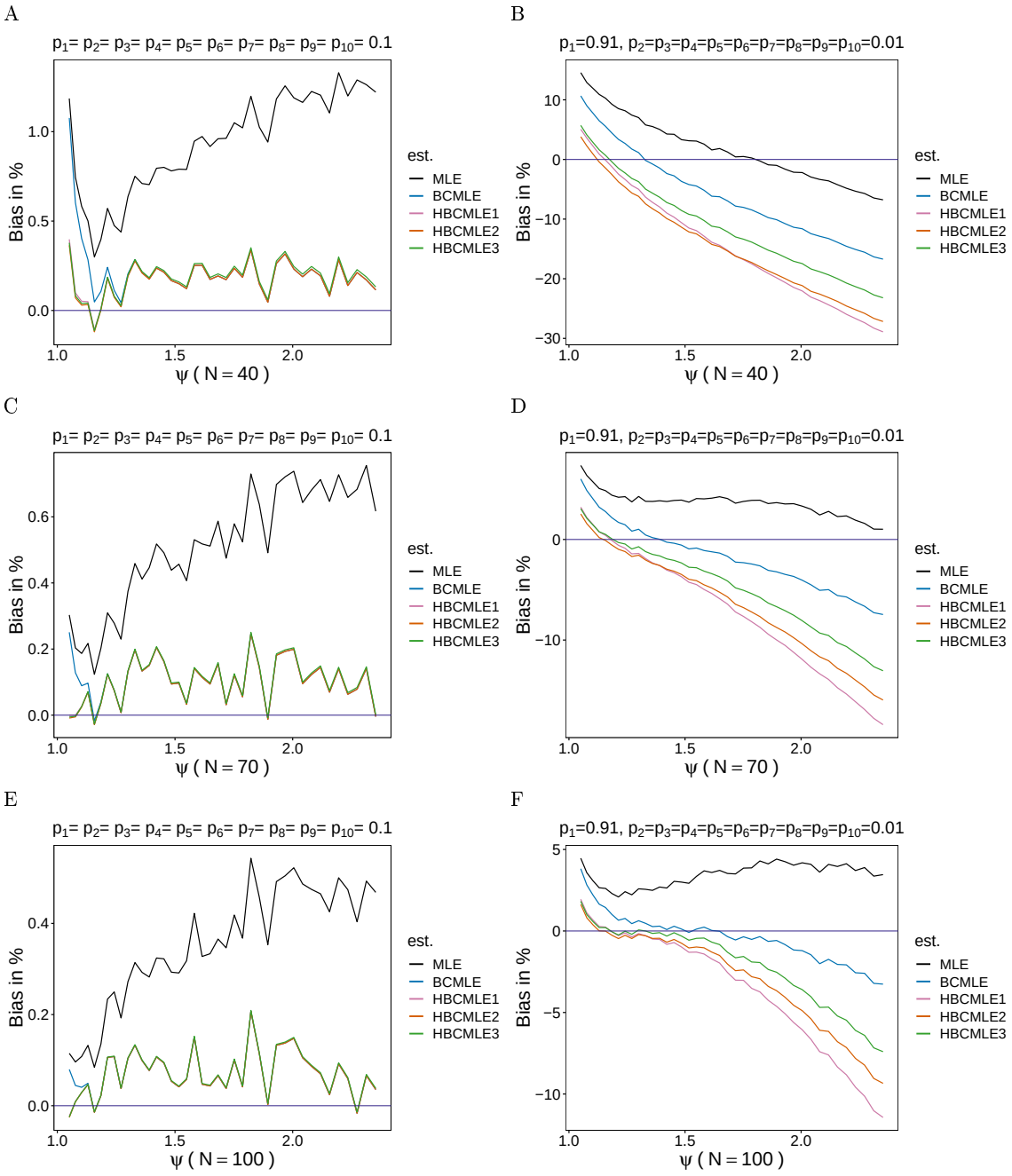


Figure 5: Same as Figure 3 but for different lineage-frequency distributions.

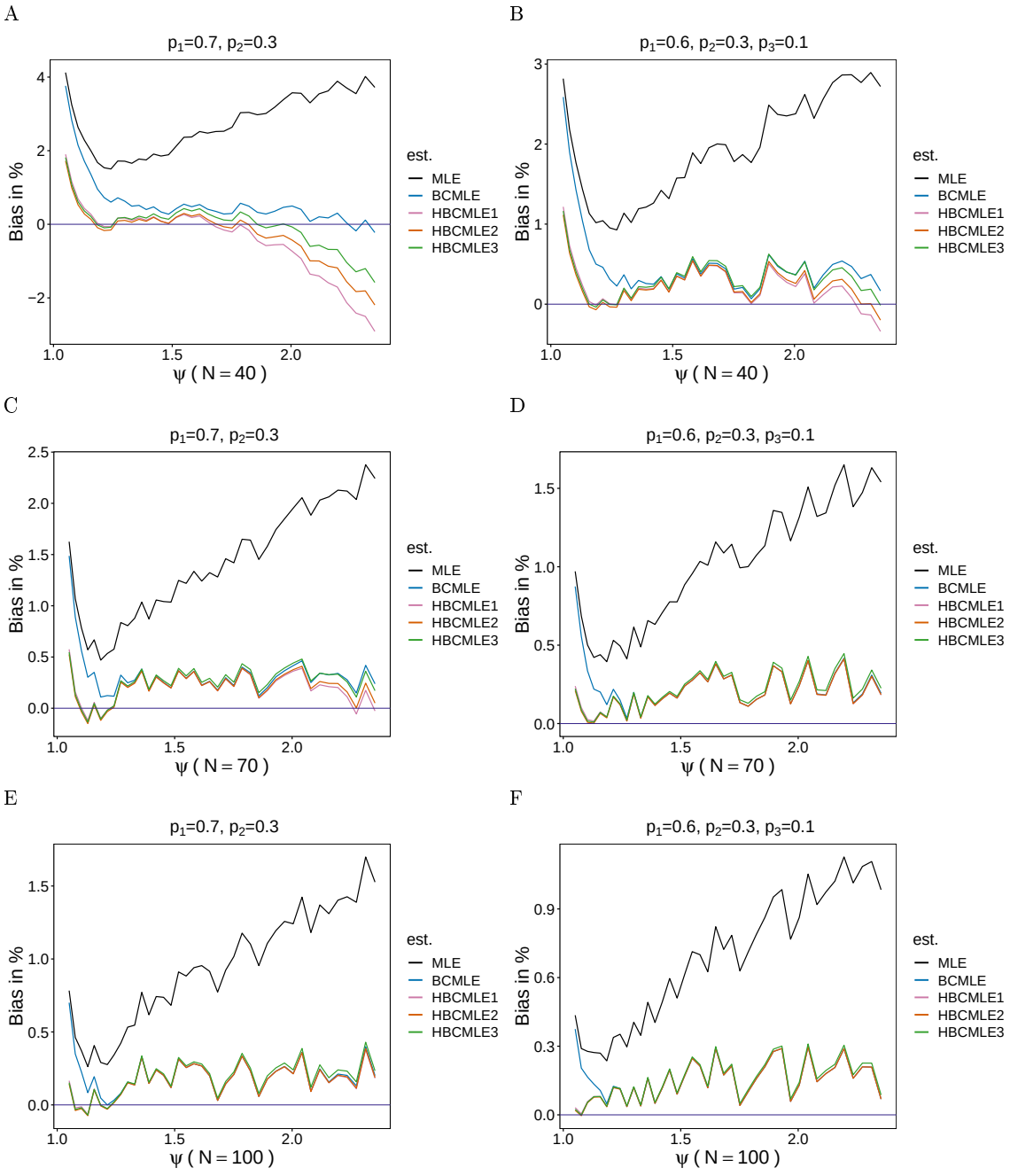


Figure 6: **Bias of alternative estimators in %**. Similar to Figure 3.

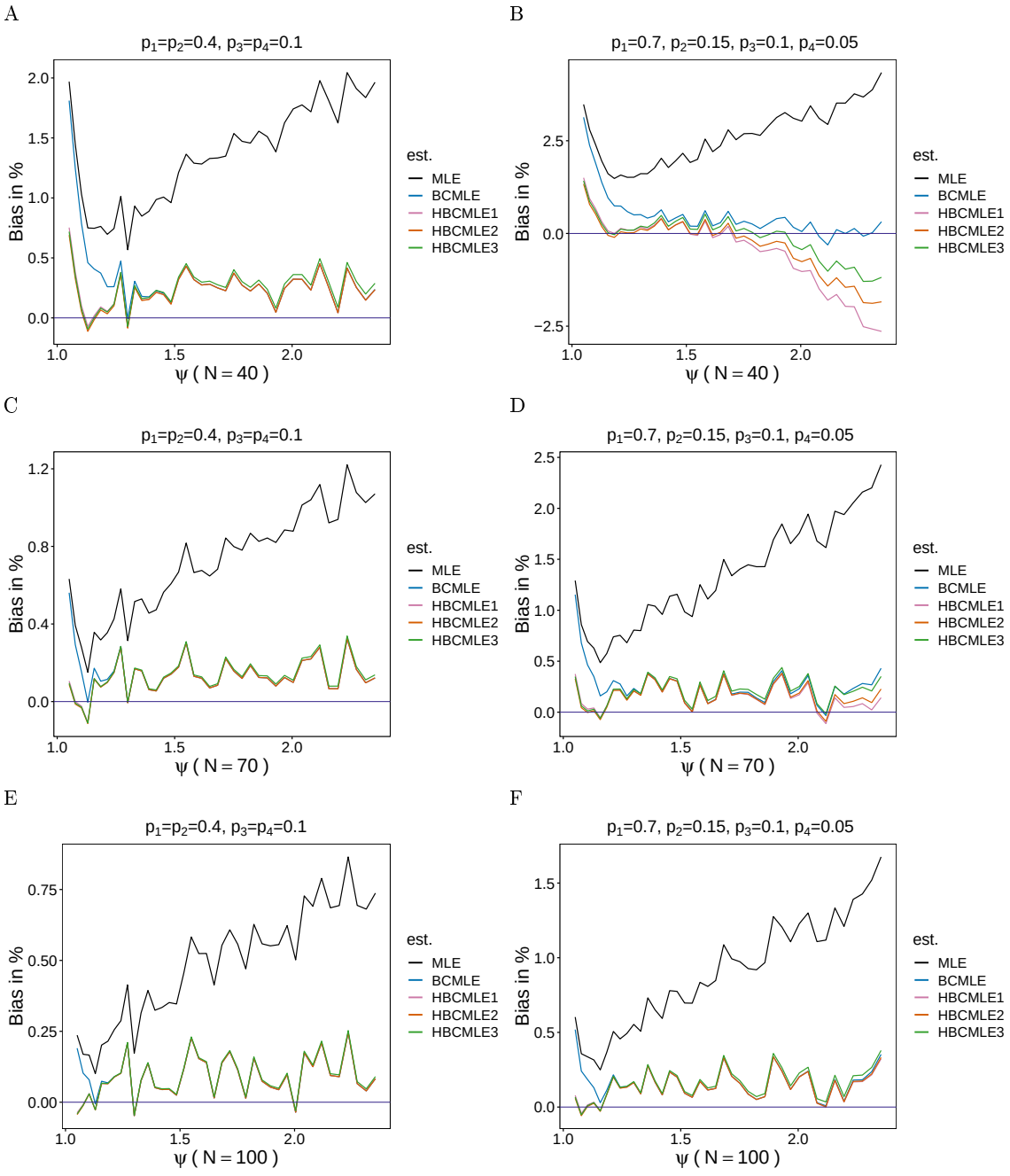


Figure 7: Same as Figure 3 but for different lineage-frequency distributions.

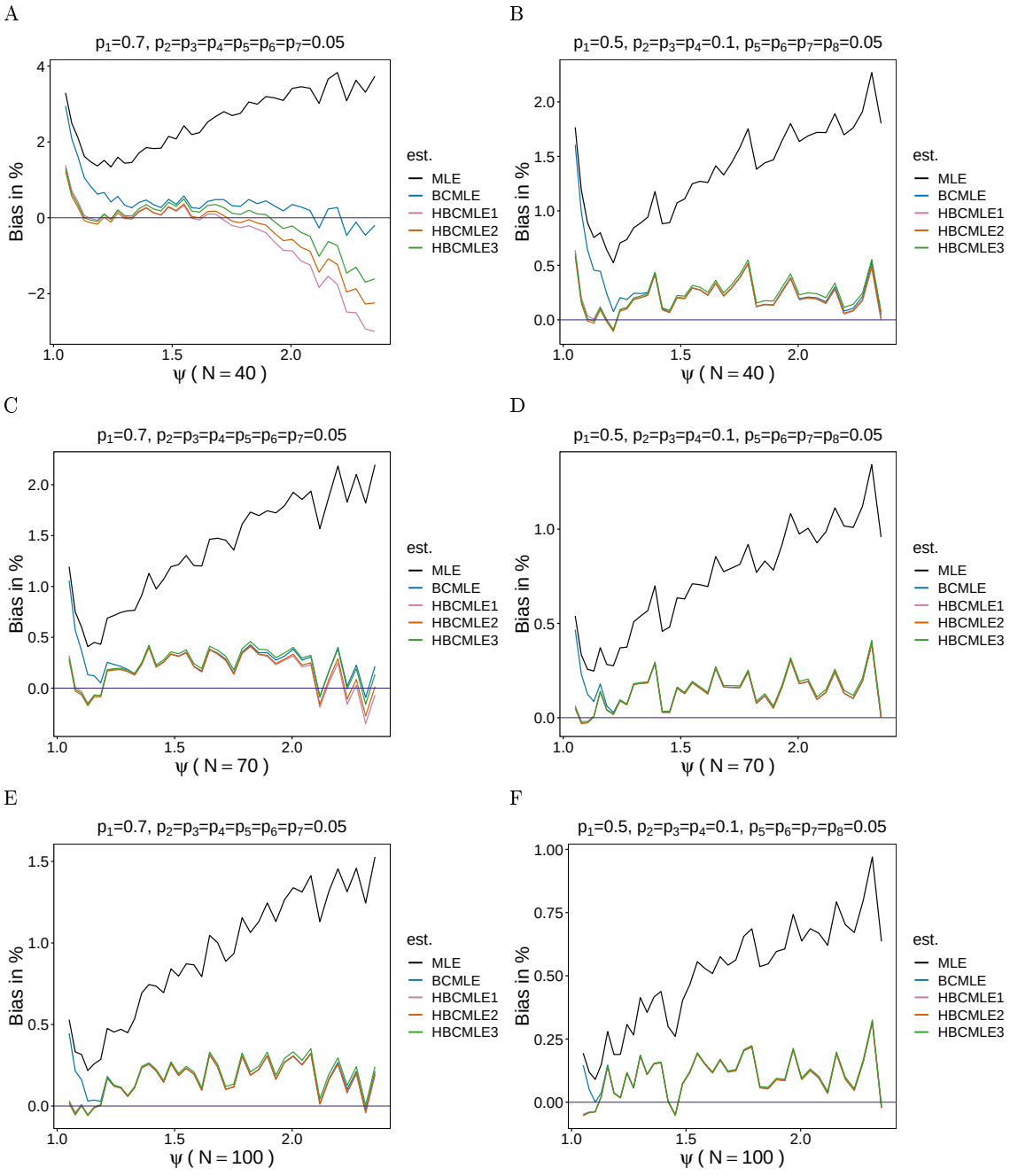


Figure 8: Same as Figure 3 but for different lineage-frequency distributions.

3.2 CV in percent

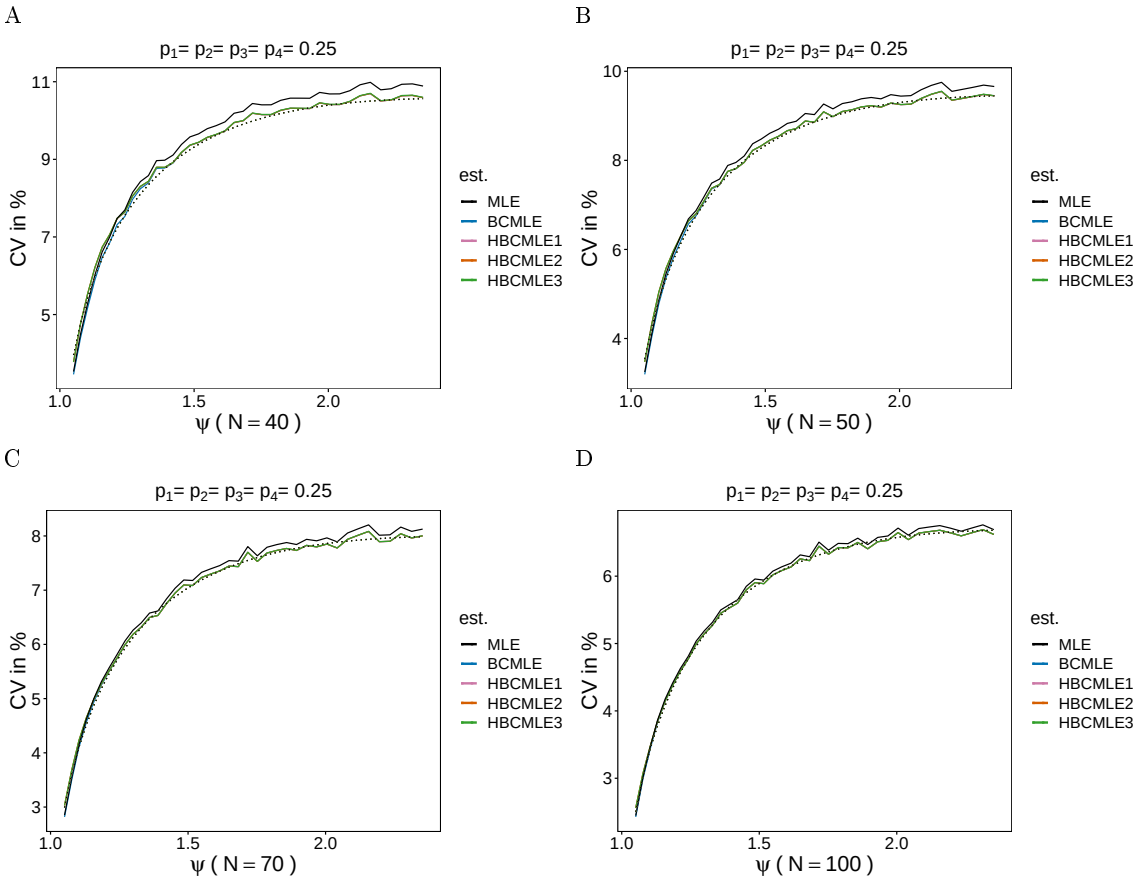
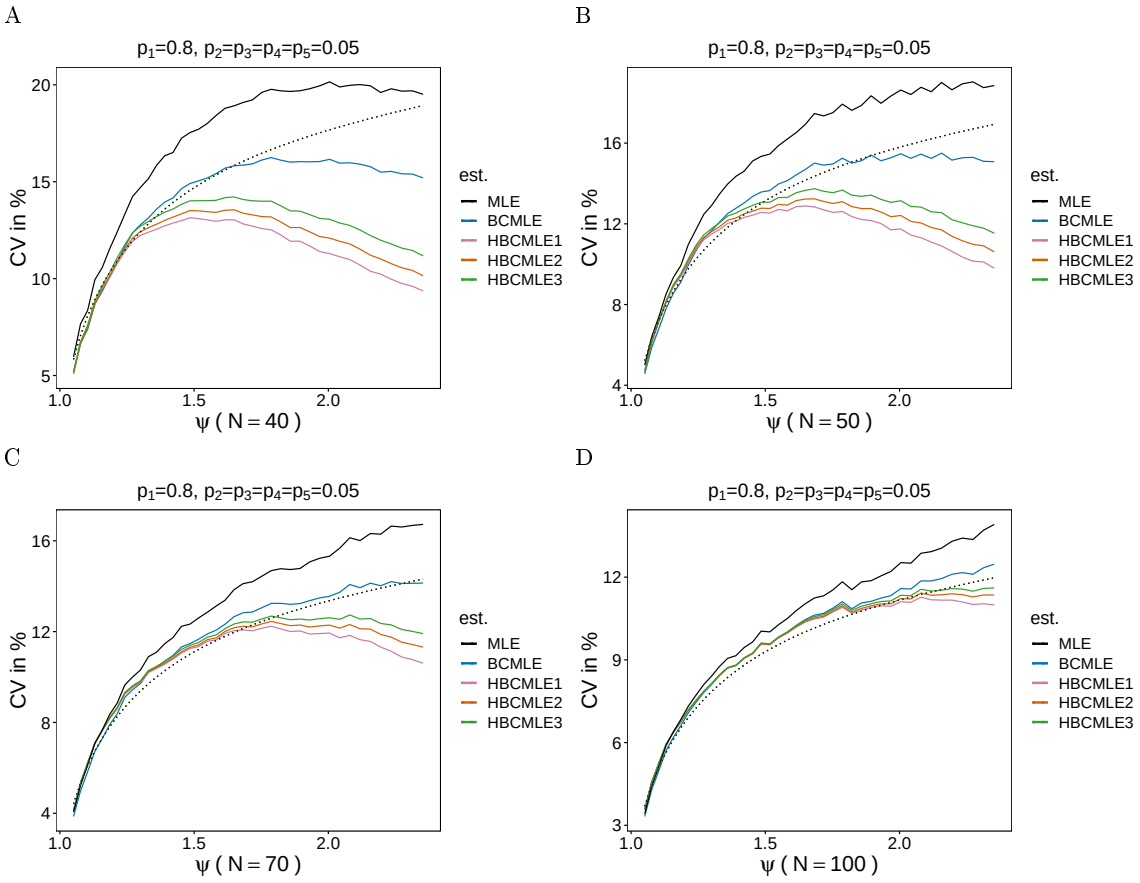
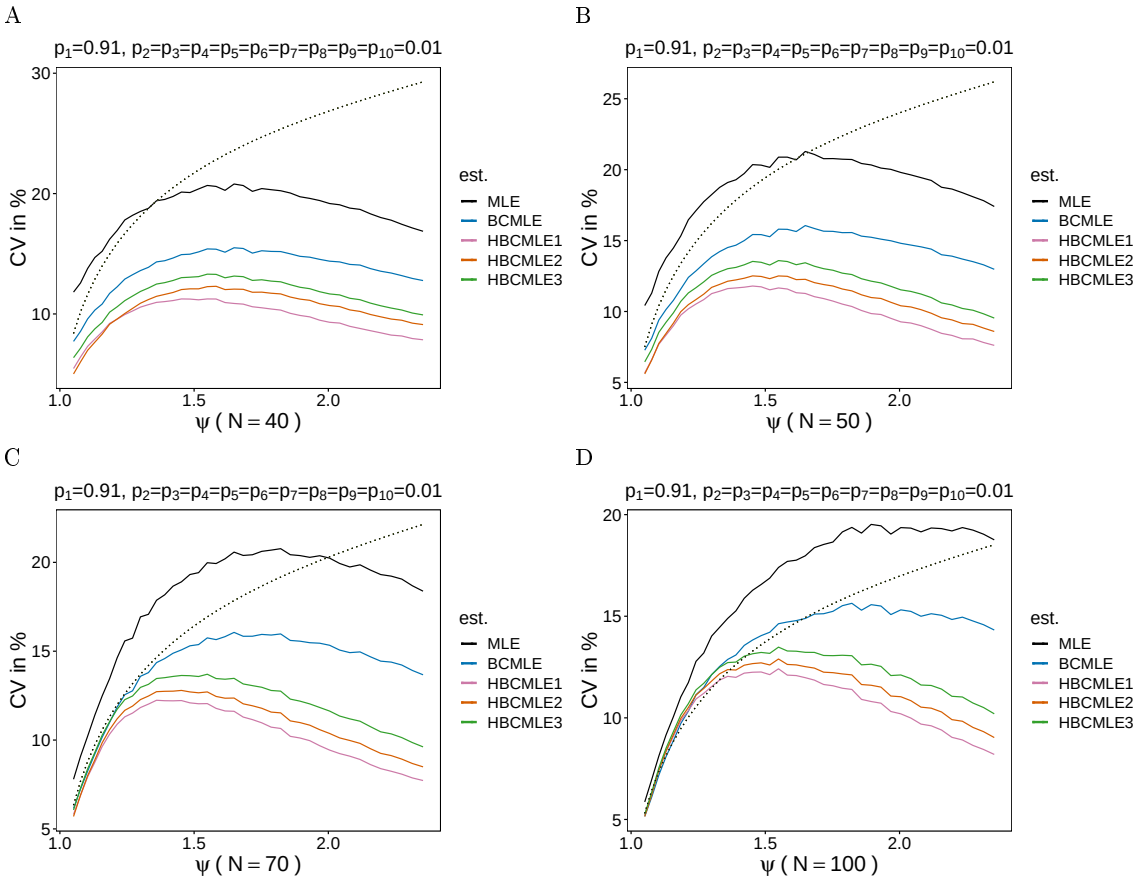
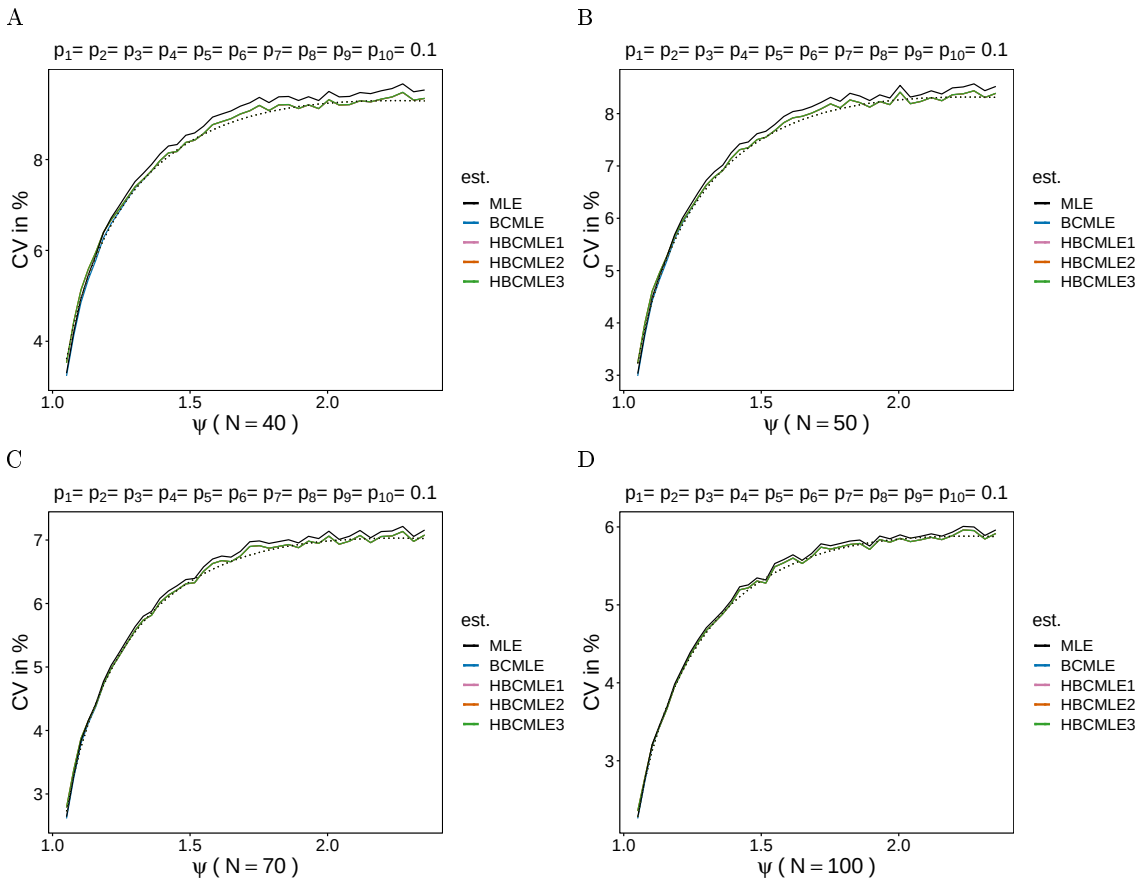
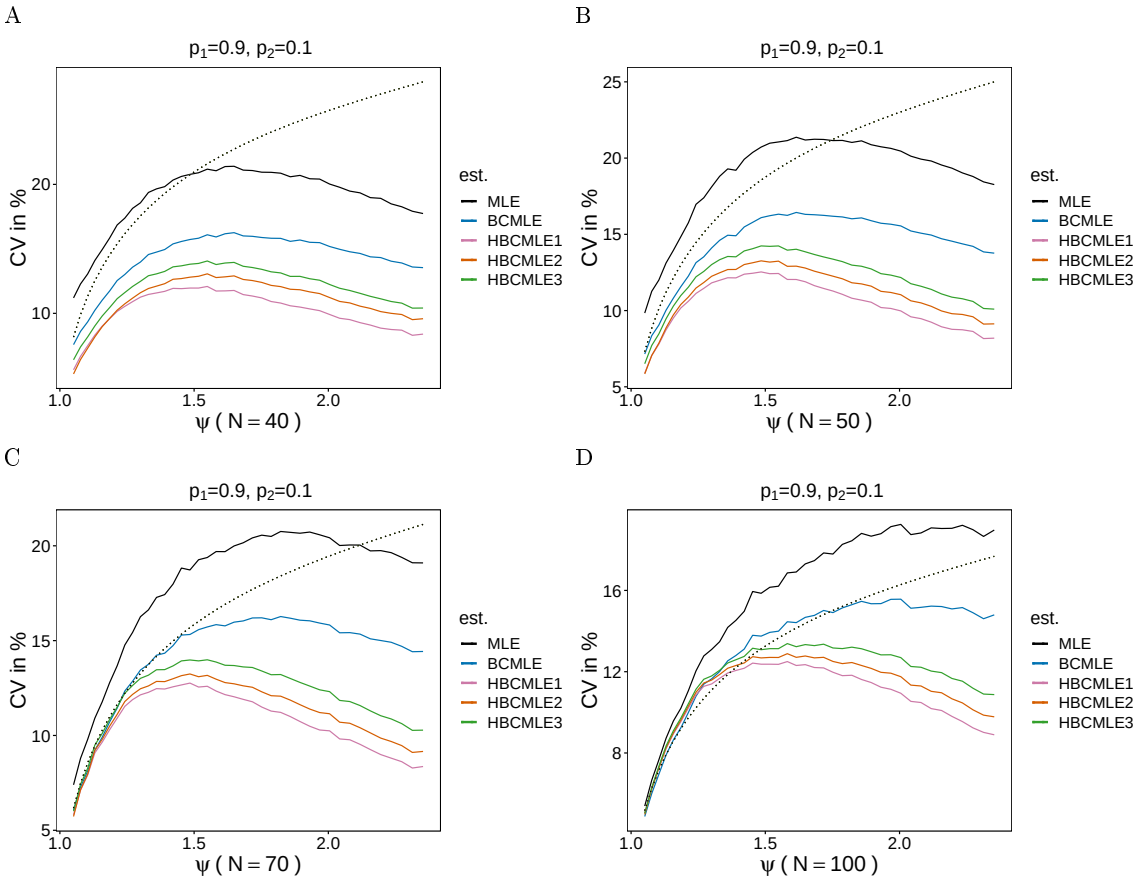


