Results:

Scenario 1:

Data generation model: $\pi_{ijl} = exp(1 + 1.36i + x_{ijl} + \delta_{ij})$

Missing model: $logit(R_{ijl} = 0|Y_{ij}, X_{ij}) = -1.8 + X_{ijl} + i$

Independent working correlation matrix

Table 1: The results of unadjusted CRA and adjusted CRA in scenario 1

	mis	Methods	Est	MCSD	SD	coverage	non_con
k=25, m=25							
1	2	UCRA	0.883	0.110	0.134	0.06	0
2	2	CRA	1.320	0.170	0.169	0.95	0
k=25, m=50							
11	2	UCRA	0.881	0.078	0.094	0.00	0
12	2	CRA	1.316	0.119	0.119	0.96	0
k=50, m=25							
21	2	UCRA	0.883	0.077	0.094	0.00	0
22	2	CRA	1.320	0.119	0.119	0.95	0
k=50, m=50							
31	2	UCRA	0.884	0.054	0.067	0.00	0
32	2	CRA	1.321	0.084	0.084	0.95	0

UCRA: unadjusted complete record analysis. CRA: adjusted complete record analysis. Est: the Average estimate. MCSD: Monte Carlo sd.

Table 2: The results of IPWs in scenario 1

				Est		MCSD		SD		coverage		non_con	
mis	True	misp	Methods	No	Stab	No	Stab	No	Stab	No	Stab	No	Stab
k=25, m=25													
2	1.32	0.31	IPW1	1.397	1.325	0.306	0.197	0.334	0.194	0.95	0.95	116	0
2			IPW2	1.367	1.325	0.324	0.197	0.342	0.189	0.95	0.94	0	0
2			IPW1__CLU	1.397	1.325	0.306	0.197	0.335	0.194	0.95	0.95	116	0
2			IPW2__CLU	1.367	1.325	0.325	0.197	0.344	0.189	0.95	0.94	0	0
k=25, m=50													
2	1.32	0.308	IPW1	1.357	1.323	0.223	0.141	0.250	0.137	0.96	0.95	88	0
2			IPW2	1.337	1.323	0.237	0.141	0.253	0.133	0.95	0.94	0	0
2			IPW1__CLU	1.359	1.323	0.223	0.141	0.250	0.137	0.96	0.95	92	0
2			IPW2__CLU	1.337	1.323	0.238	0.141	0.254	0.133	0.95	0.94	0	0
k=50, m=25													
2	1.316	0.309	IPW1	1.361	1.321	0.221	0.141	0.251	0.138	0.96	0.95	88	0
2			IPW2	1.345	1.321	0.230	0.141	0.258	0.135	0.96	0.94	0	0
2			IPW1__CLU	1.363	1.321	0.221	0.141	0.252	0.138	0.96	0.95	89	0
2			IPW2__CLU	1.345	1.321	0.231	0.141	0.258	0.135	0.96	0.94	0	0
k=50, m=50													
2	1.321	0.308	IPW1	1.349	1.323	0.158	0.097	0.186	0.098	0.97	0.96	63	0
2			IPW2	1.334	1.323	0.174	0.097	0.190	0.096	0.96	0.95	0	0
2			IPW1__CLU	1.349	1.323	0.159	0.097	0.186	0.098	0.97	0.96	63	0
2			IPW2__CLU	1.334	1.323	0.175	0.097	0.190	0.096	0.96	0.95	0	0

- No: without weight adjustment.

- stab: with weight stabilization.
- IPW1: ipw by using packages geepack and lme4, without cluster effects. IPW2: ipw by using package CRTgeeDR, without cluster effects. IPW1_CLU: ipw by using packages geepack and lme4, with cluster effects. IPW2_CLU: ipw by using package CRTgeeDR, with cluster effects.

Scenario 2:

Data generation model: $\pi_{ijl} = exp(1 + 1.36i + x_{ijl} + \delta_{ij})$; Missing model: (with cluster effects) $logit(R_{ijl} = 0|Y_{ij}, X_{ij}) = -1.8 + X_{ijl} + i + \theta_{ij}$,$\theta_{ij} \sim N(0, 0.004)$

; Exchangeable working correlation matrix

Table 3: The results of unadjusted CRA and adjusted CRA in scenario 2

	mis	Methods	Est	MCSD	SD	coverage	non_con
k=25, m=25							
1	2	UCRA	0.884	0.110	0.146	0.07	0
2	2	CRA	1.321	0.170	0.169	0.94	0
k=25, m=50							
11	2	UCRA	0.881	0.078	0.099	0.00	0
12	2	CRA	1.316	0.119	0.119	0.95	0
k=50, m=25							
21	2	UCRA	0.885	0.076	0.109	0.01	0
22	2	CRA	1.319	0.119	0.119	0.94	0
k=50, m=50							
31	2	UCRA	0.884	0.055	0.078	0.00	0
32	2	CRA	1.321	0.084	0.084	0.95	0

Table 4: The results of IPWs in scenario 2

				Est		MCSD		SD		coverage		non_con	
mis	True	misp	Methods	No	Stab	No	Stab	No	Stab	No	Stab	No	Stab
k=25, m=25													
2	1.321	0.31	IPW1	1.398	1.325	0.307	0.197	0.334	0.194	0.95	0.93	116	0
2			IPW2	1.369	1.326	0.320	0.200	0.327	0.190	0.90	0.93	0	0
2			IPW1_CLU	1.397	1.326	0.307	0.197	0.335	0.194	0.95	0.94	116	0
2			IPW2_CLU	1.369	1.326	0.320	0.200	0.329	0.190	0.90	0.93	0	0
k=25, m=50													
2	1.319	0.308	IPW1	1.358	1.323	0.225	0.141	0.250	0.137	0.94	0.93	88	0
2			IPW2	1.346	1.325	0.227	0.142	0.242	0.135	0.84	0.93	0	0
2			IPW1_CLU	1.361	1.323	0.224	0.141	0.250	0.137	0.94	0.93	92	0
2			IPW2_CLU	1.348	1.325	0.227	0.142	0.236	0.135	0.83	0.93	0	0
k=50, m=25													
2	1.316	0.309	IPW1	1.361	1.321	0.221	0.141	0.251	0.138	0.96	0.95	88	0
2			IPW2	1.346	1.323	0.225	0.140	0.248	0.136	0.93	0.94	0	0
2			IPW1_CLU	1.362	1.321	0.221	0.141	0.252	0.138	0.96	0.95	89	0
2			IPW2_CLU	1.346	1.323	0.226	0.140	0.248	0.136	0.93	0.94	0	0
k=50, m=50													
2	1.321	0.308	IPW1	1.349	1.323	0.159	0.097	0.186	0.098	0.96	0.95	63	0
2			IPW2	1.341	1.323	0.163	0.098	0.183	0.097	0.87	0.95	0	0
2			IPW1_CLU	1.350	1.323	0.159	0.097	0.186	0.098	0.96	0.95	63	0
2			IPW2_CLU	1.341	1.323	0.163	0.098	0.183	0.097	0.87	0.95	0	0

Scenario 3:

Data generation model:

- Intervention arm: $\pi_{ijl} = exp(1 + 1.36i + x_{