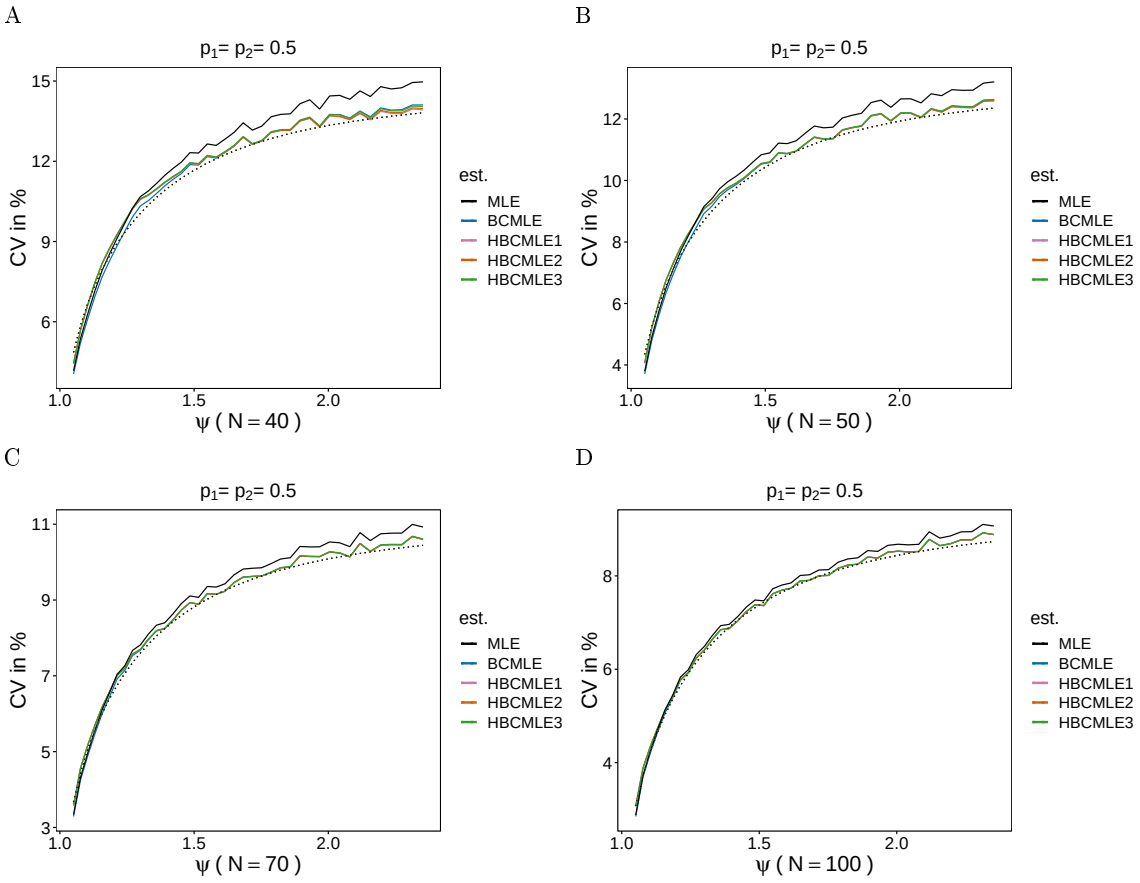
Figure 9: **CV of alternative estimators in %**. Same as Figure **??** but for CV

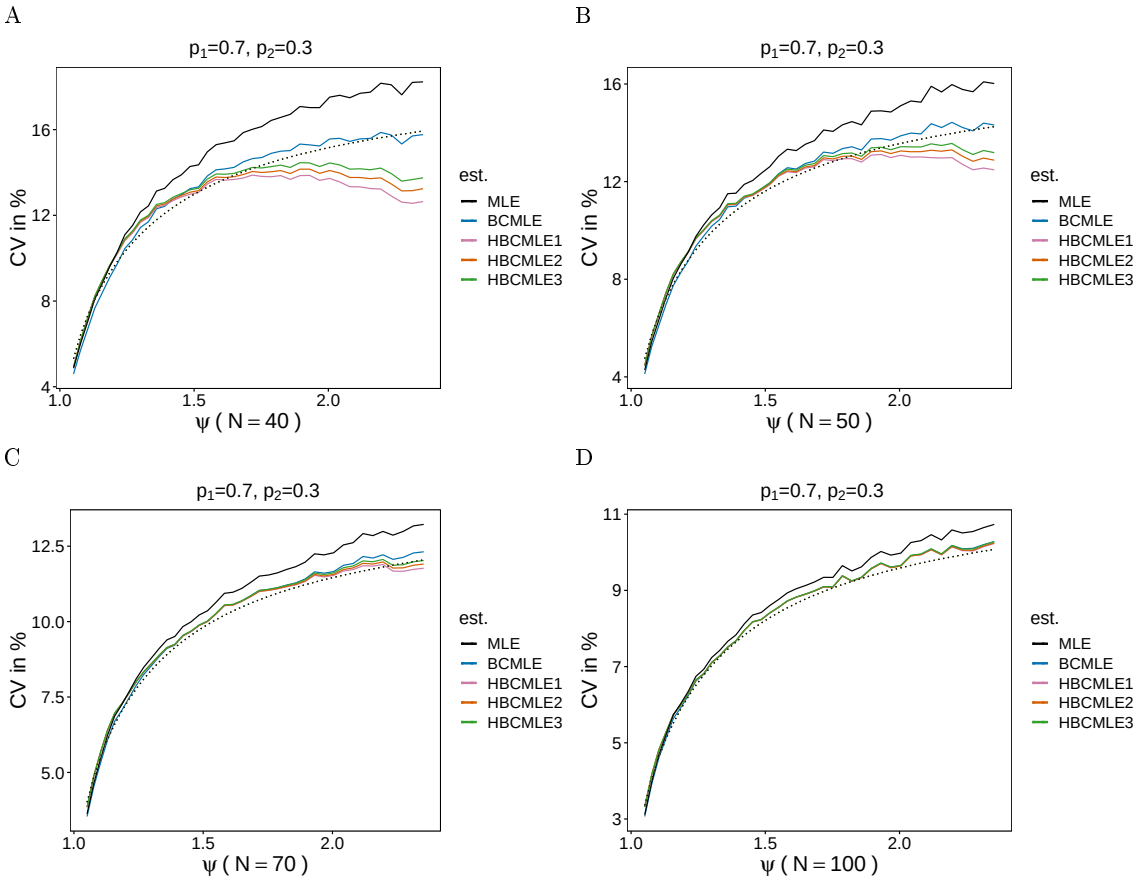

A
p_1=0.8, p_2=p_3=p_4=p_5=0.05
CV in %
ψ (N = 40)
B
p_1=0.8, p_2=p_3=p_4=p_5=0.05
CV in %
ψ (N = 50)
C
p_1=0.8, p_2=p_3=p_4=p_5=0.05
CV in %
ψ (N = 70)
D
p_1=0.8, p_2=p_3=p_4=p_5=0.05
CV in %
ψ (N = 100)
est.
MLE
BCMLE
HBCMLE1
HBCMLE2
HBCMLE3

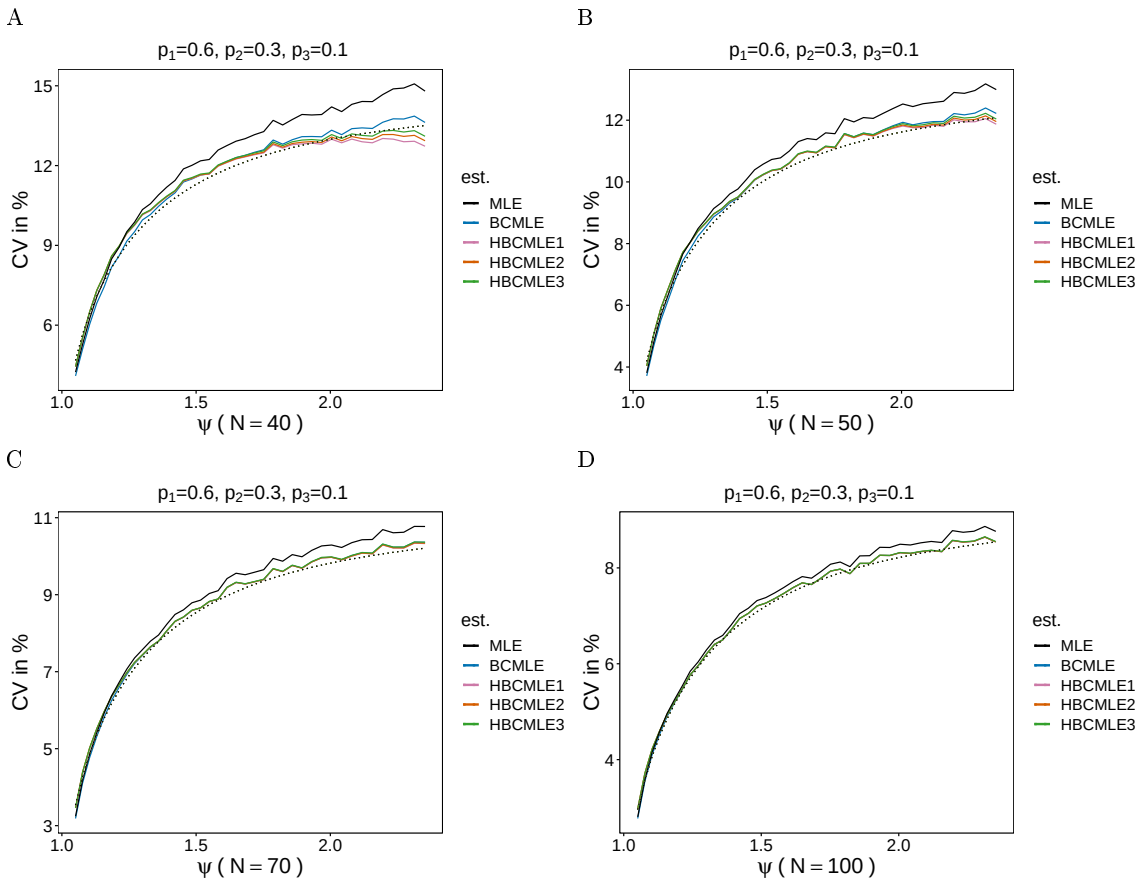

A
p_1=0.91, p_2=p_3=p_4=p_5=p_6=p_7=p_8=p_9=p_{10}=0.01
CV in %
ψ (N = 40)
B
p_1=0.91, p_2=p_3=p_4=p_5=p_6=p_7=p_8=p_9=p_{10}=0.01
CV in %
ψ (N = 50)
C
p_1=0.91, p_2=p_3=p_4=p_5=p_6=p_7=p_8=p_9=p_{10}=0.01
CV in %
ψ (N = 70)
D
p_1=0.91, p_2=p_3=p_4=p_5=p_6=p_7=p_8=p_9=p_{10}=0.01
CV in %
ψ (N = 100)
est.
MLE
BCMLE
HBCMLE1
HBCMLE2
HBCMLE3

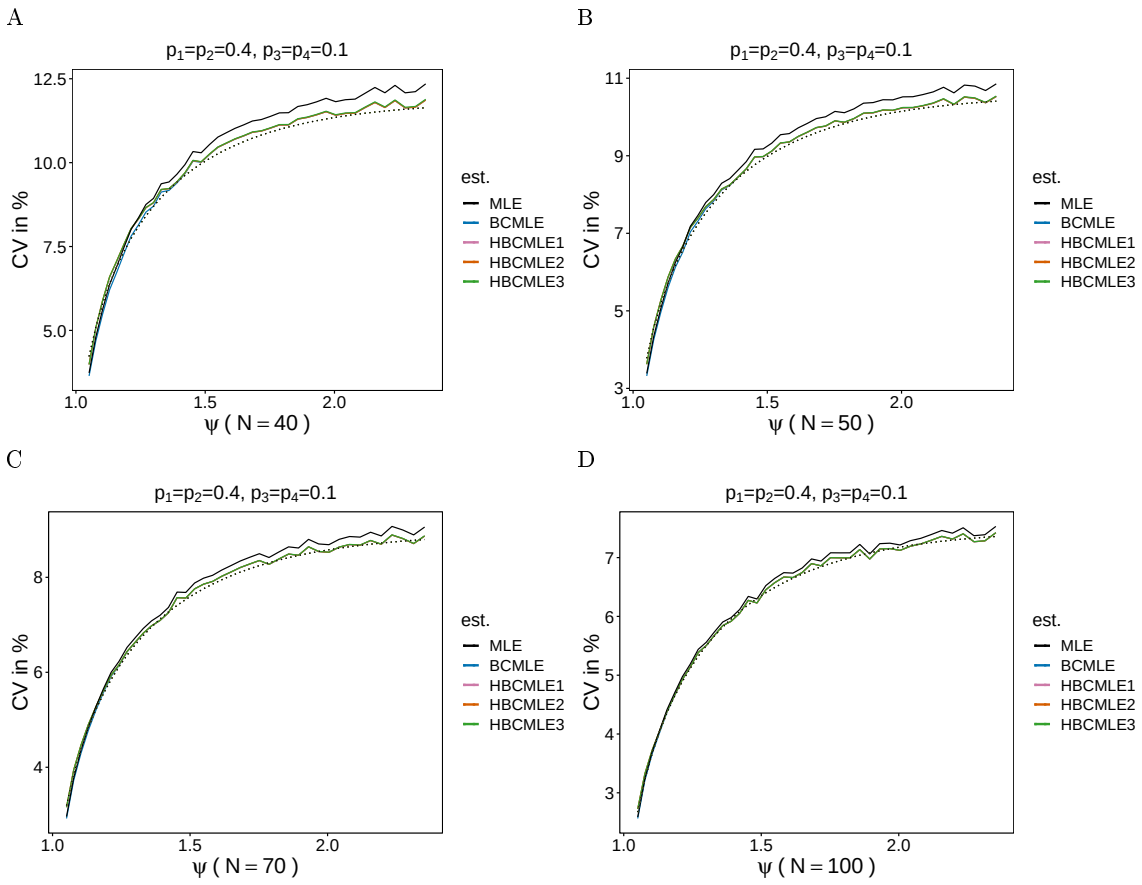

A
B
C
D
p1= p2= p3= p4= p5= p6= p7= p8= p9= p10= 0.1
CV in %
ψ (N = 40)
ψ (N = 50)
ψ (N = 70)
ψ (N = 100)
est.
MLE
BCMLE
HBCMLE1
HBCMLE2
HBCMLE3

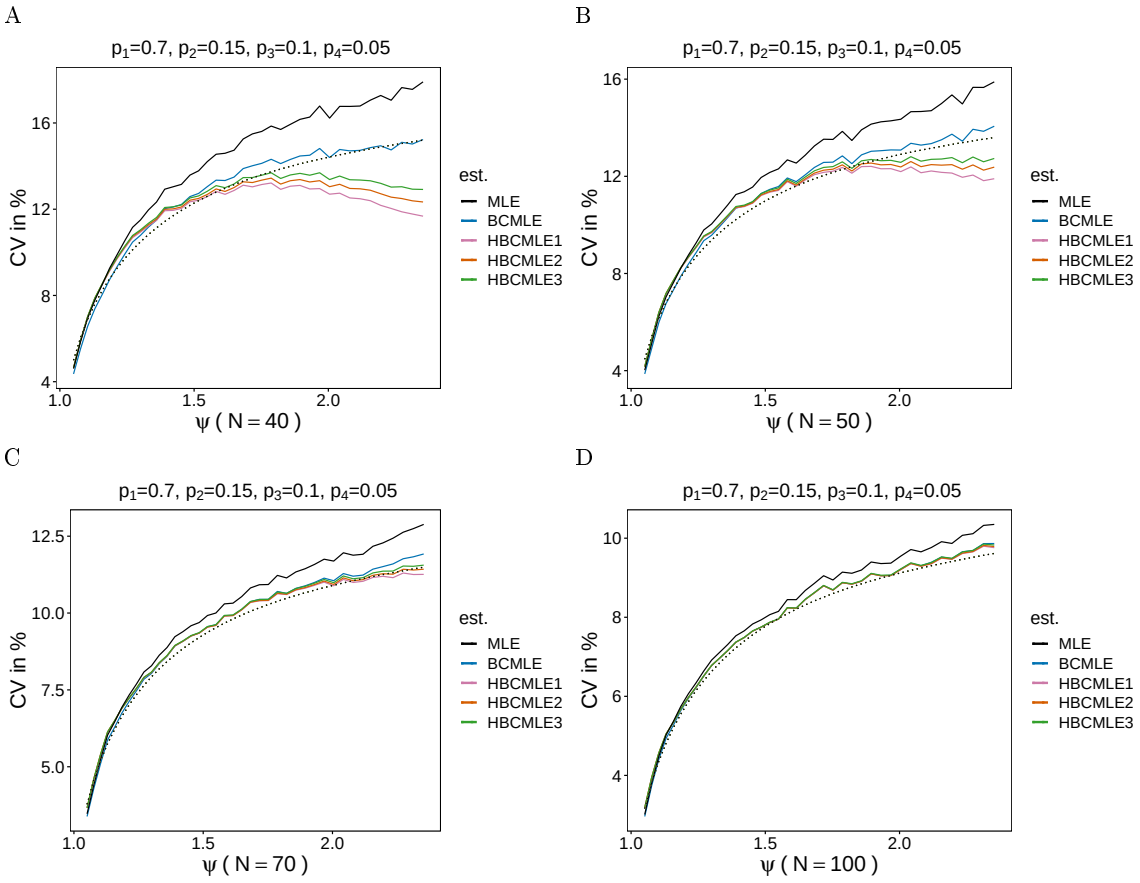

A
B
C
D
p1=0.9, p2=0.1
CV in %
ψ (N = 40)
ψ (N = 50)
ψ (N = 70)
ψ (N = 100)
est.
MLE
BCMLE
HBCMLE1
HBCMLE2
HBCMLE3

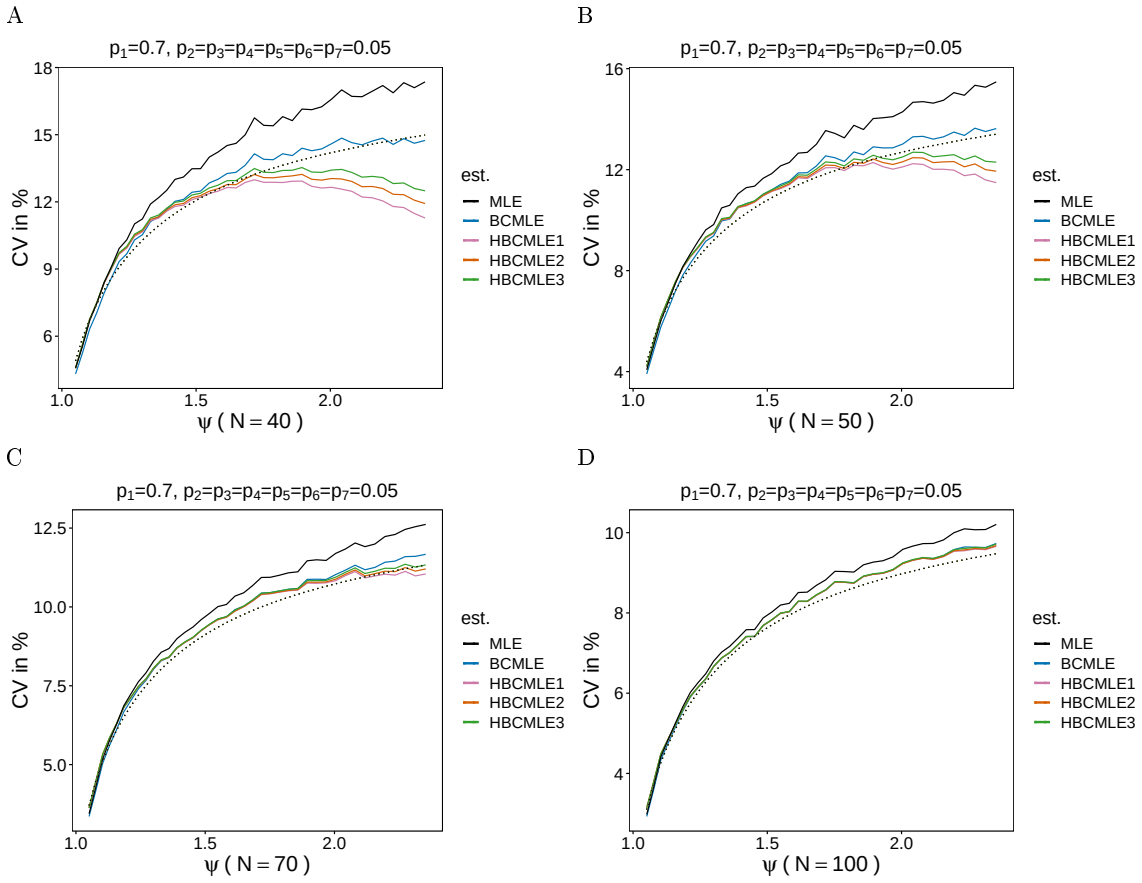

A
p_1= p_2= 0.5
CV in %
ψ (N = 40)
B
p_1= p_2= 0.5
CV in %
ψ (N = 50)
C
p_1= p_2= 0.5
CV in %
ψ (N = 70)
D
p_1= p_2= 0.5
CV in %
ψ (N = 100)
est.
MLE
BCMLE
HBCMLE1
HBCMLE2
HBCMLE3

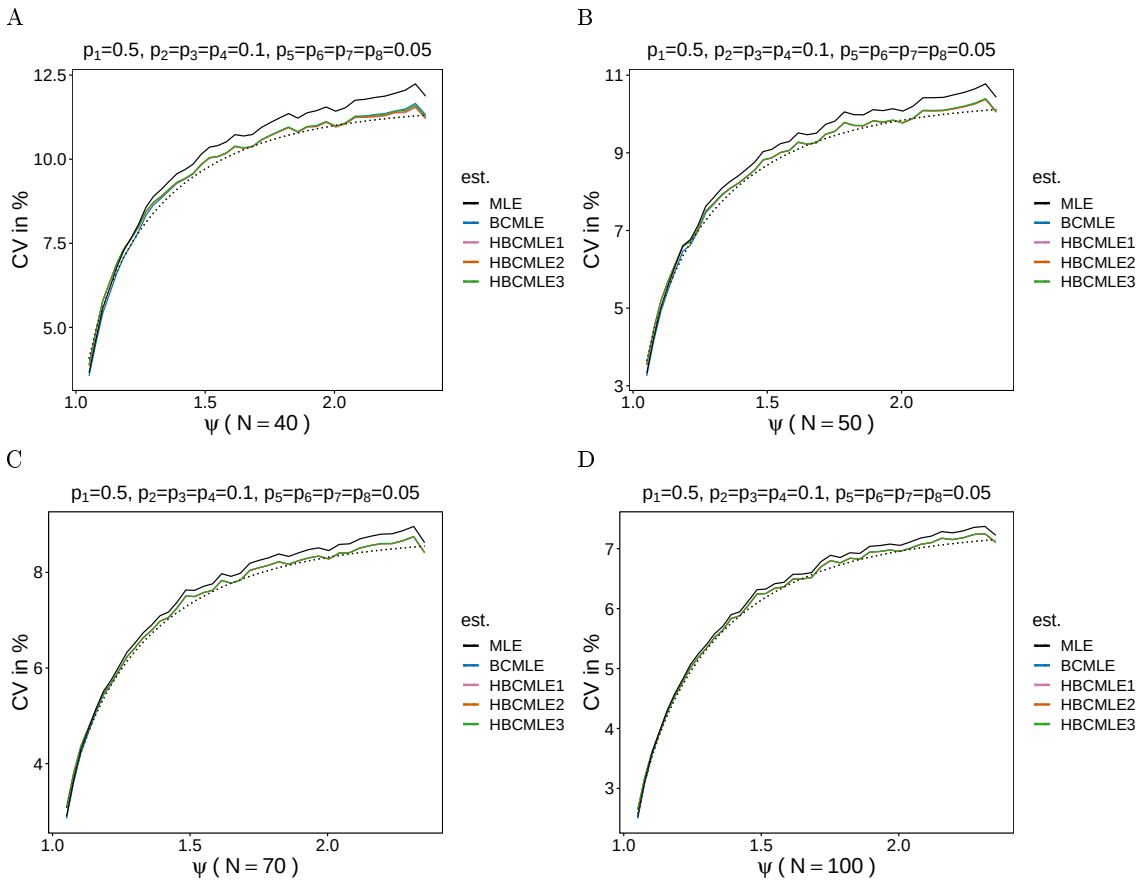

A
B
C
D
p1=0.7, p2=0.3
CV in %
ψ (N = 40)
ψ (N = 50)
ψ (N = 70)
ψ (N = 100)
est.
MLE
BCMLE
HBCMLE1
HBCMLE2
HBCMLE3


A
B
C
D
p1=0.6, p2=0.3, p3=0.1
CV in %
ψ (N = 40)
ψ (N = 50)
ψ (N = 70)
ψ (N = 100)
est.
MLE
BCMLE
HBCMLE1
HBCMLE2
HBCMLE3


A
p_1=p_2=0.4, p_3=p_4=0.1
CV in %
ψ (N = 40)
B
p_1=p_2=0.4, p_3=p_4=0.1
CV in %
ψ (N = 50)
C
p_1=p_2=0.4, p_3=p_4=0.1
CV in %
ψ (N = 70)
D
p_1=p_2=0.4, p_3=p_4=0.1
CV in %
ψ (N = 100)
est.
MLE
BCMLE
HBCMLE1
HBCMLE2
HBCMLE3


A
B
C
D
p1=0.7, p2=0.15, p3=0.1, p4=0.05
CV in %
ψ (N = 40)
ψ (N = 50)
ψ (N = 70)
ψ (N = 100)
est.
MLE
BCMLE
HBCMLE1
HBCMLE2
HBCMLE3


A
B
C
D
p1=0.7, p2=p3=p4=p5=p6=p7=0.05
CV in %
ψ (N = 40)
ψ (N = 50)
ψ (N = 70)
ψ (N = 100)
est.
MLE
BCMLE
HBCMLE1
HBCMLE2
HBCMLE3


A
B
C
D
p1=0.5, p2=p3=p4=0.1, p5=p6=p7=p8=0.05
CV in %
ψ (N = 40)
ψ (N = 50)
ψ (N = 70)
ψ (N = 100)
est.
MLE
BCMLE
HBCMLE1
HBCMLE2
HBCMLE3

4 Ratio of bias to sd

4.1 $\hat{\lambda}$

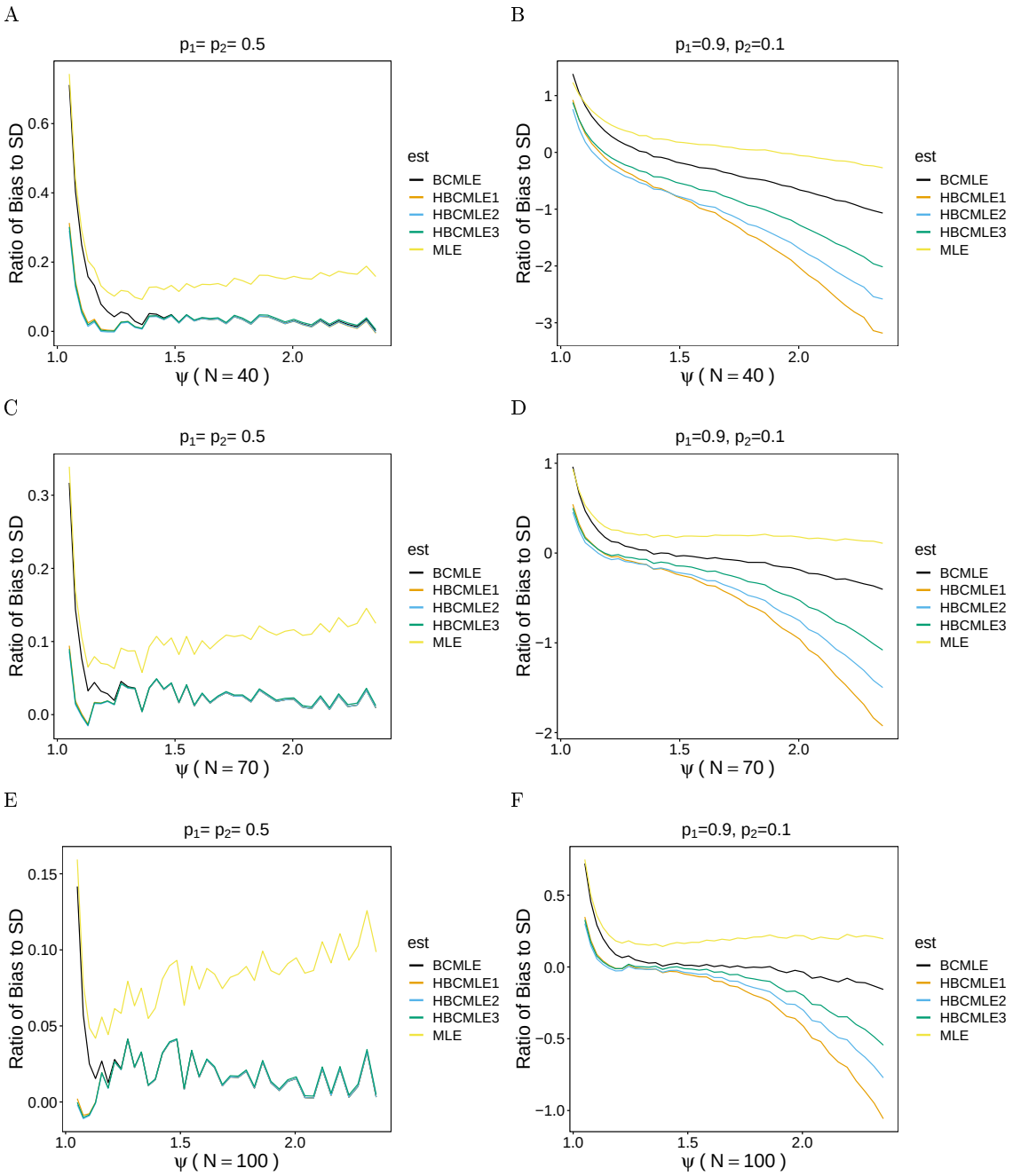


Figure 21: **Ratio of MLE to BCMLE**. The figure shows ratio of MLE to BCMLE. Different colors correspond to different ratios. The dotted and dashed lines correspond to true q 5 and r 2, respectively.

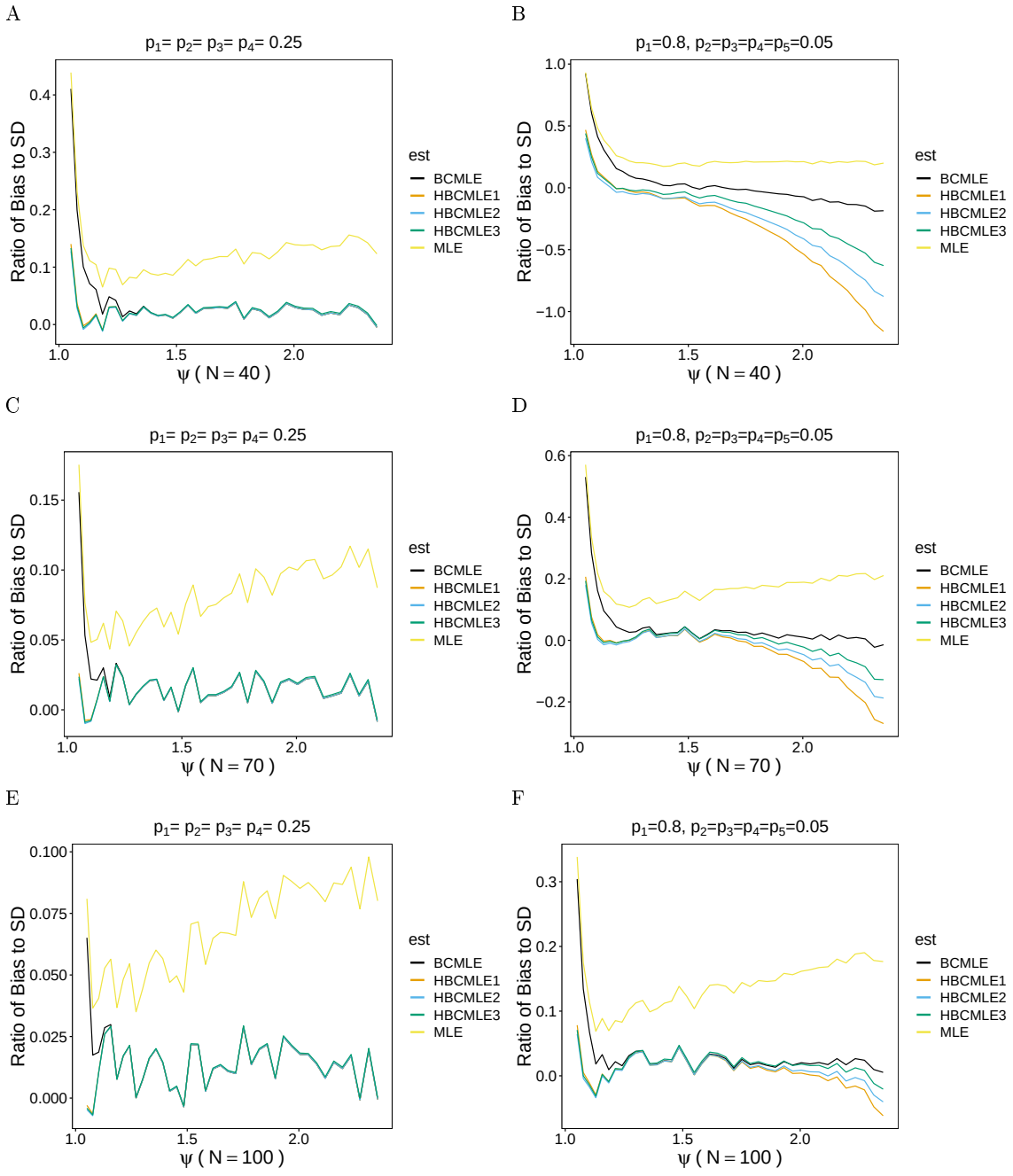


Figure 22: Same as Figure 21 but for different lineage-frequency distributions.

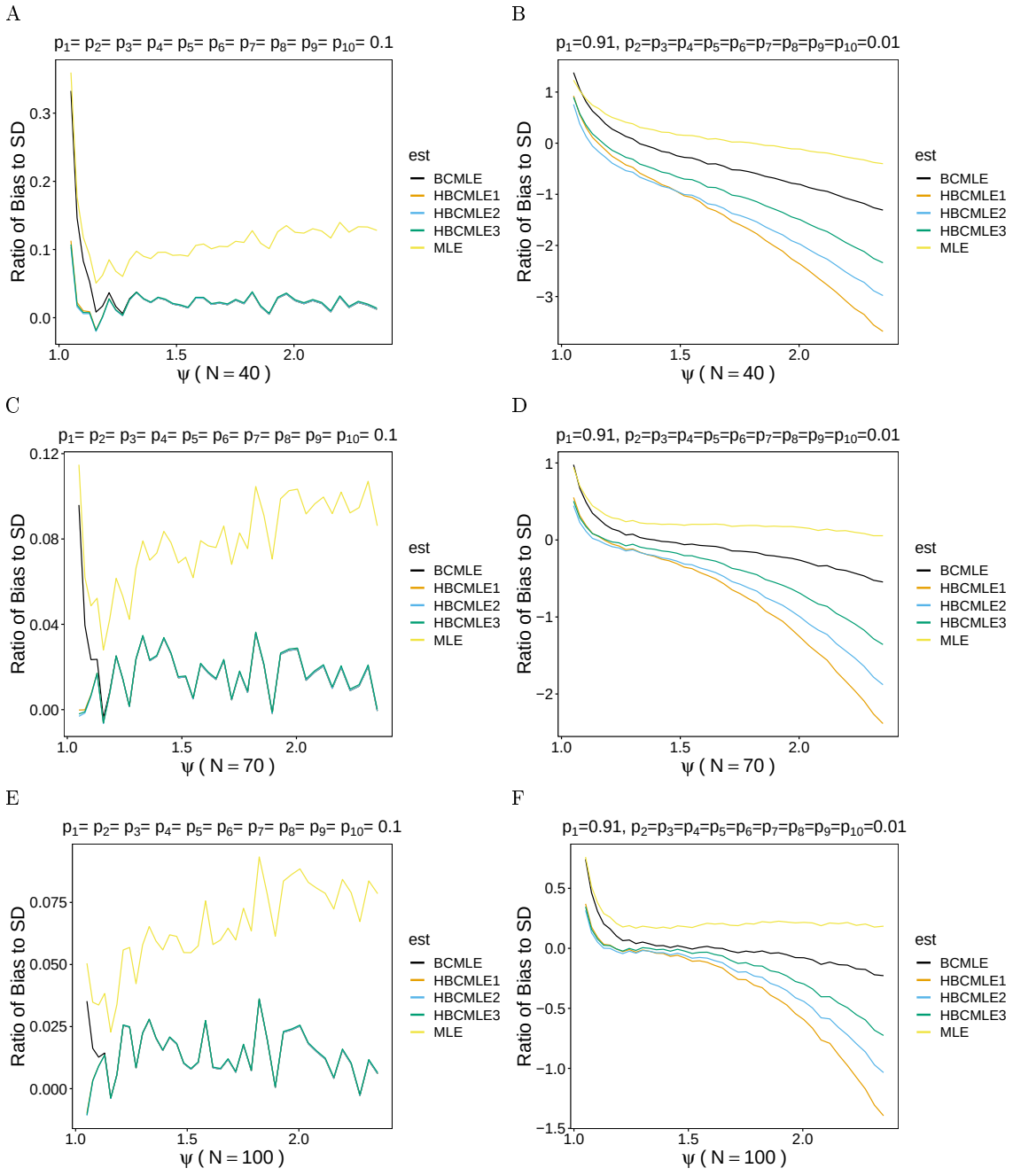


Figure 23: Same as Figure 21 but for different lineage-frequency distributions.

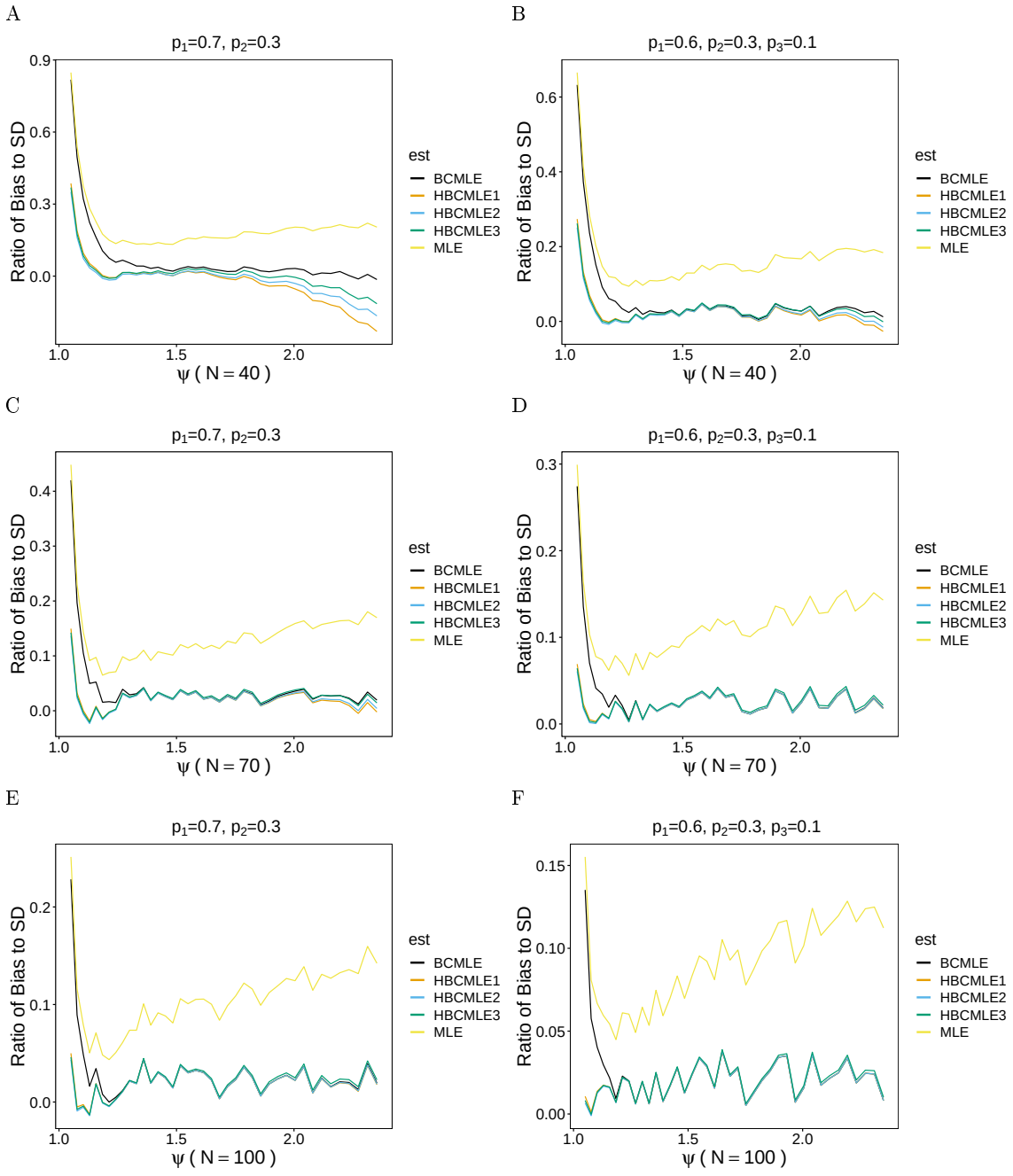


Figure 24: **Ratio of alternative estimators to BCMLE**. Similar to Figure 21

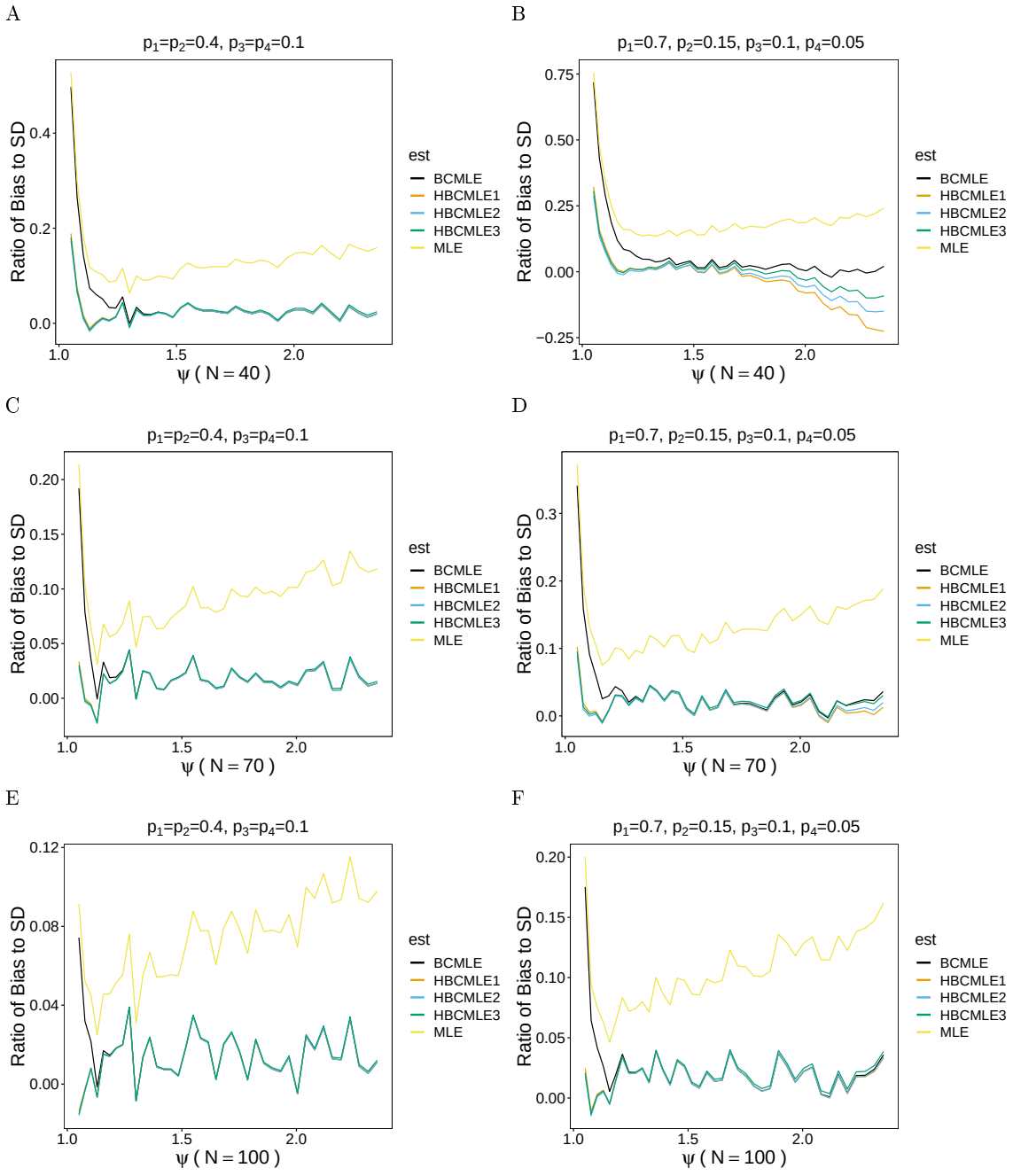


Figure 25: Same as Figure 21 but for different lineage-frequency distributions.

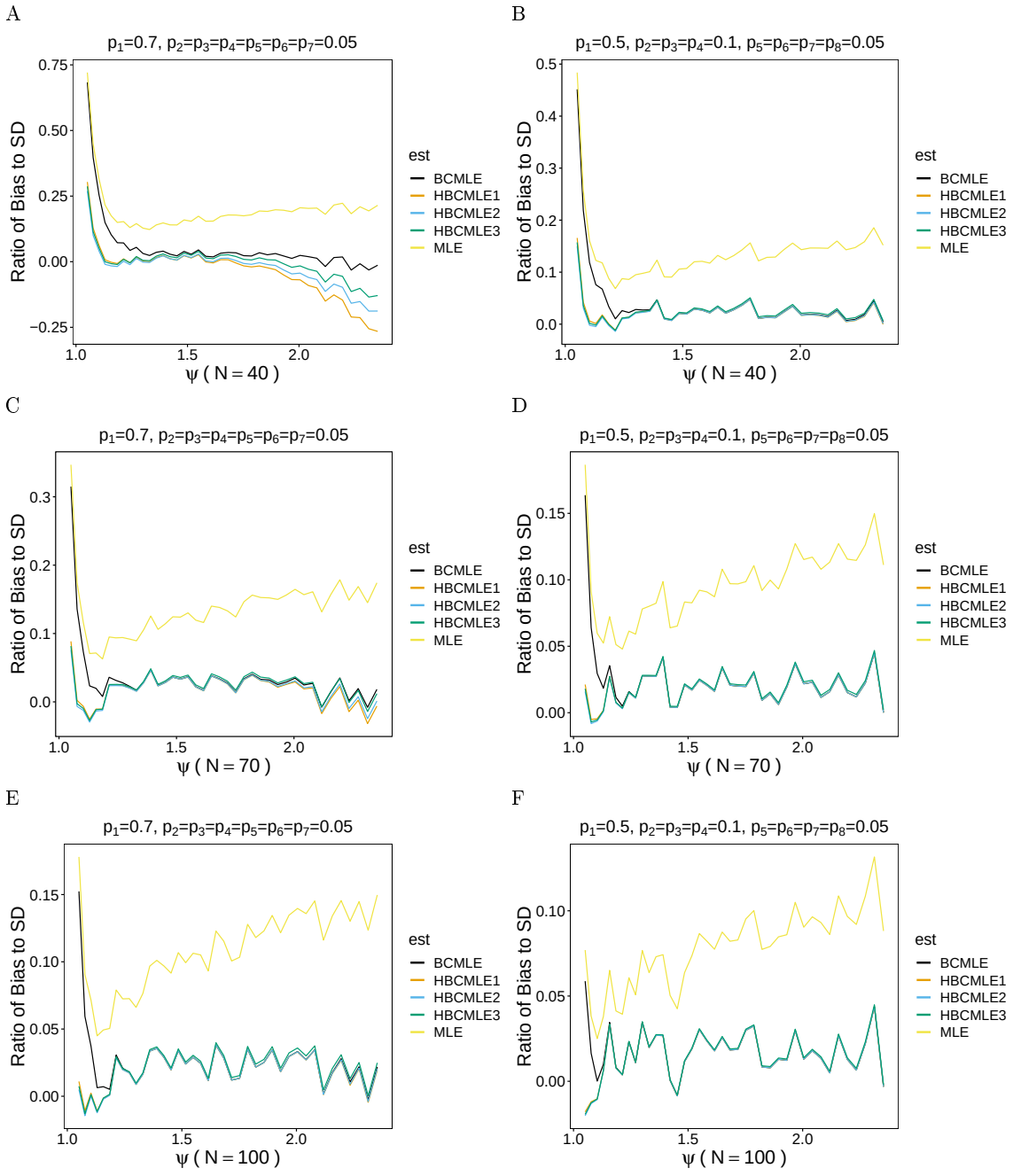


Figure 26: Same as Figure 21 but for different lineage-frequency distributions.

5 Histograms

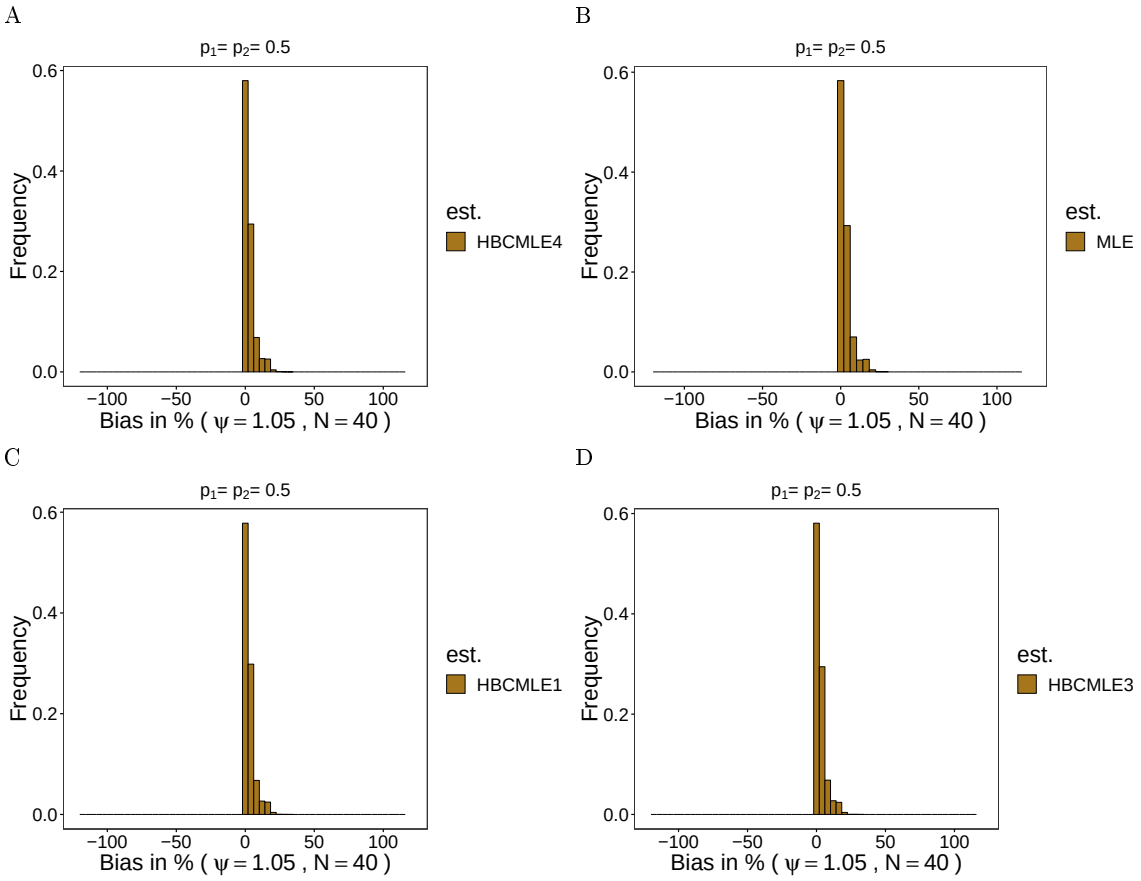


Figure 27: **Histograms**. The figure shows histograms of $10,000$ estimations of ψ for a specific sample size N and true ψ. Each plot corresponds to a different estimator.

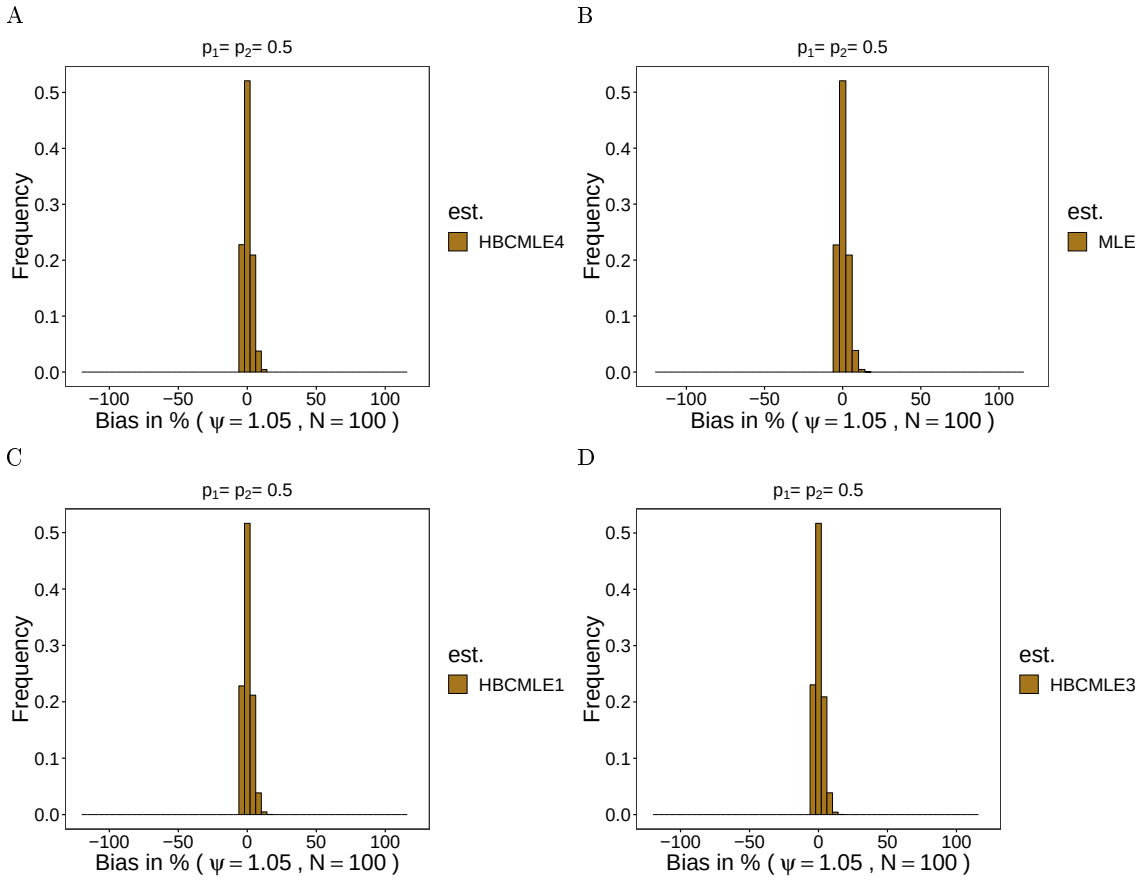


Figure 28: Same as Figure 27.

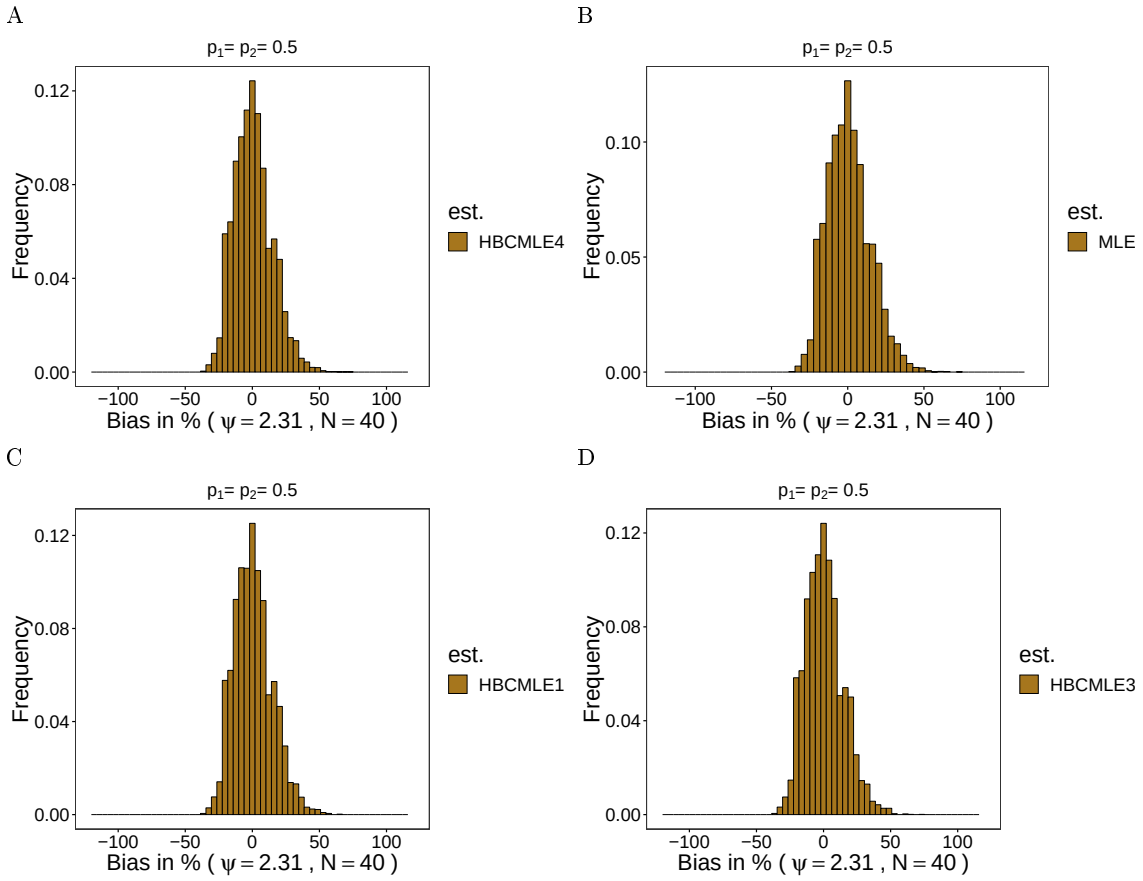


Figure 29: Same as Figure 27.

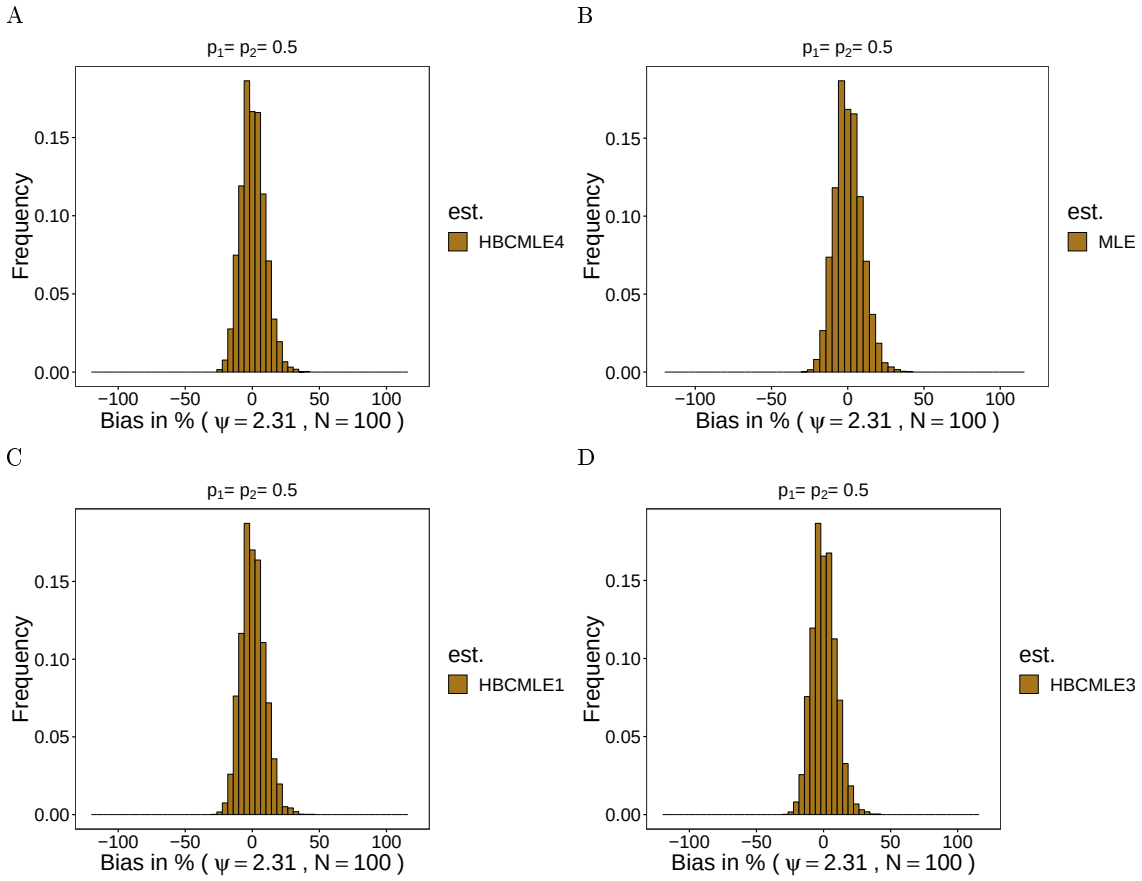


Figure 30: Same as Figure 27.

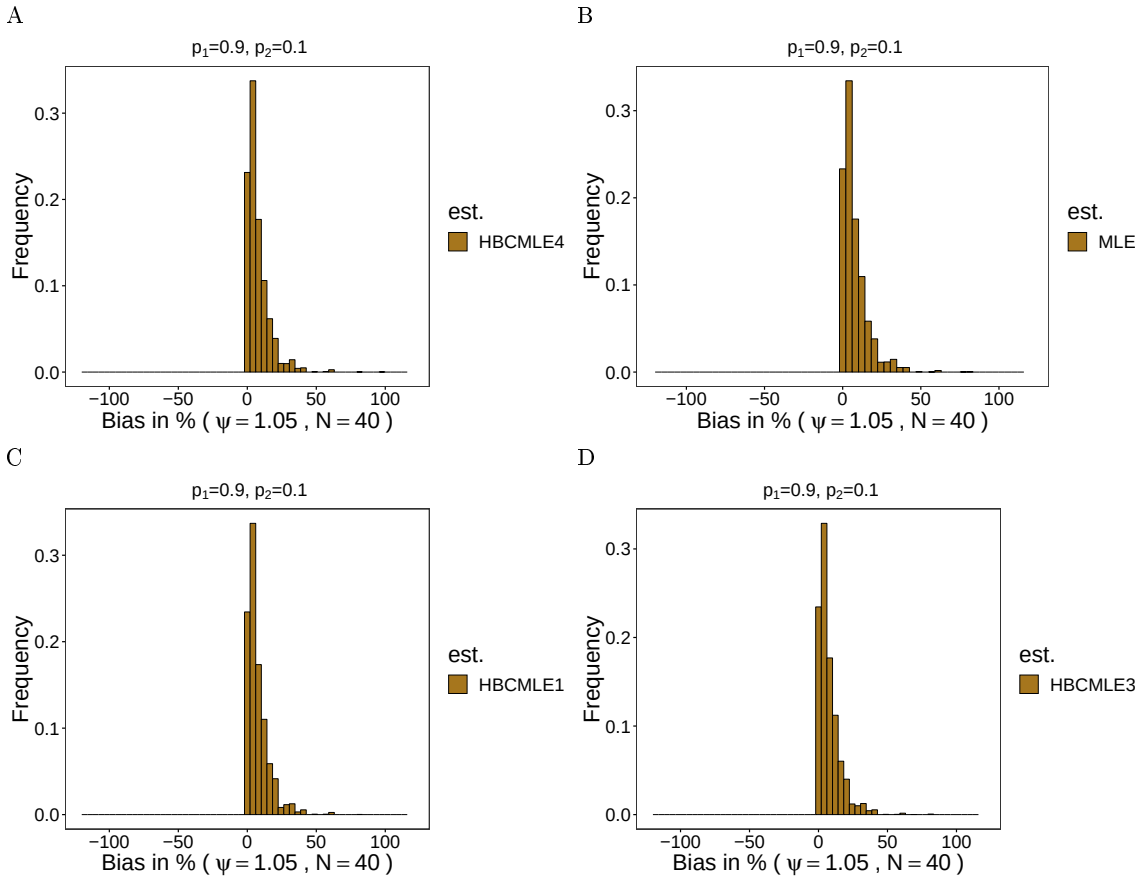


Figure 31: Same as Figure 27.

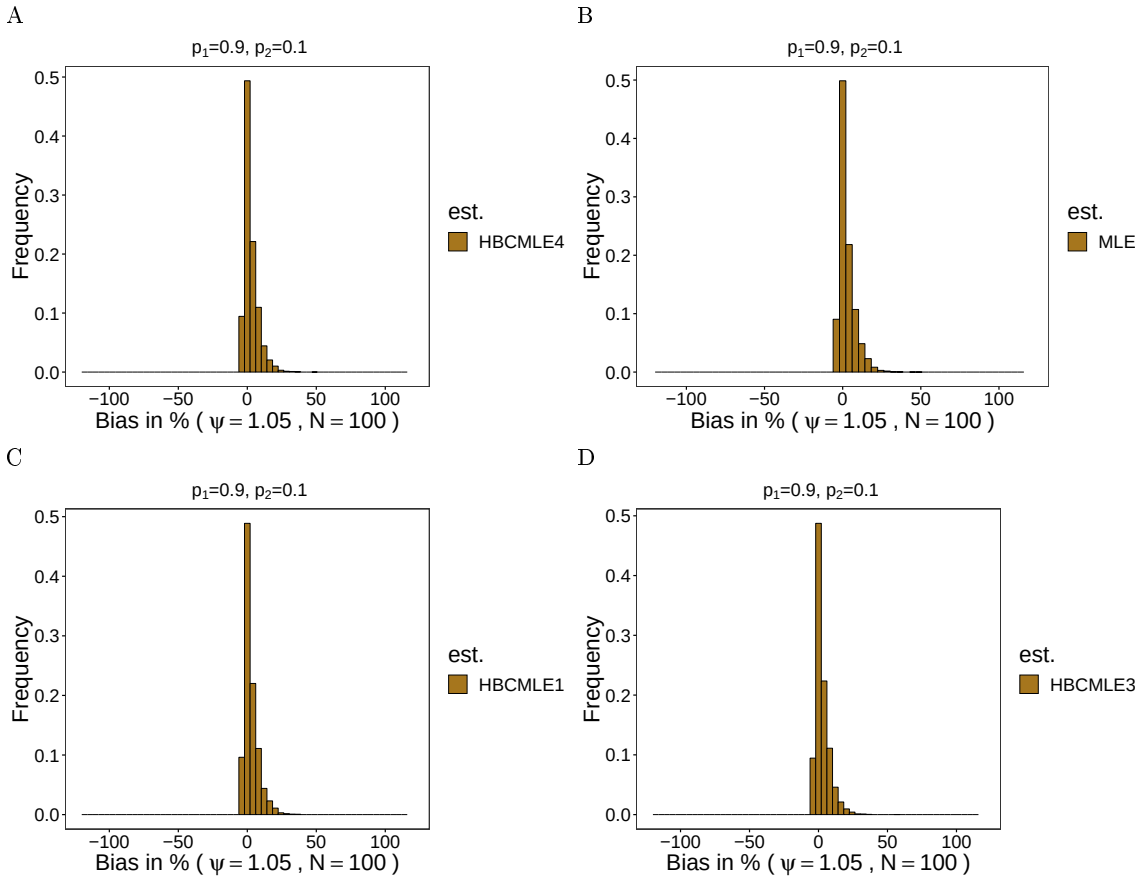


Figure 32: Same as Figure 27.

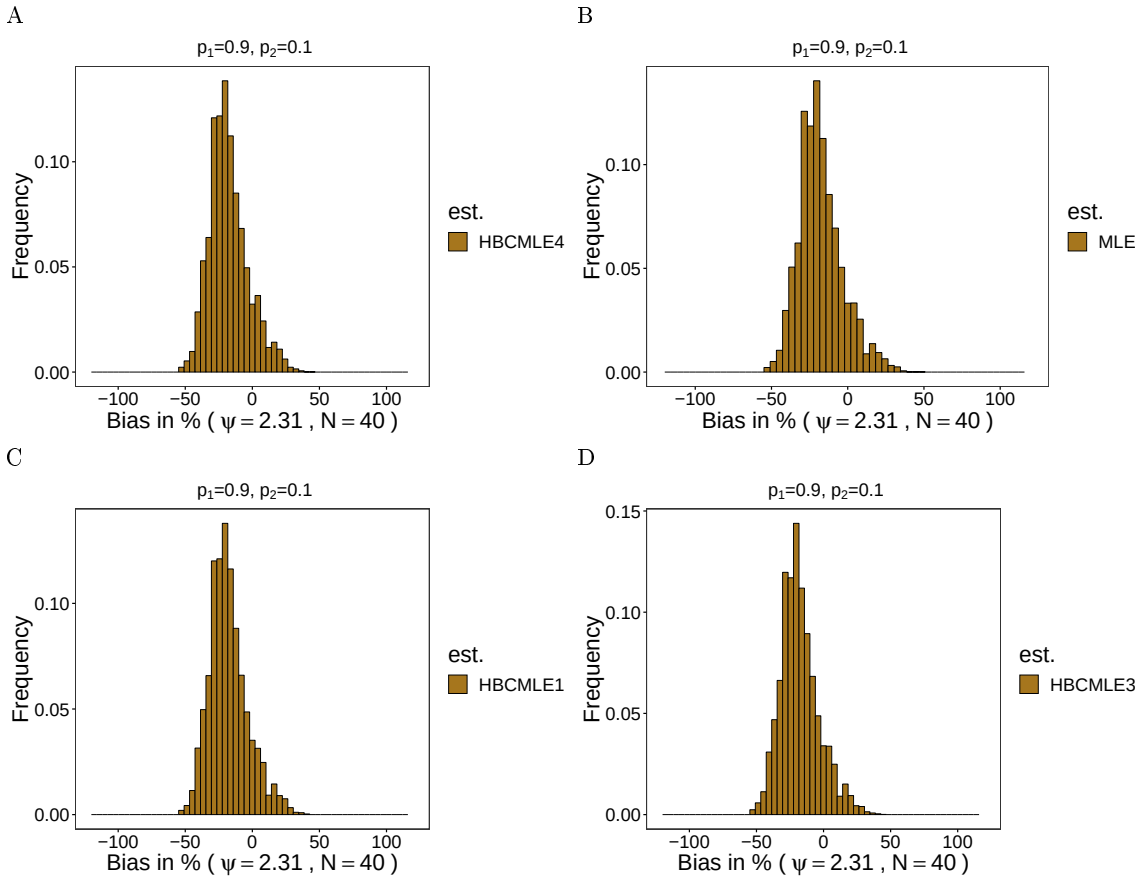


Figure 33: Same as Figure 27.

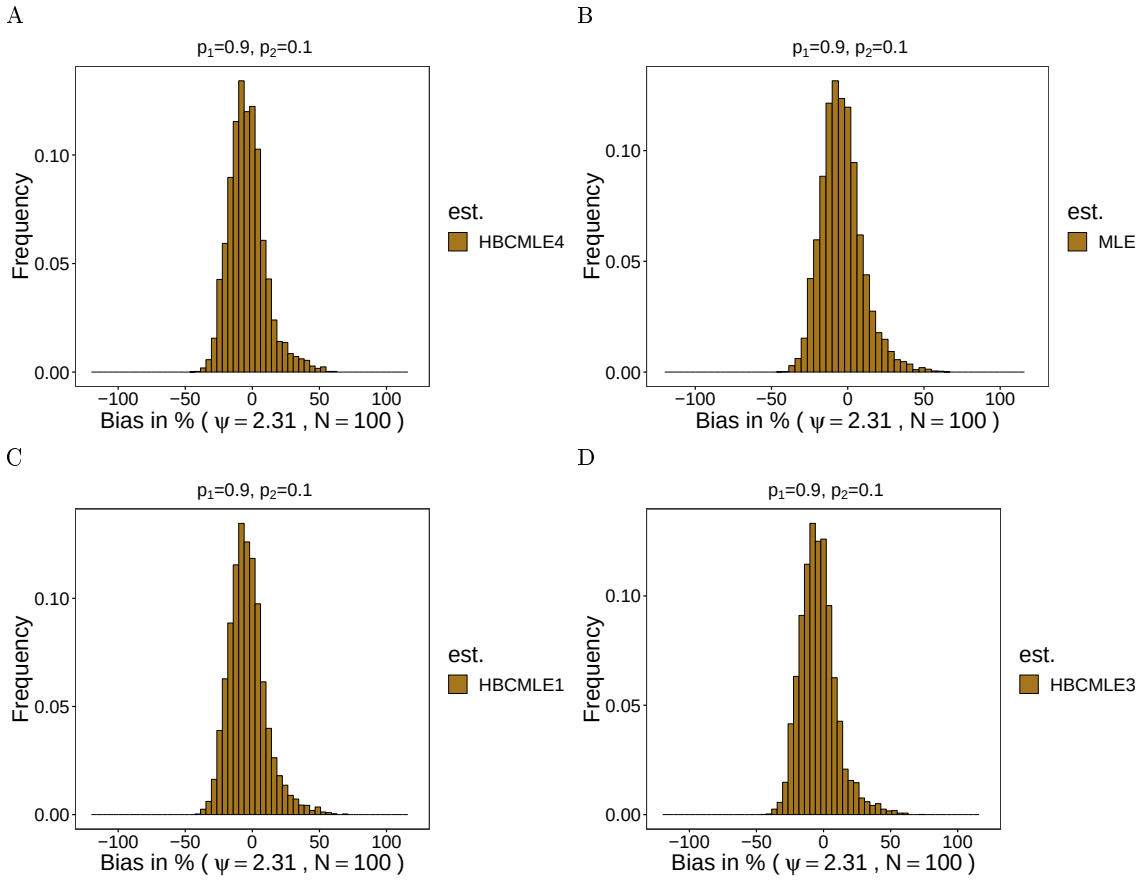


Figure 34: Same as Figure 27.

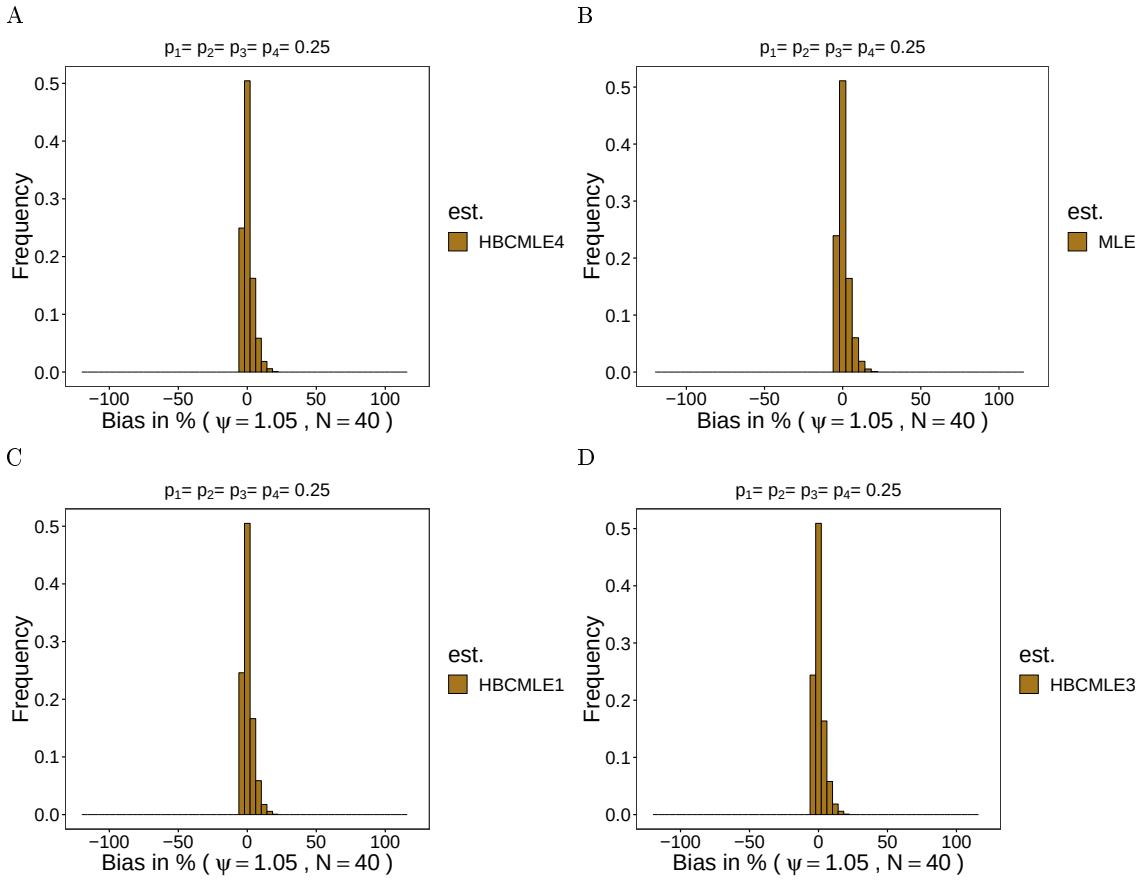


Figure 35: Same as Figure 27.

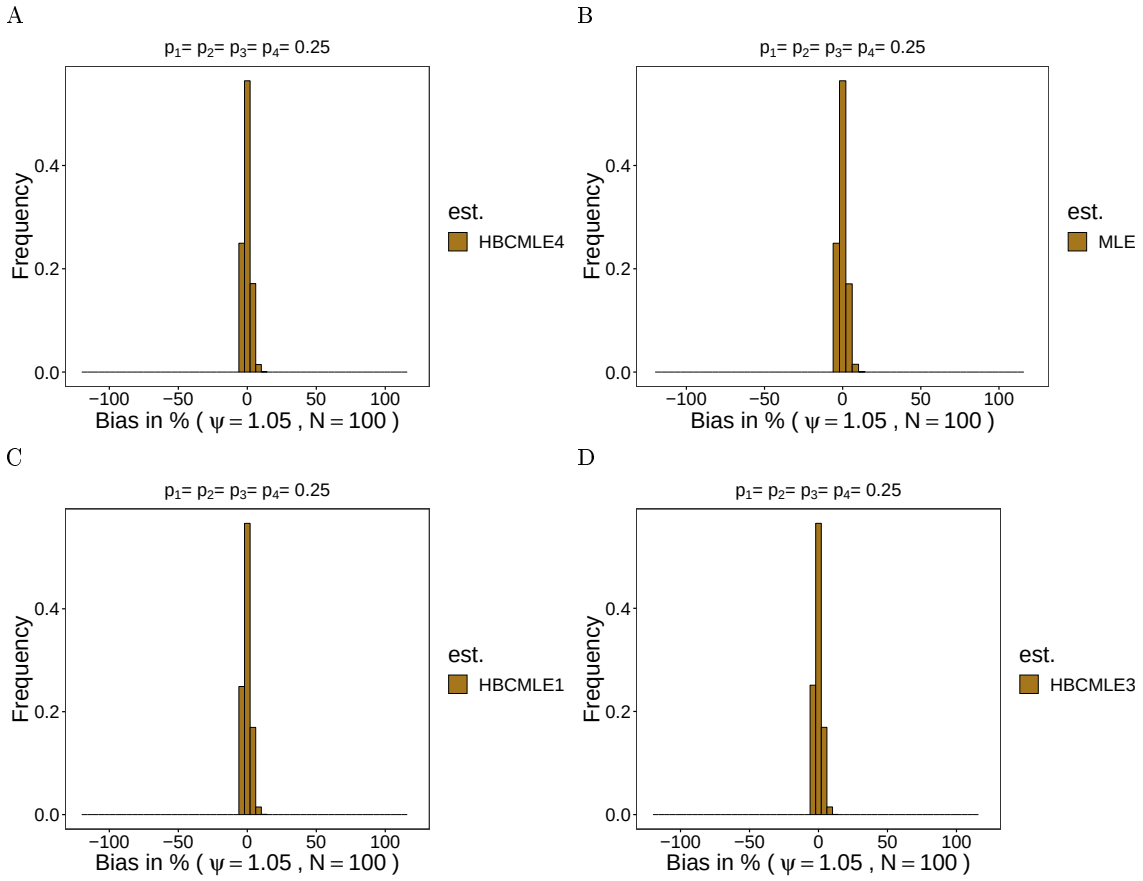


Figure 36: Same as Figure 27.

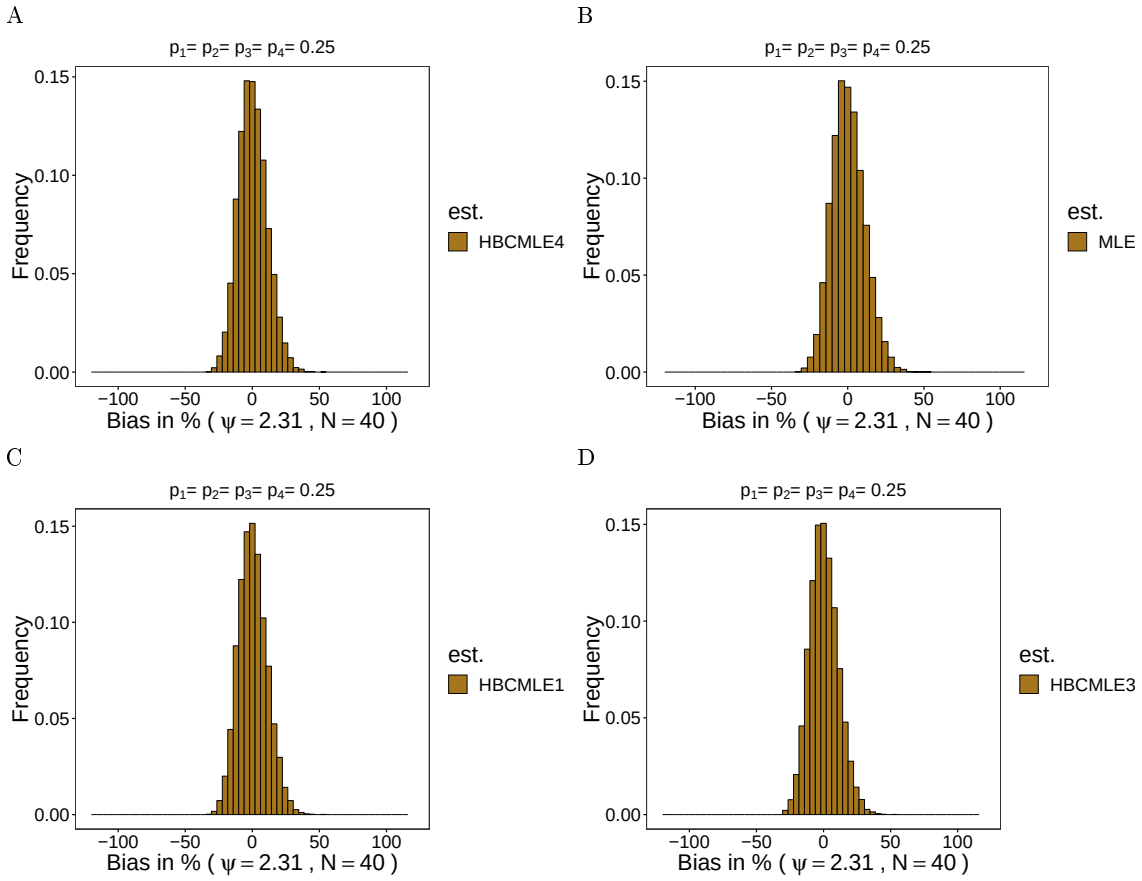


Figure 37: Same as Figure 27.

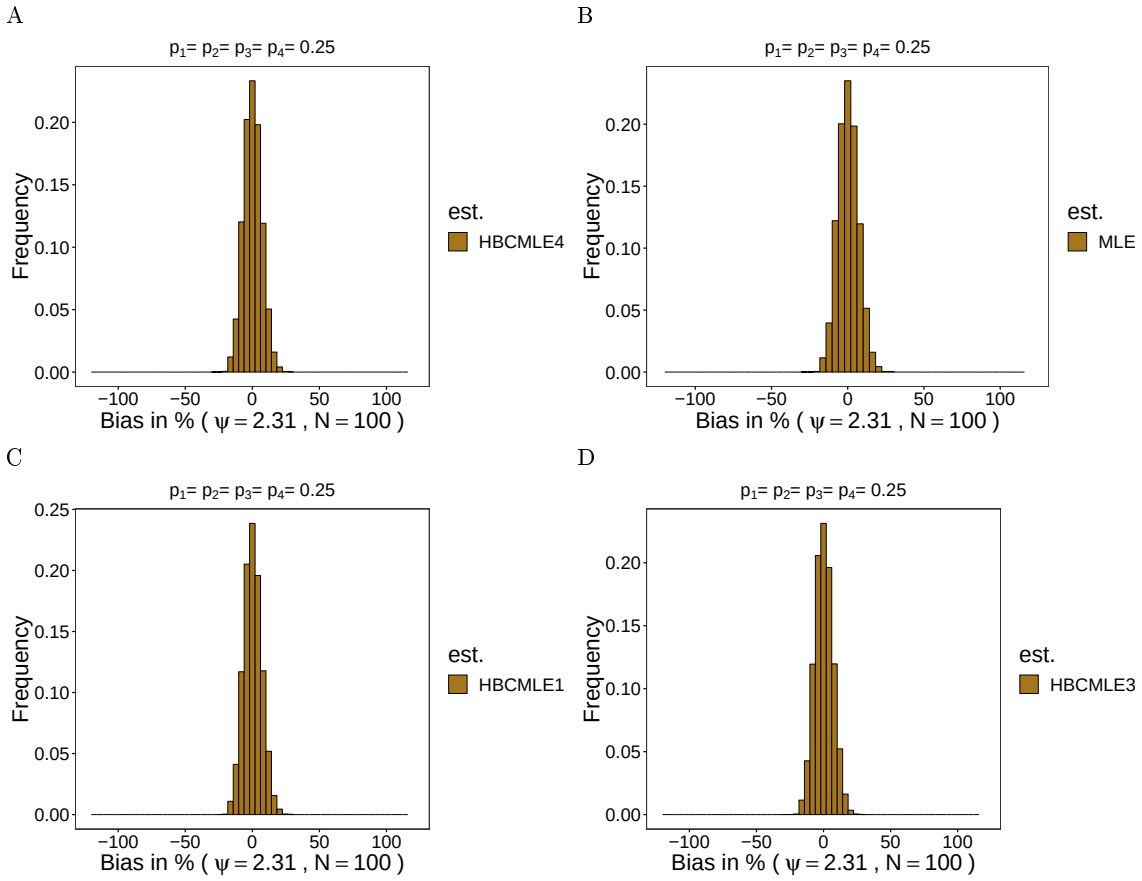


Figure 38: Same as Figure 27.

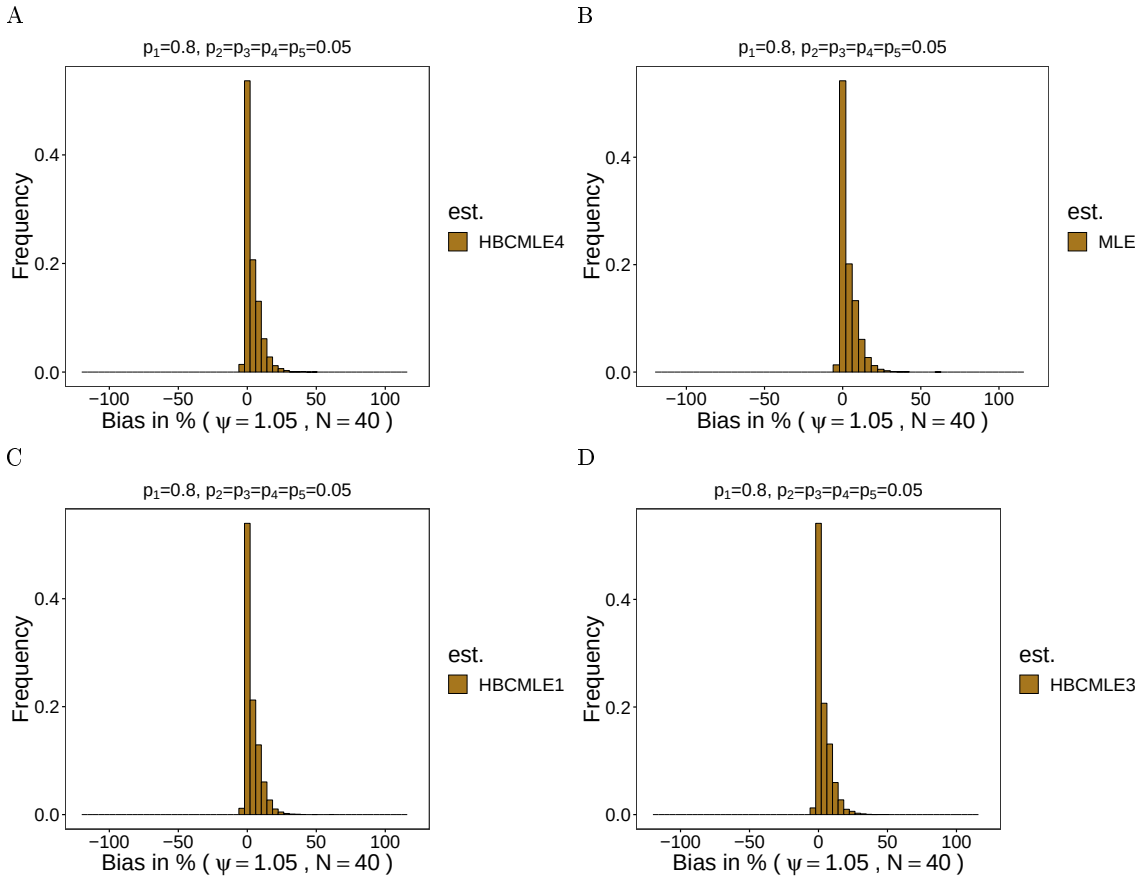


Figure 39: Same as Figure 27.

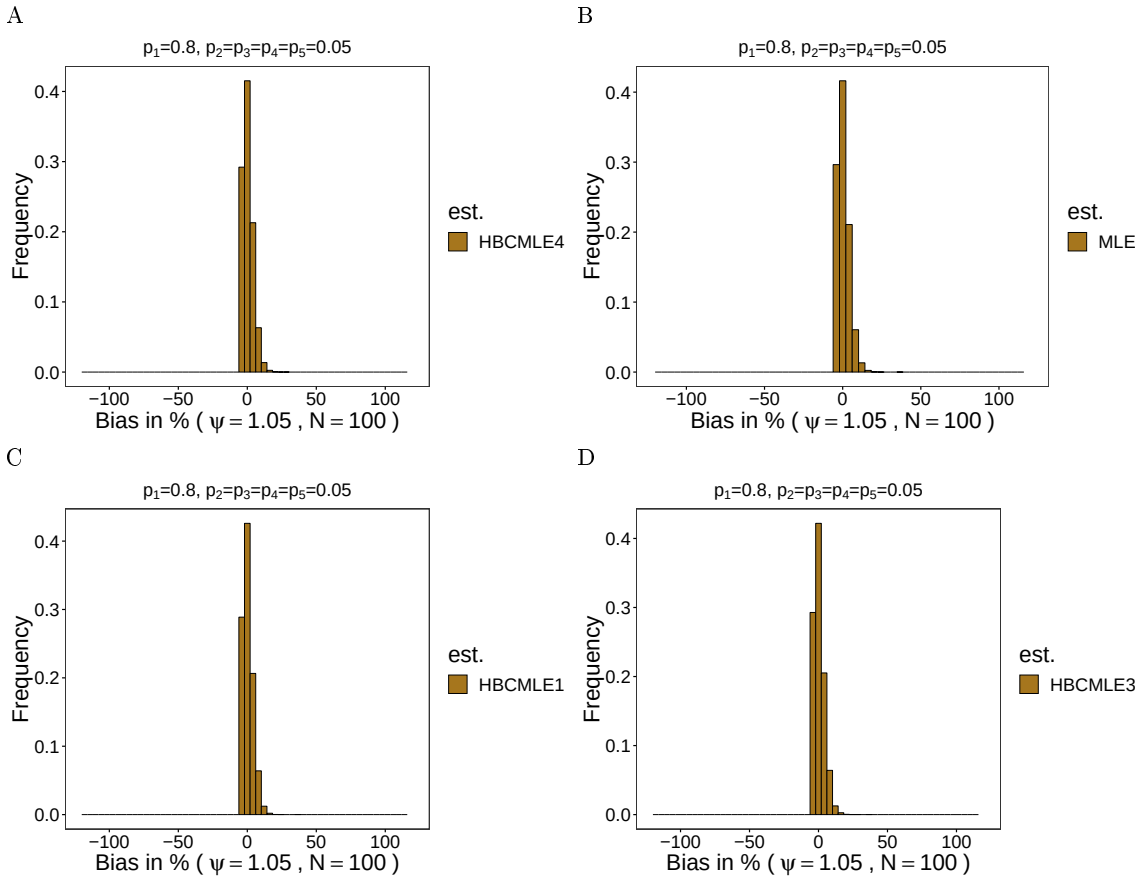


Figure 40: Same as Figure 27.

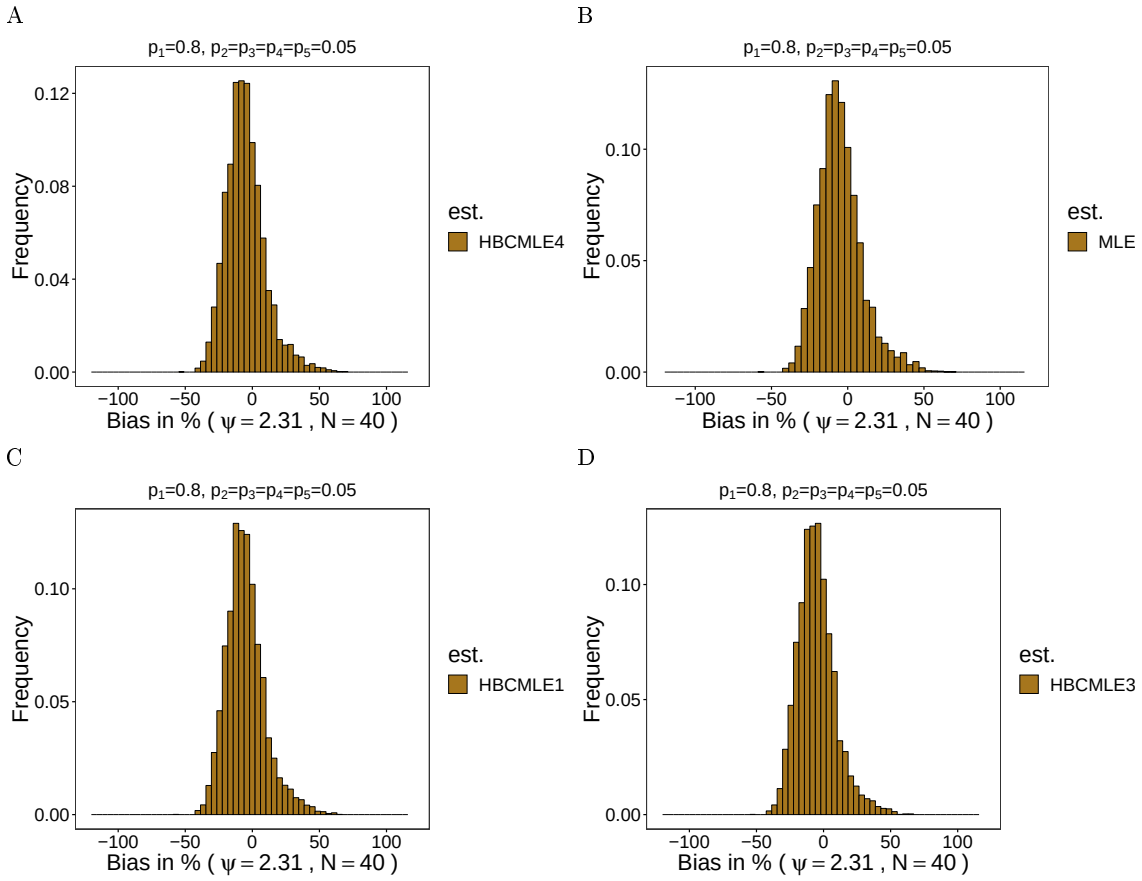


Figure 41: Same as Figure 27.

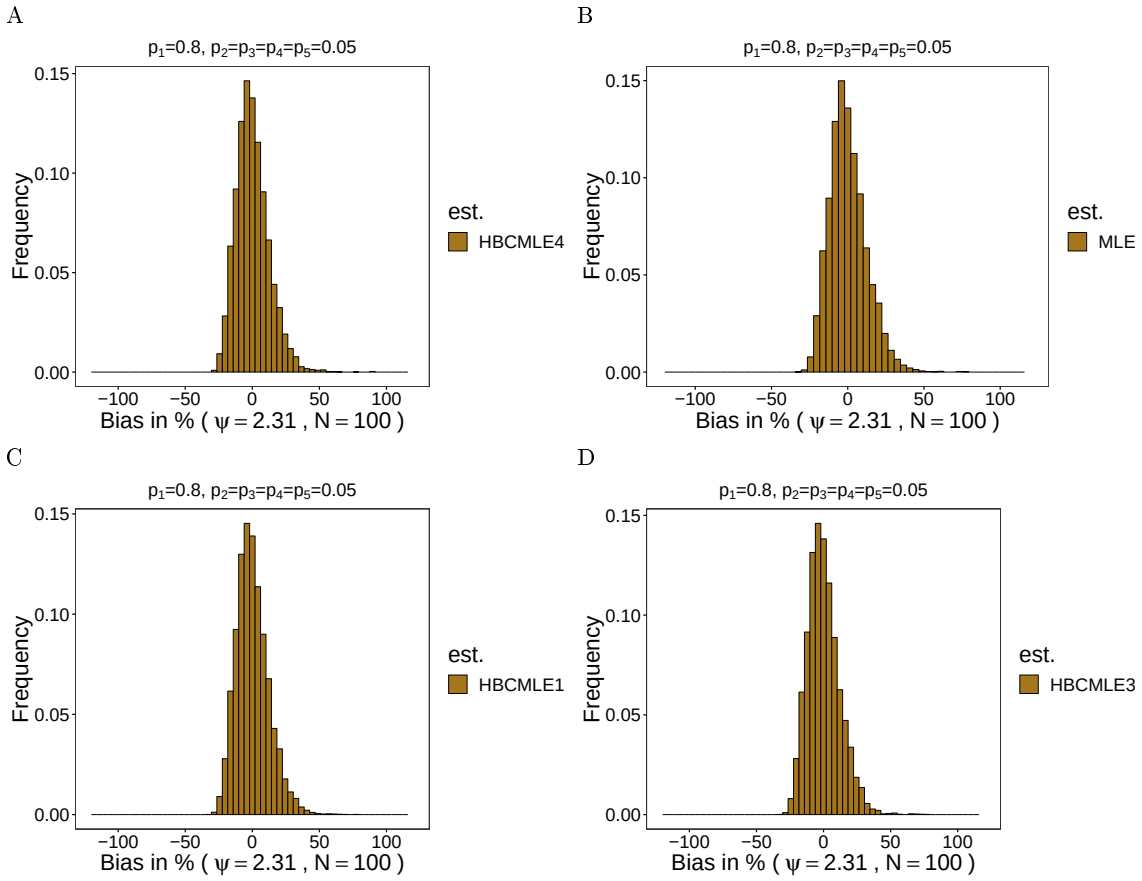


Figure 42: Same as Figure 27.

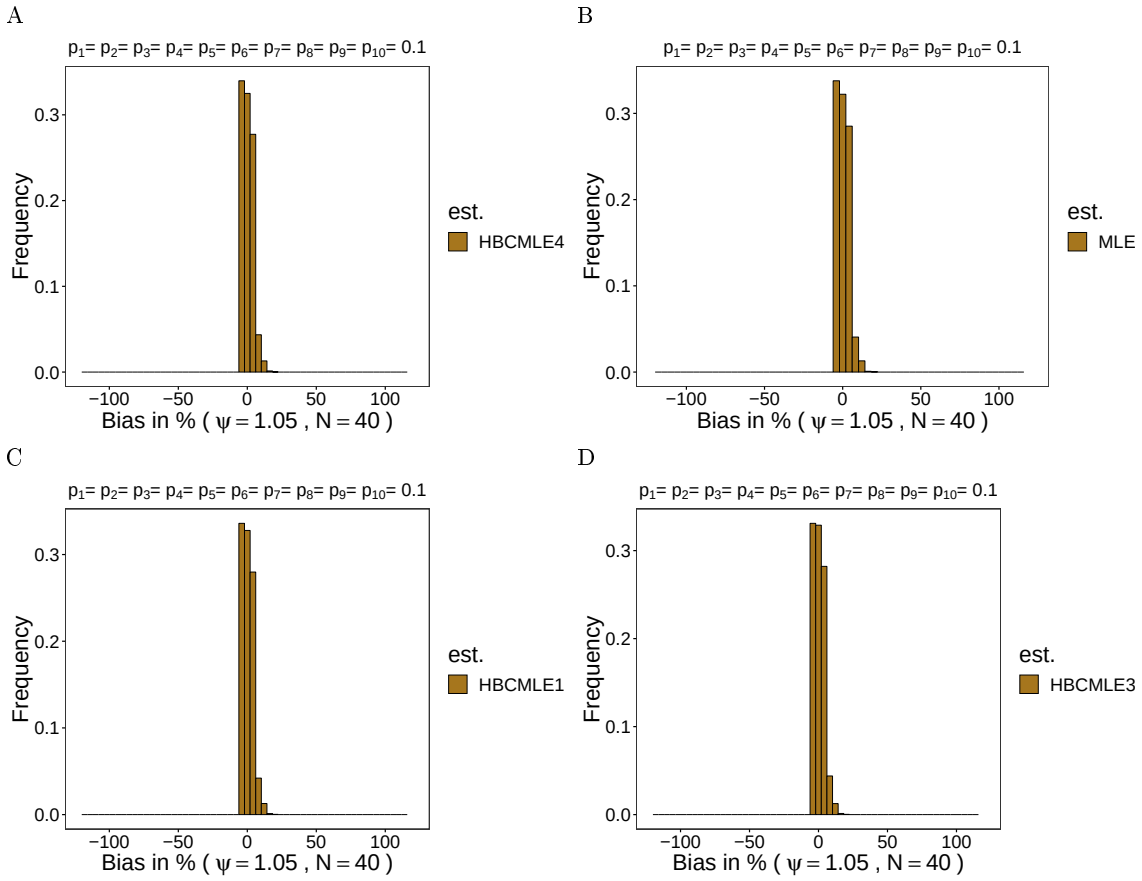


Figure 43: Same as Figure 27.

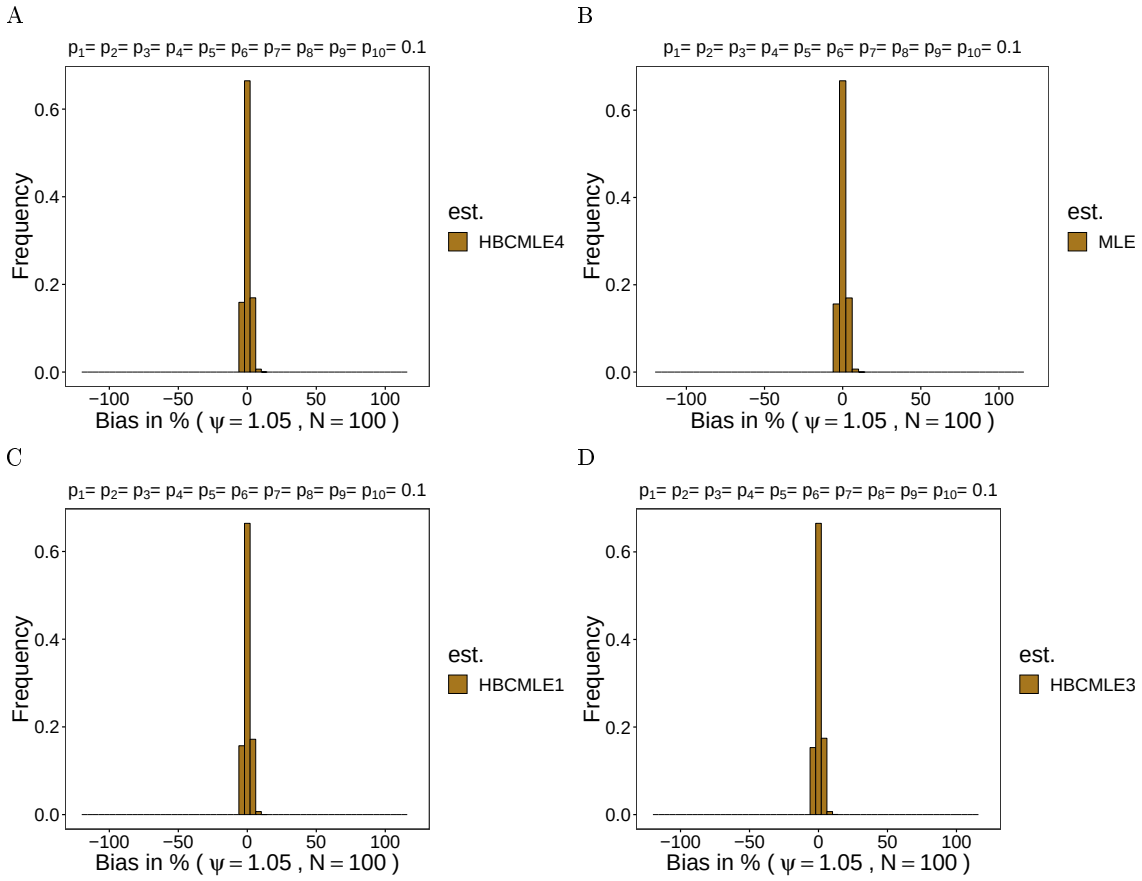


Figure 44: Same as Figure 27.

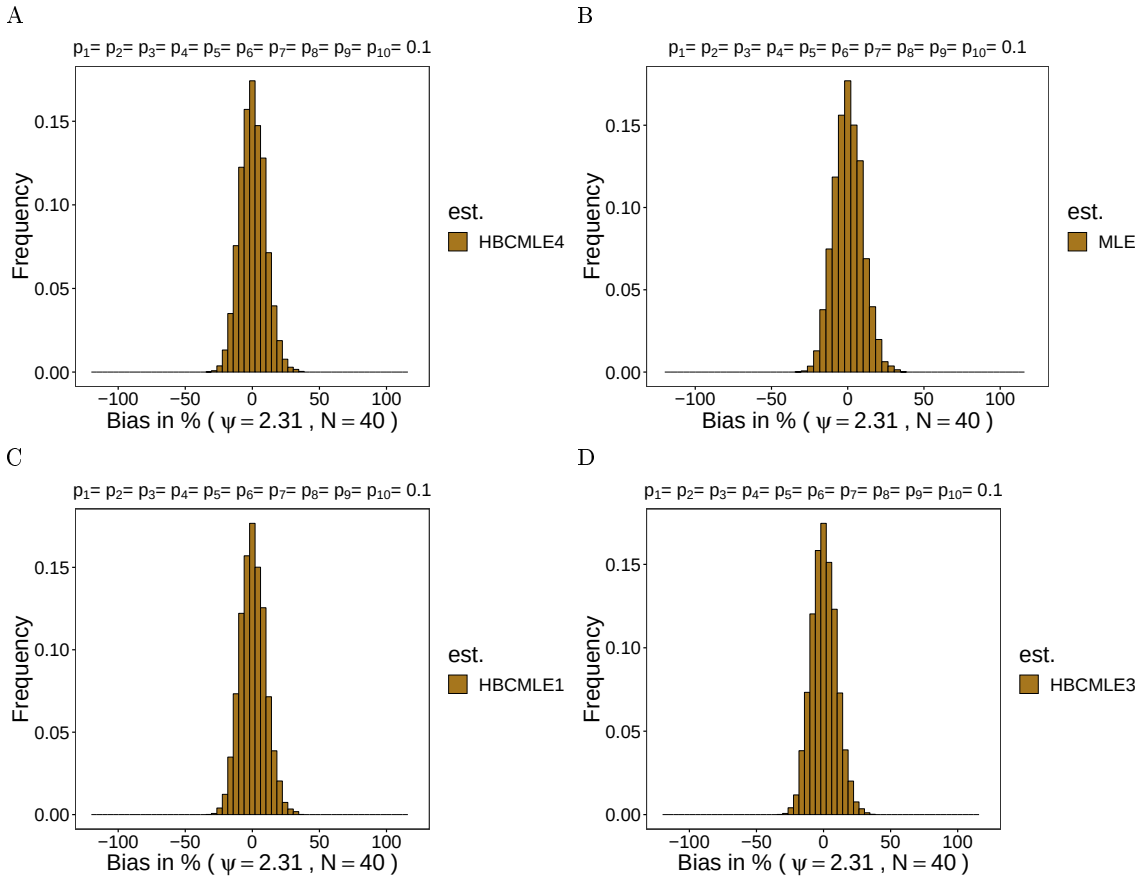


Figure 45: Same as Figure 27.

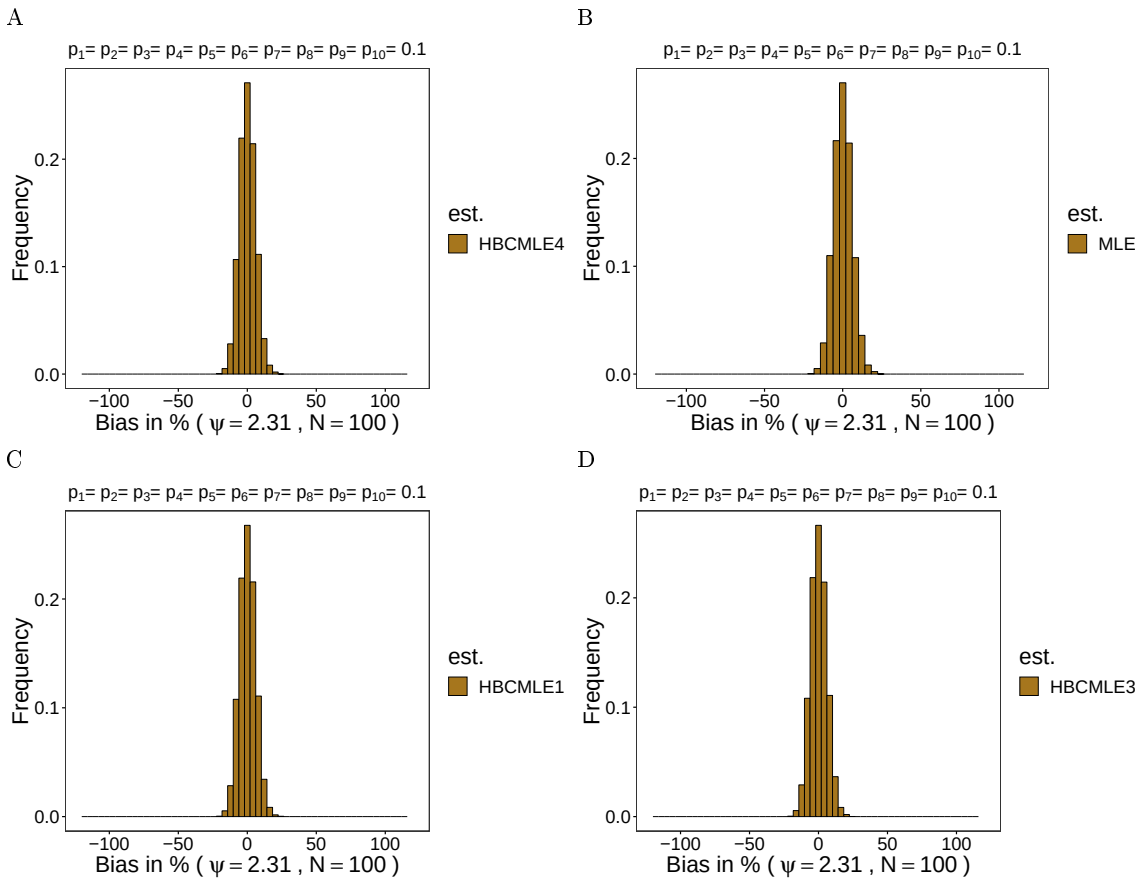


Figure 46: Same as Figure 27.

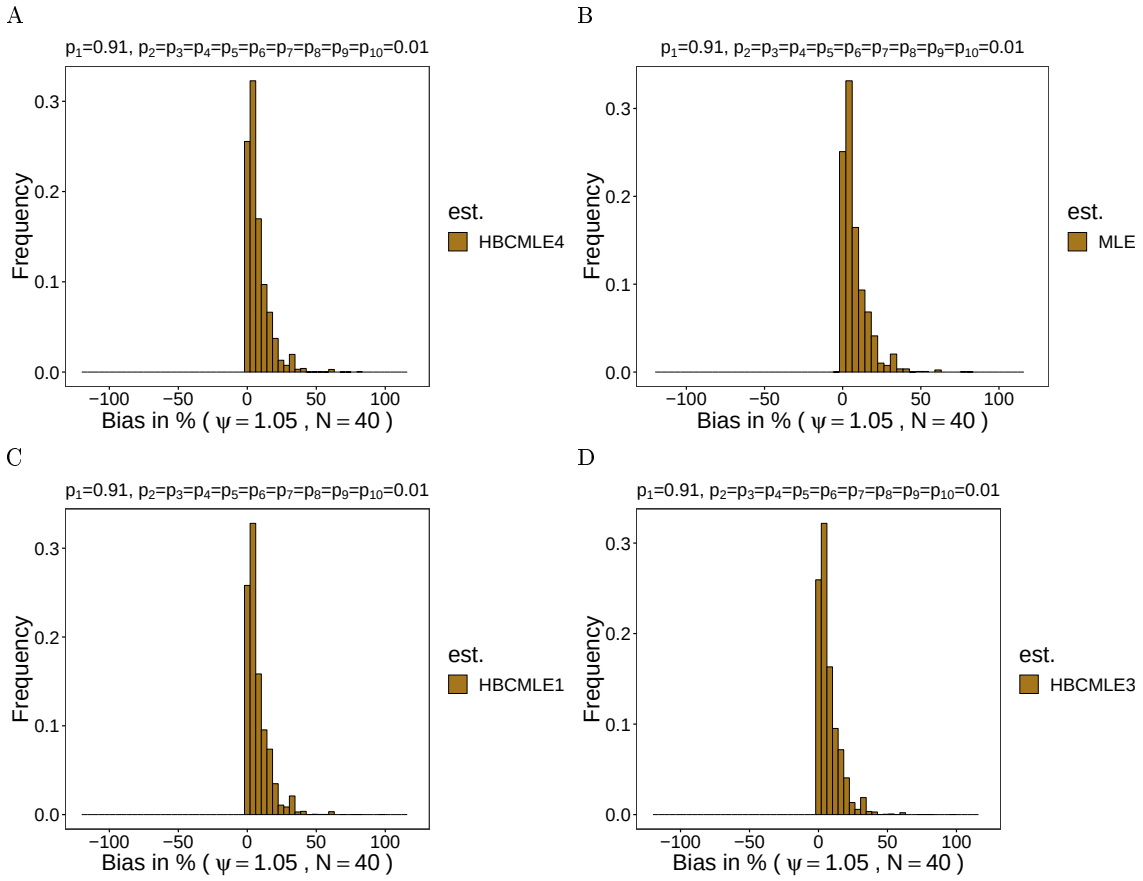


Figure 47: Same as Figure 27.

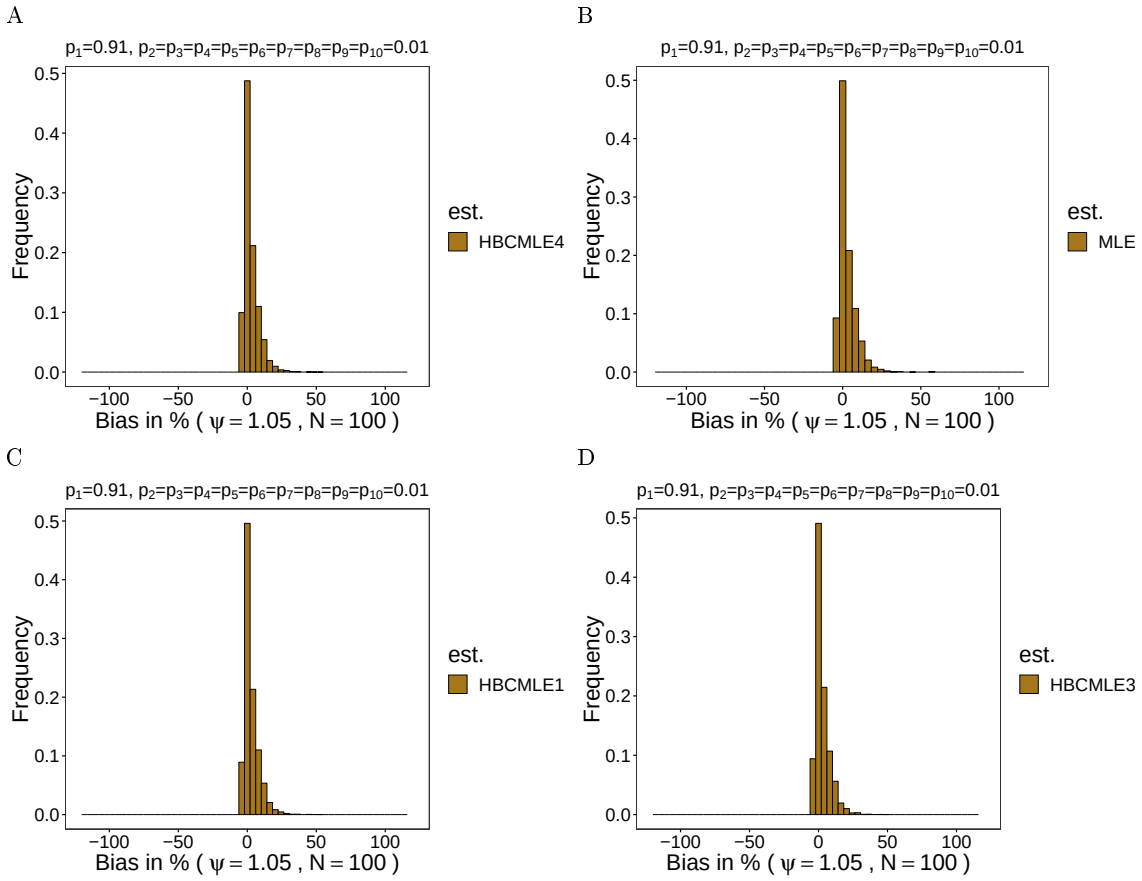


Figure 48: Same as Figure 27.

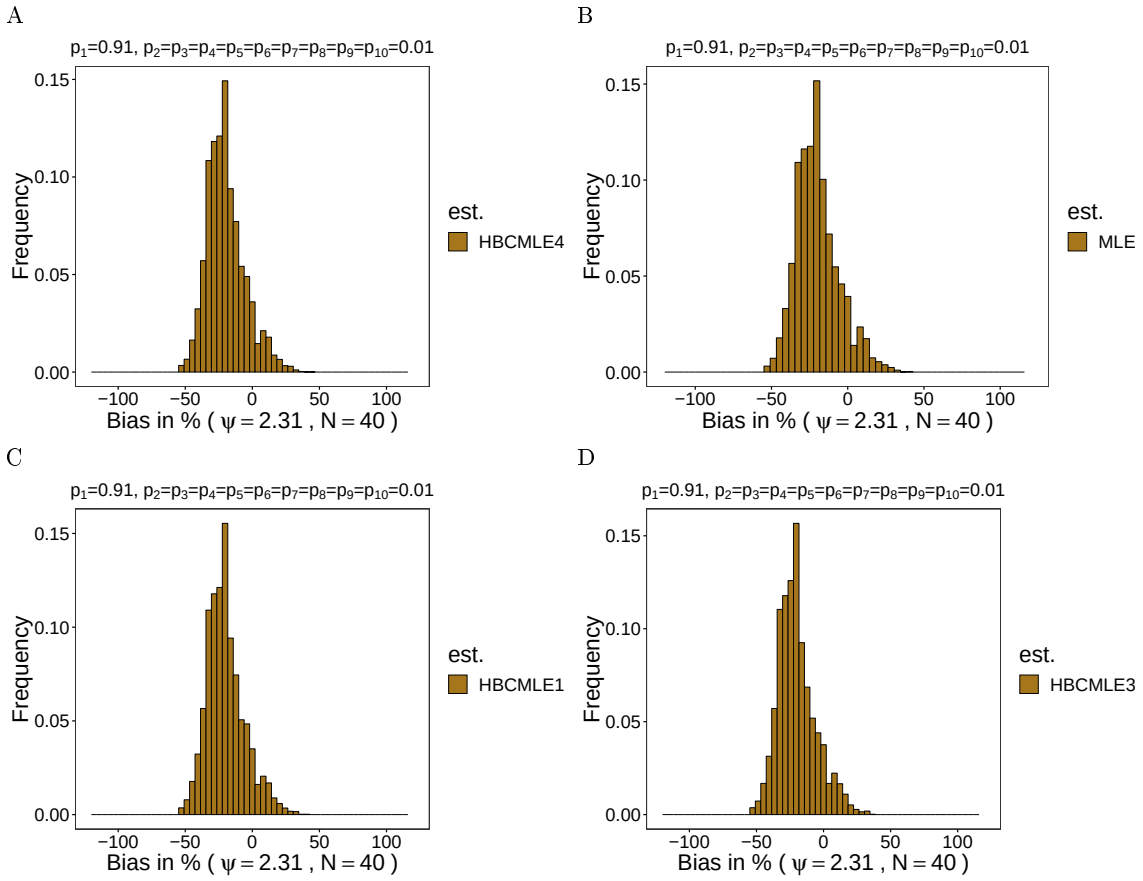


Figure 49: Same as Figure 27.

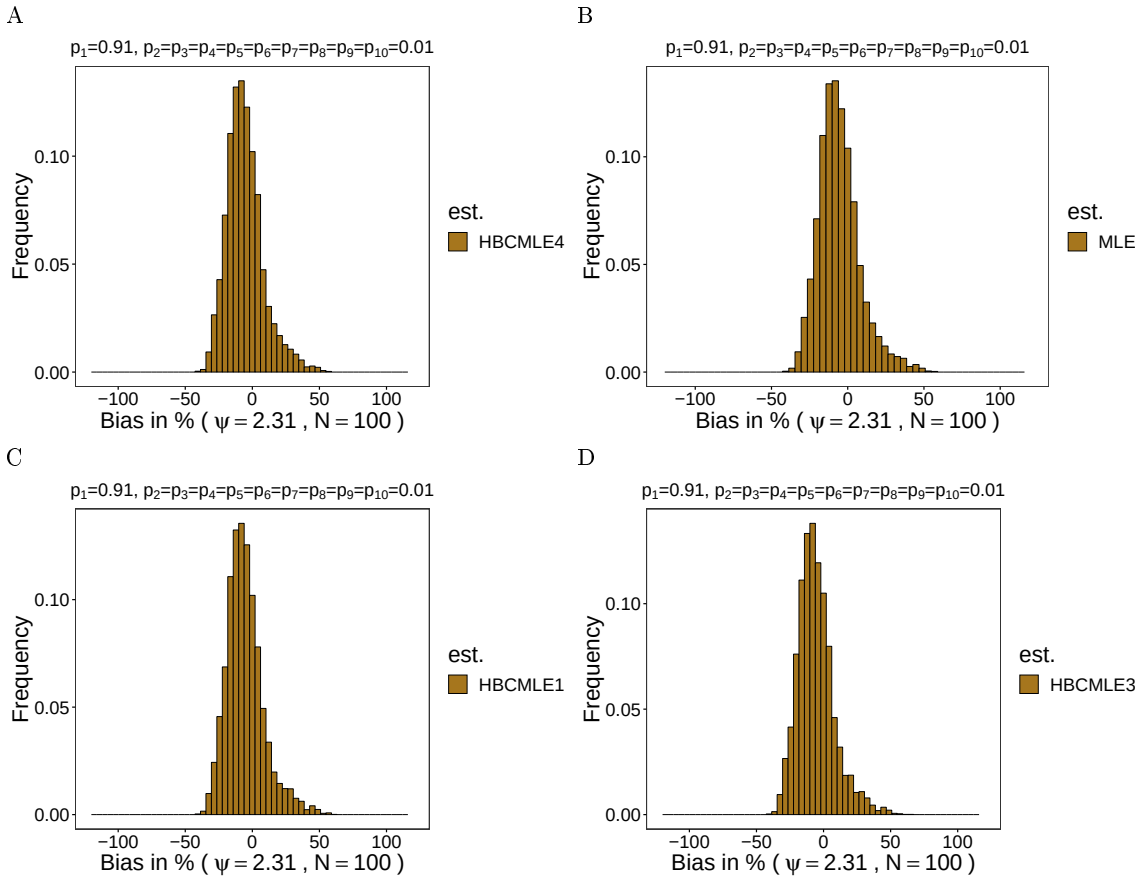


Figure 50: Same as Figure 27.

6 Boxplots

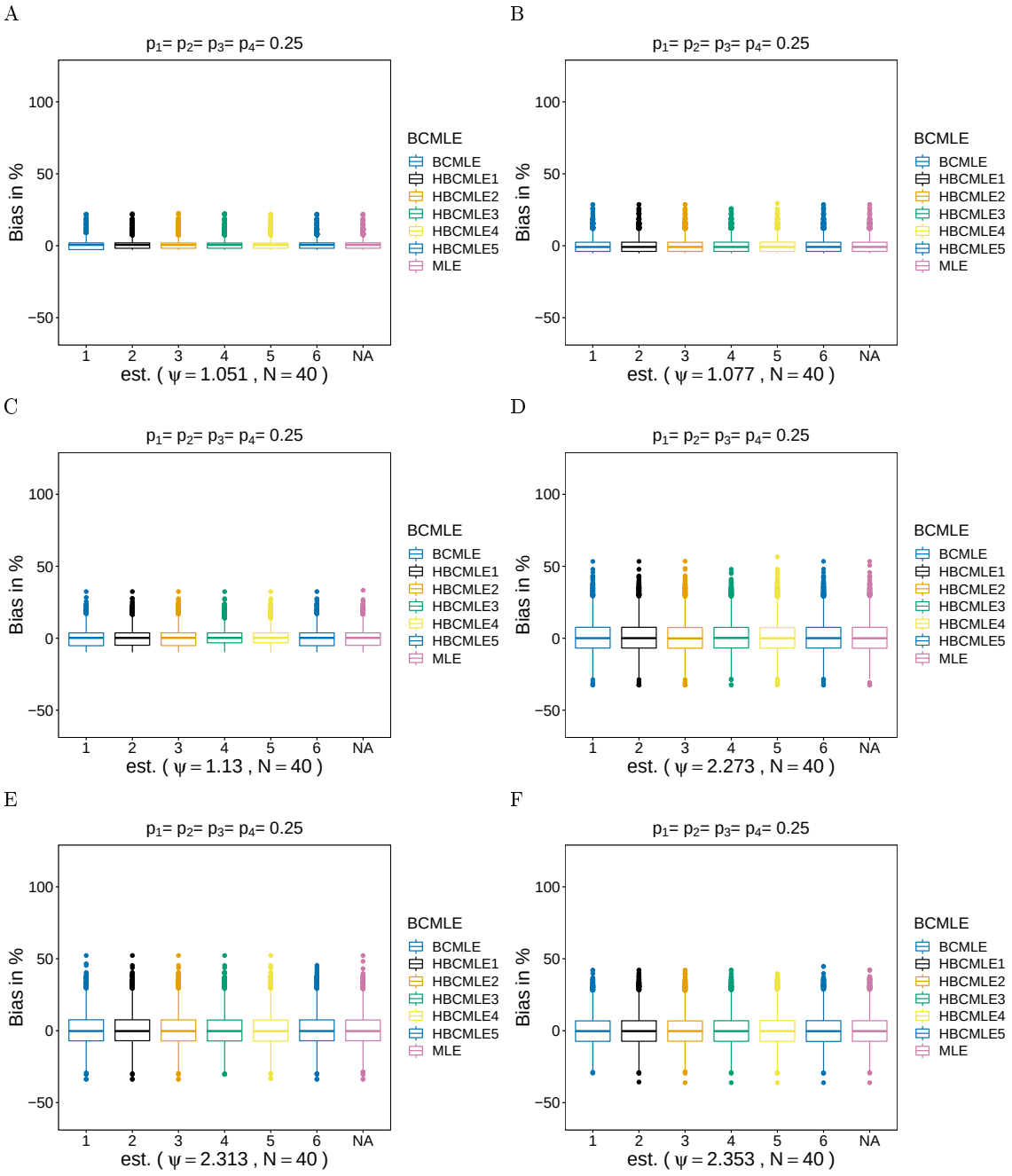


Figure 51: **Boxplots**. The figure shows boxplots of 10,000 estimations of ψ for a specific sample size N. Colors correspond to different estimators. Each plot corresponds to a different value of true parameter ψ.

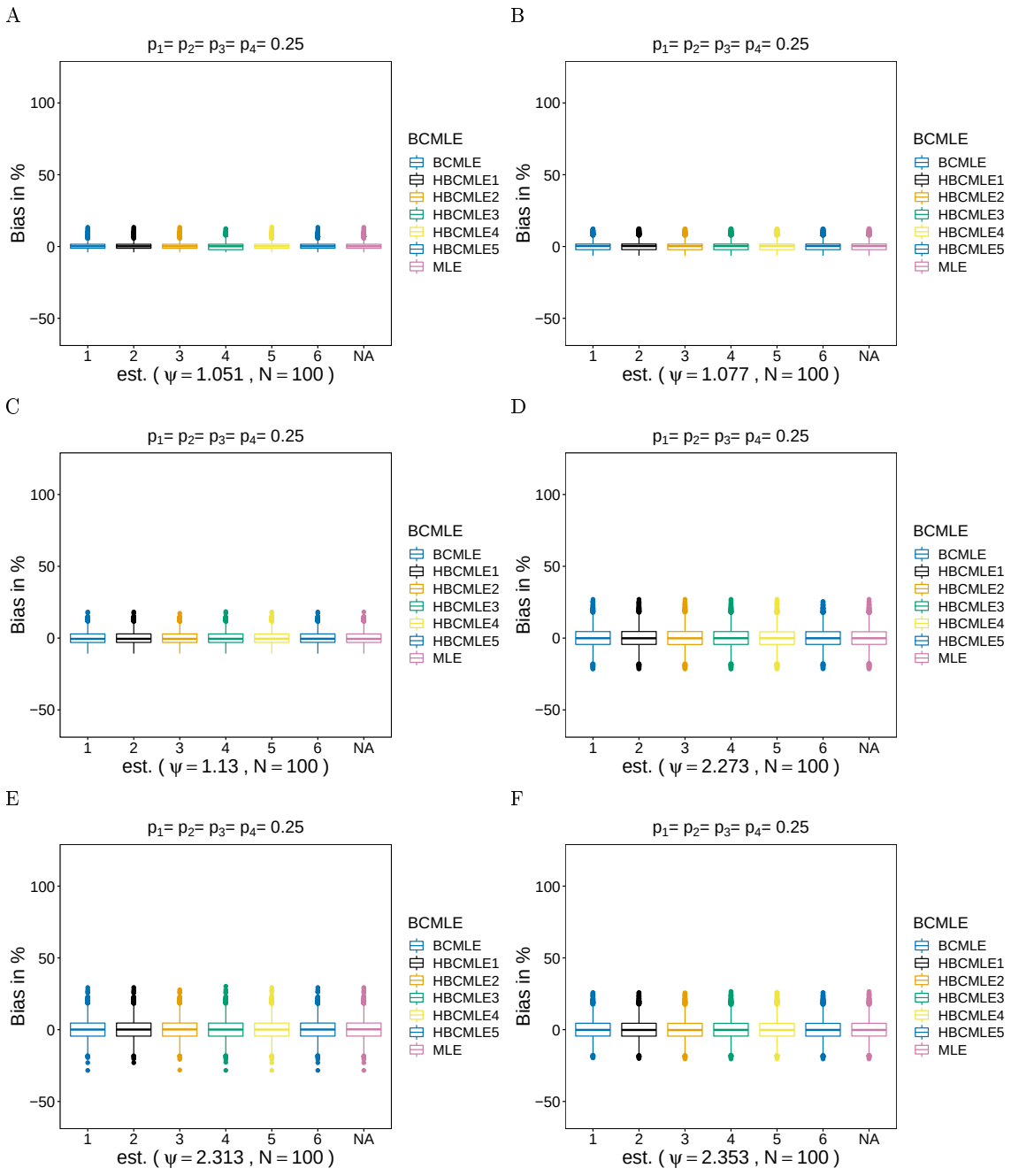


Figure 52: Same as Figure 51

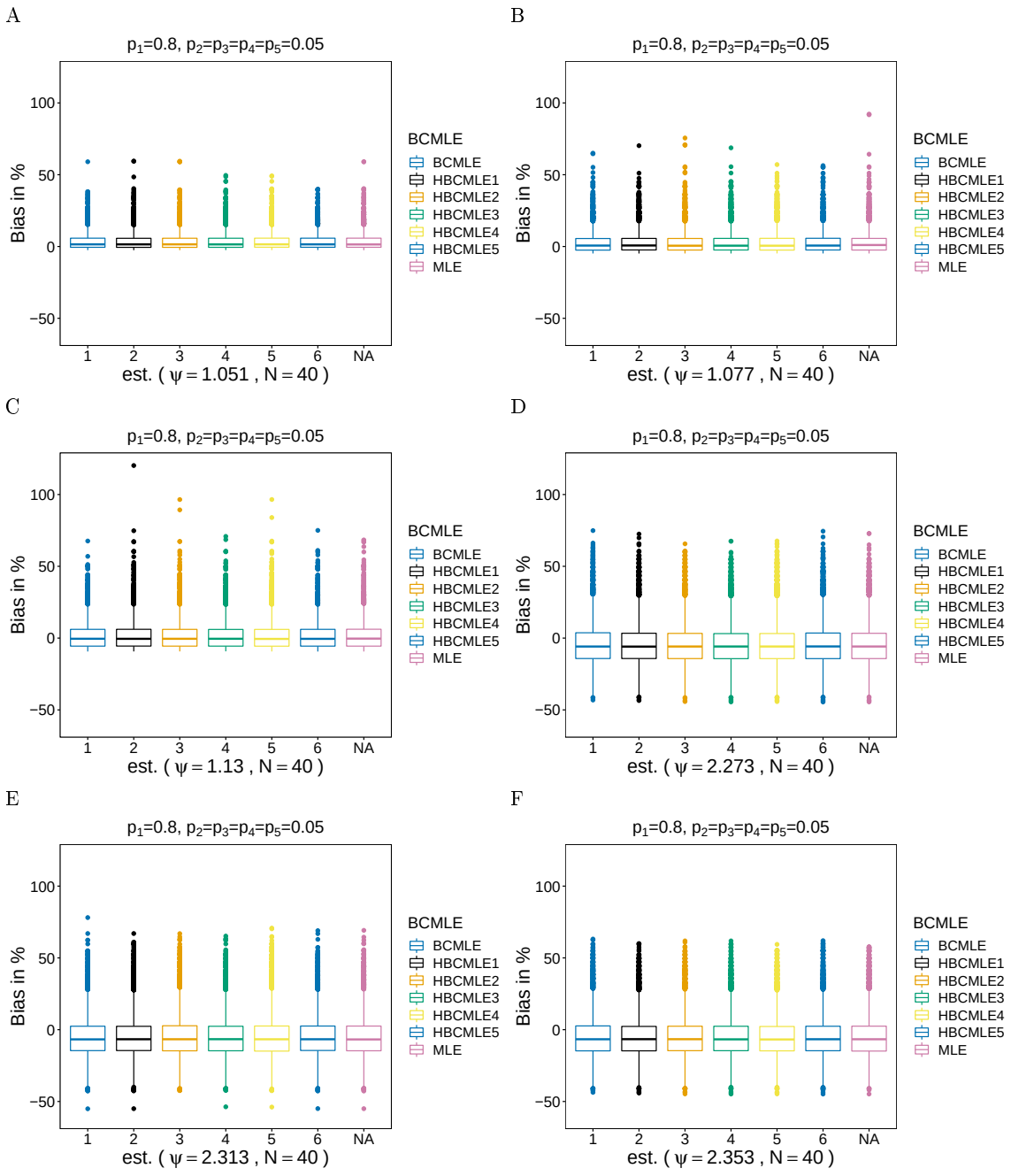


Figure 53: Same as Figure 51

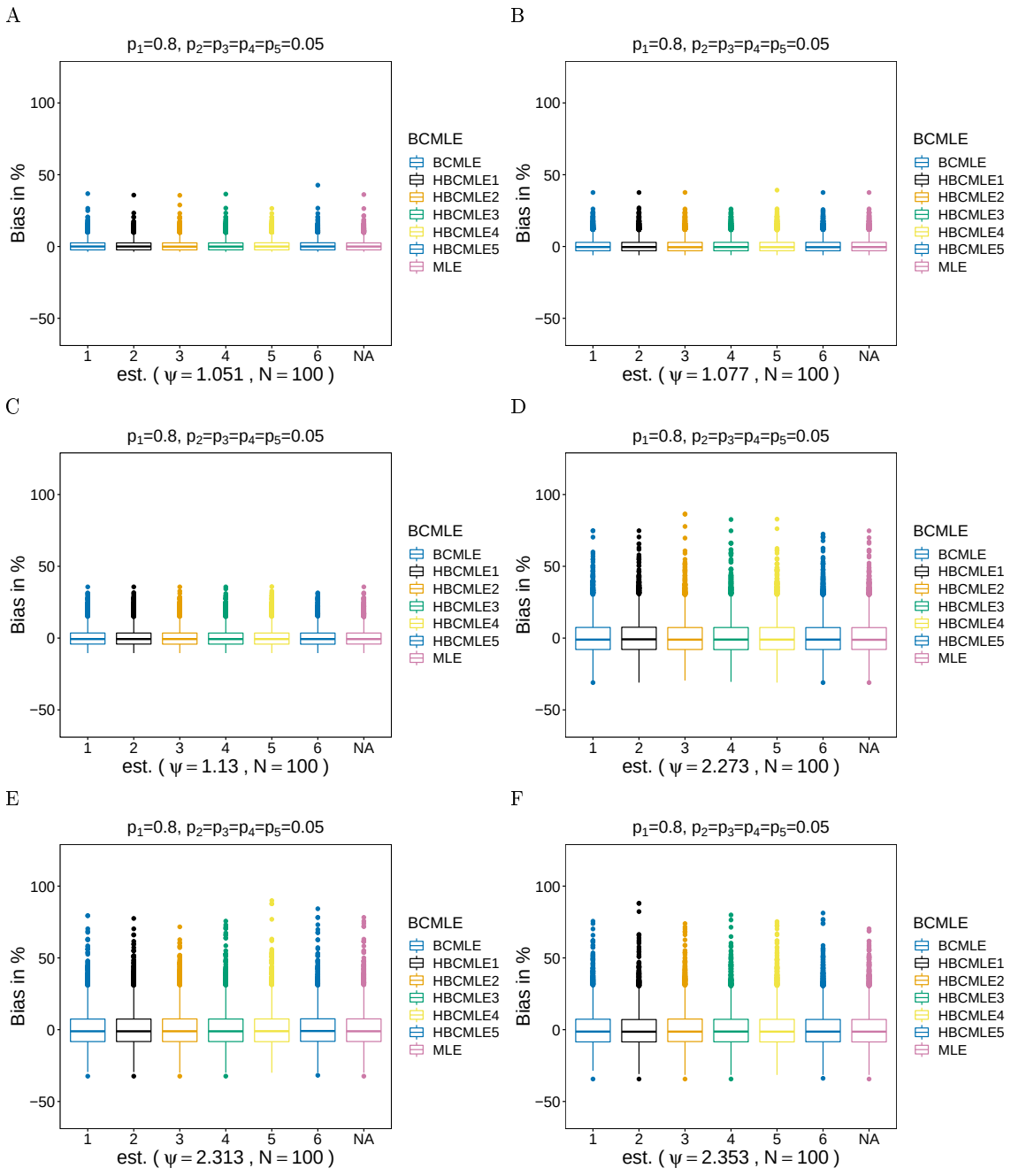


Figure 54: Same as Figure 51

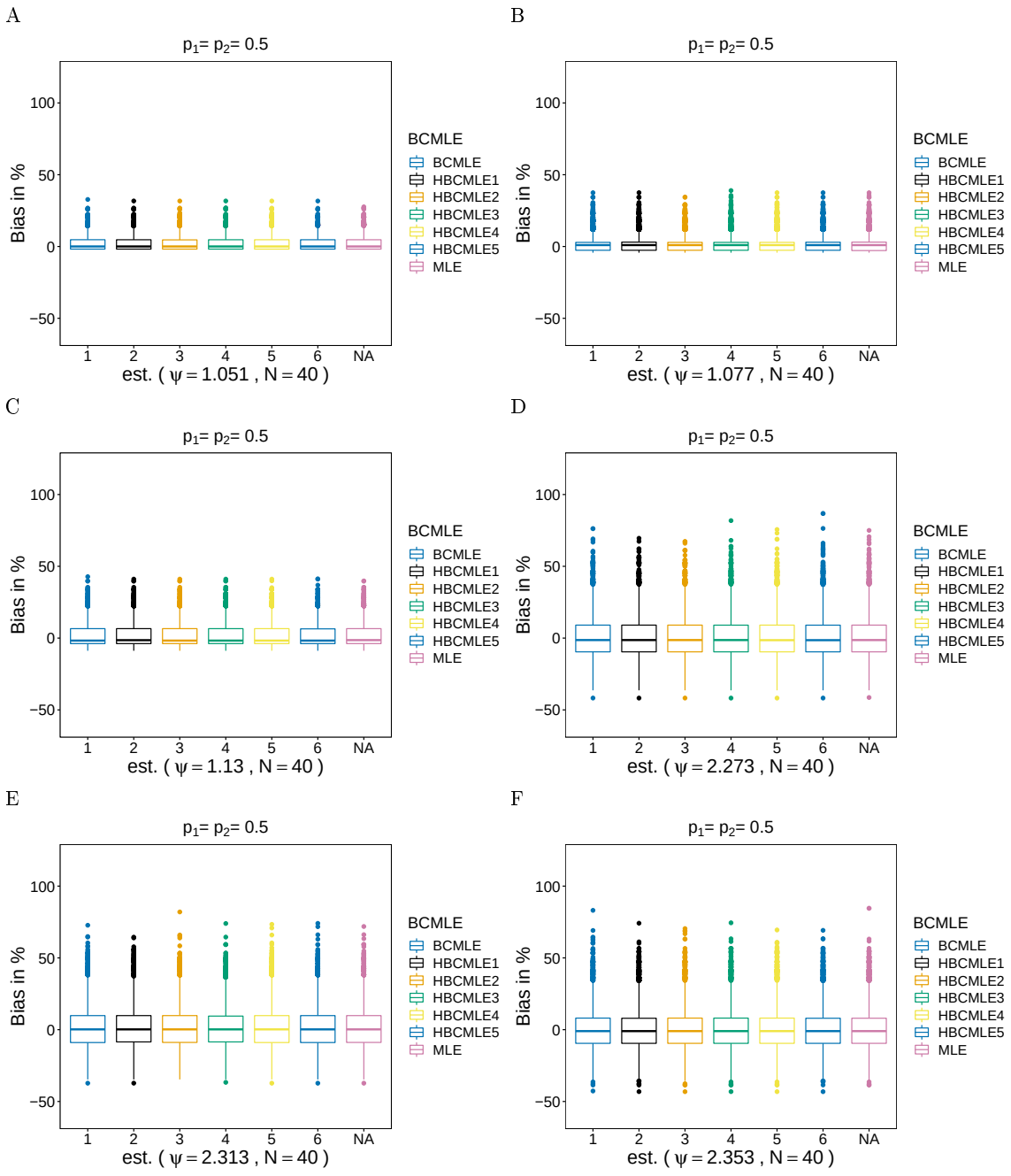


Figure 55: Same as Figure 51

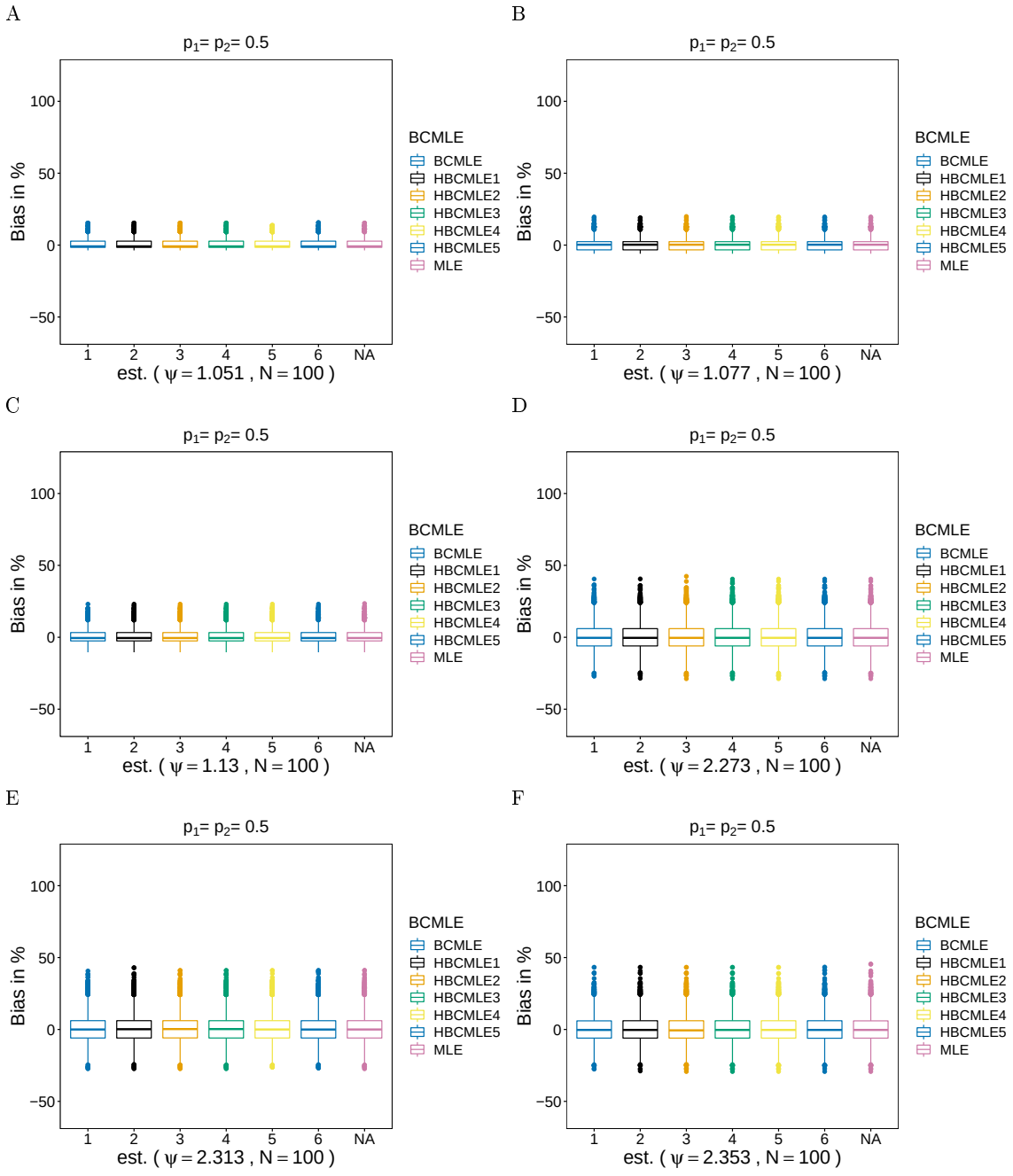


Figure 56: Same as Figure 51

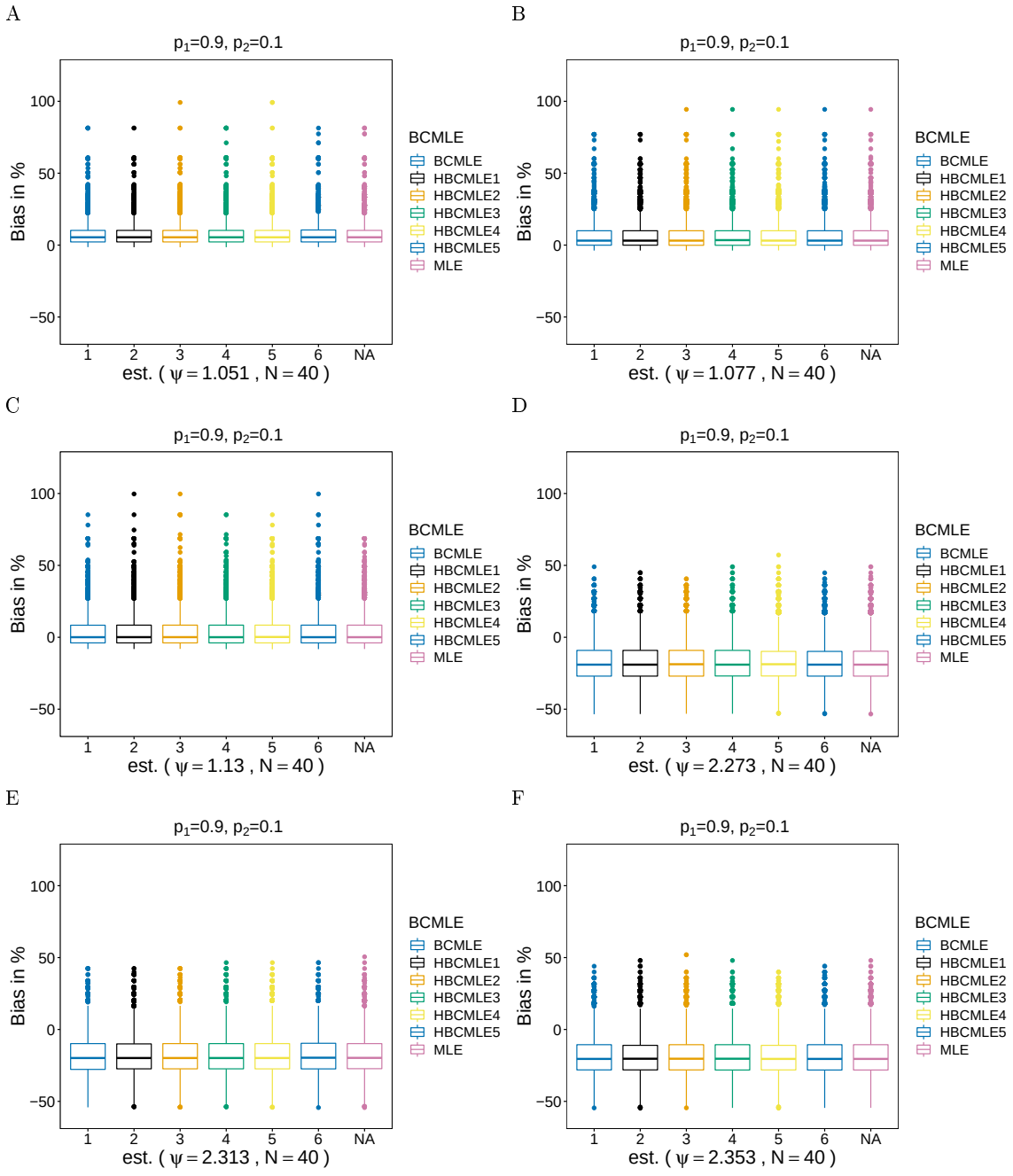


Figure 57: Same as Figure 51

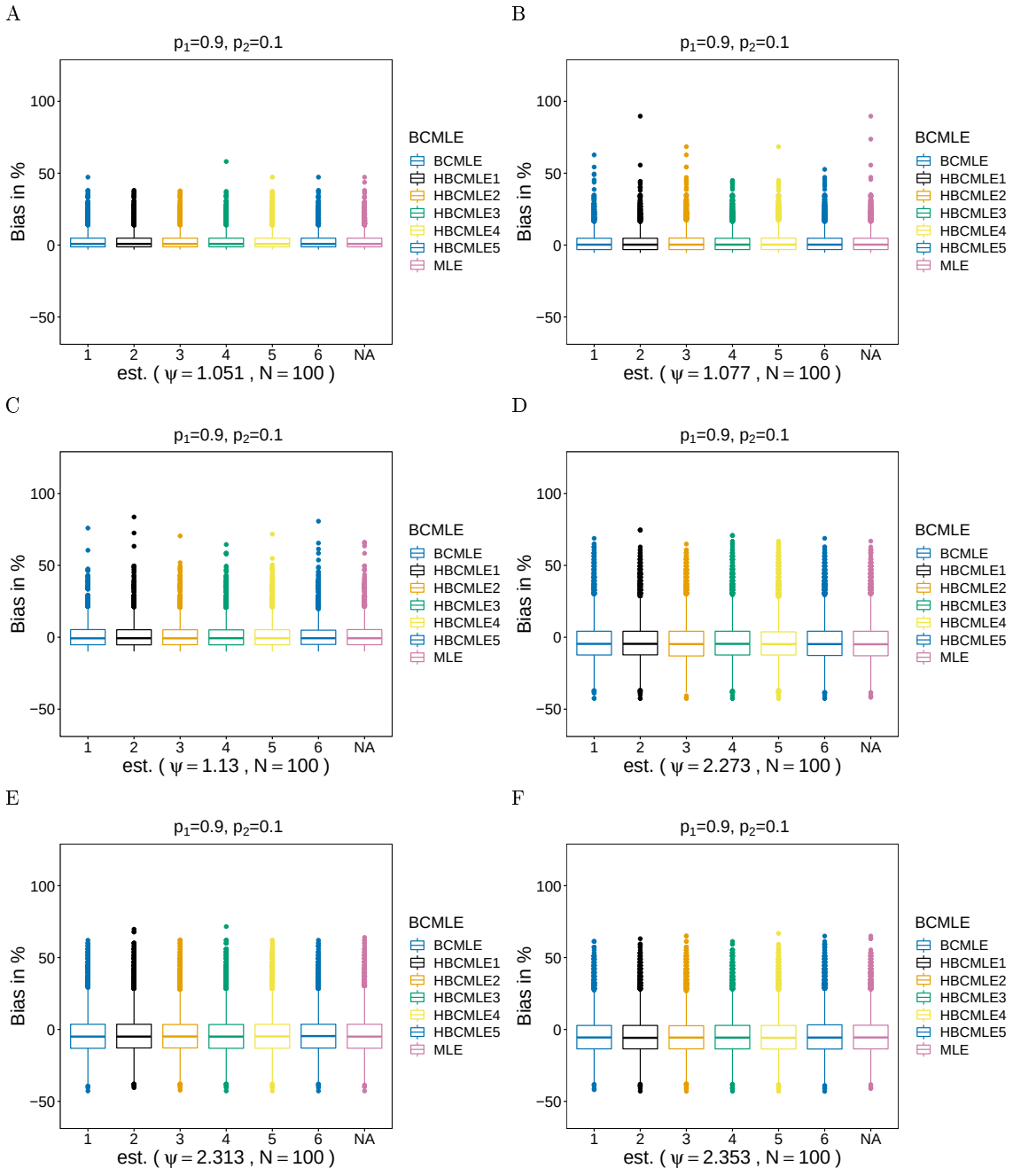


Figure 58: Same as Figure 51

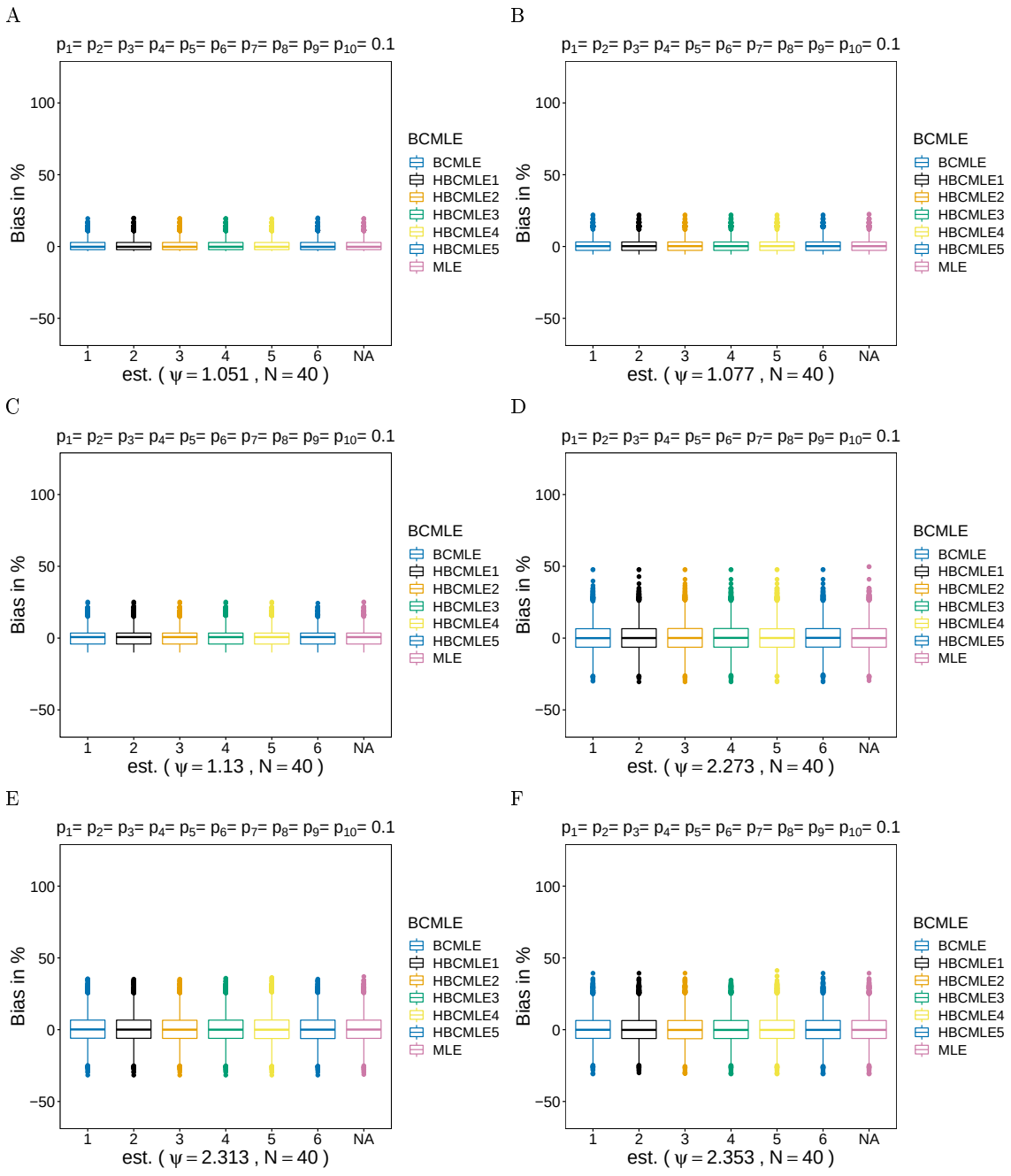


Figure 59: Same as Figure 51

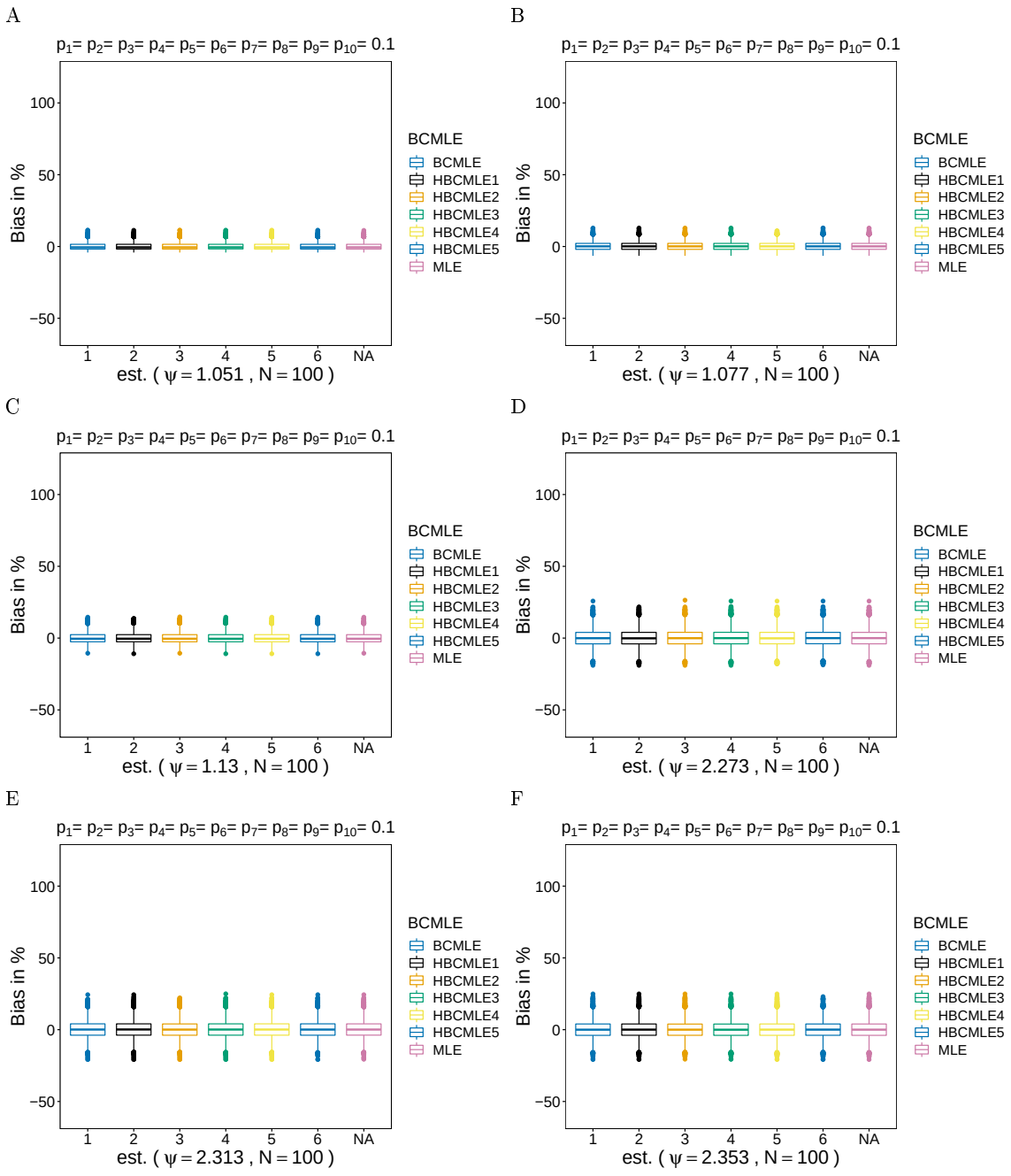


Figure 60: Same as Figure 51

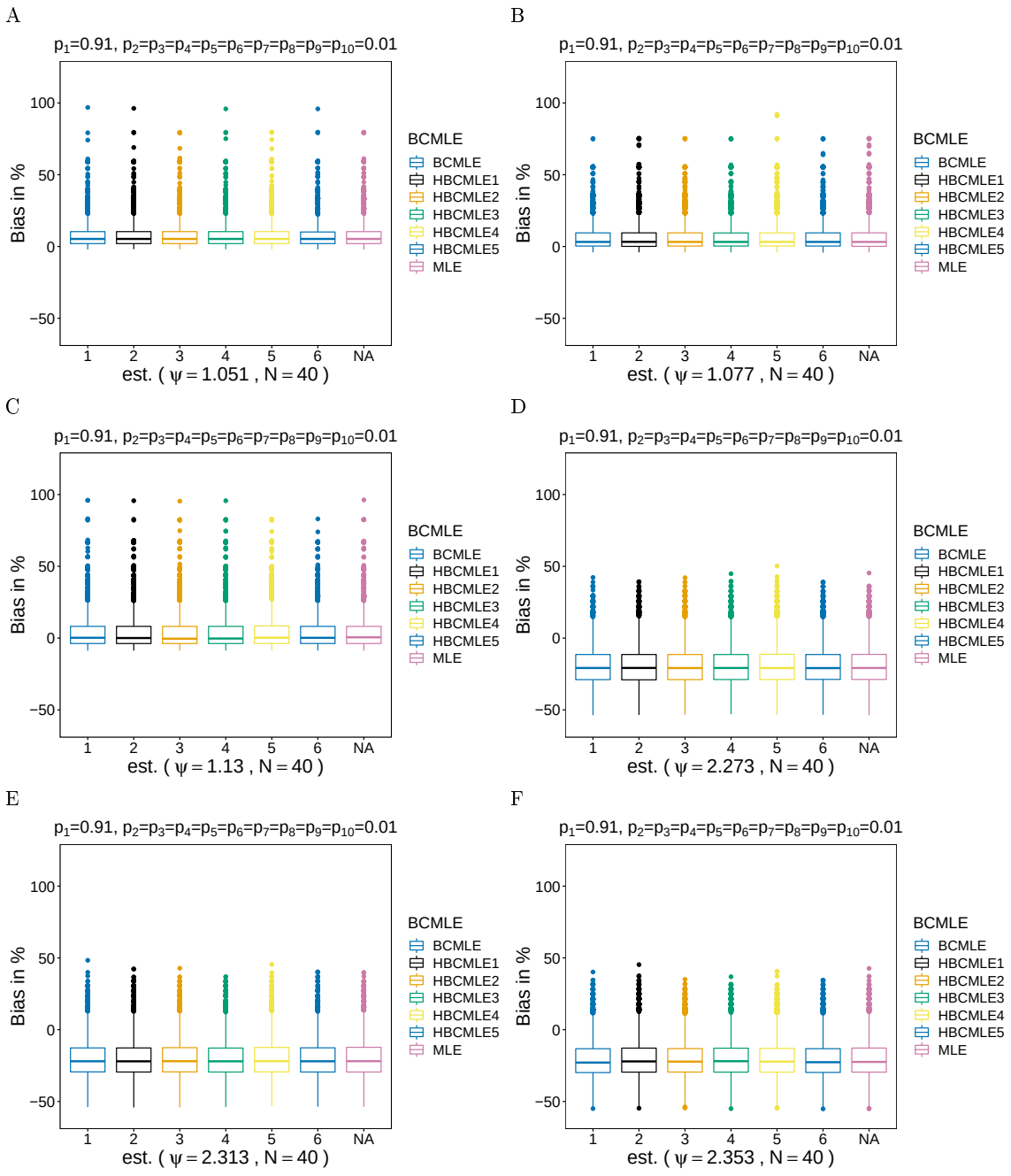


Figure 61: Same as Figure 51

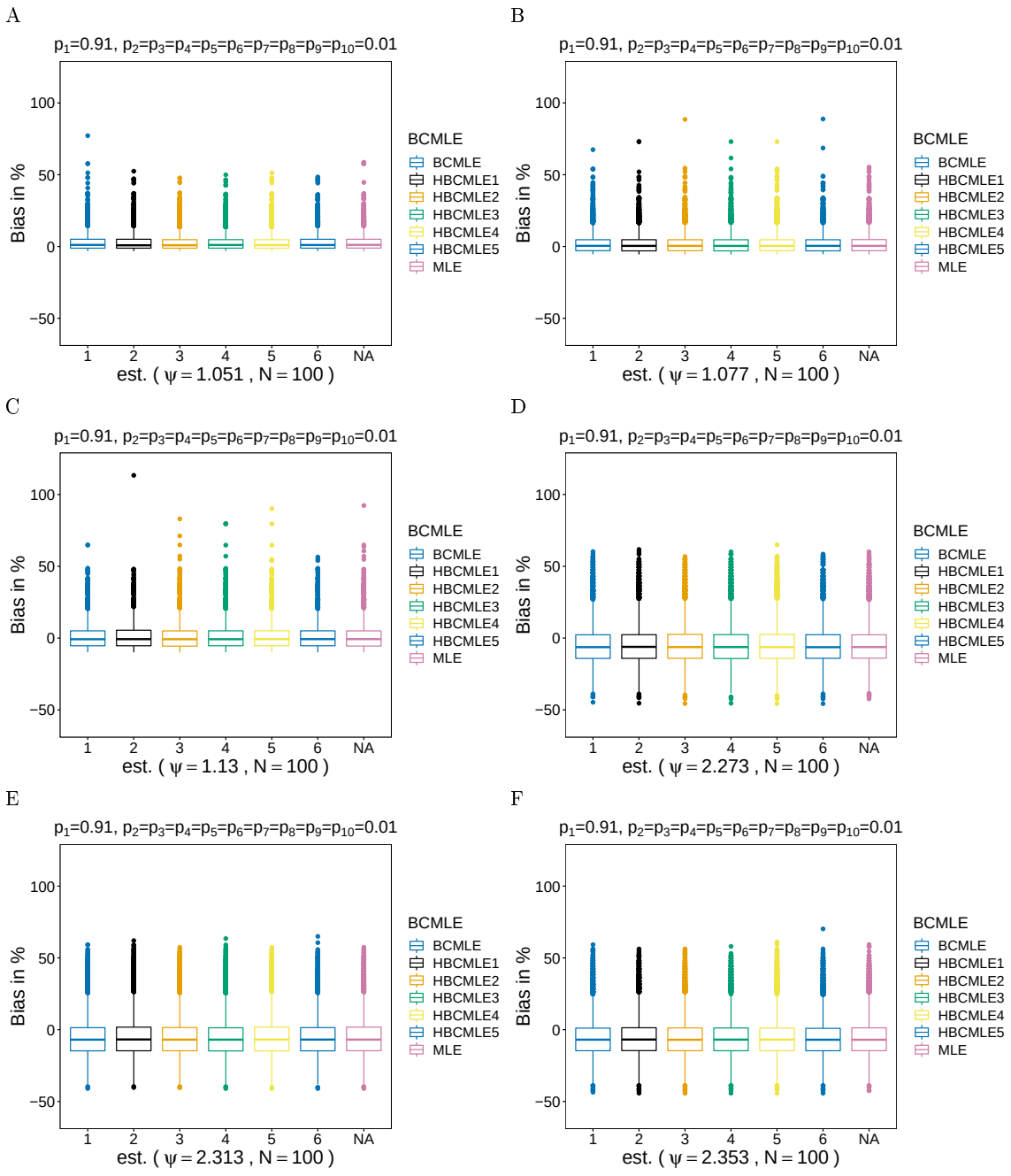


Figure 62: Same as Figure 51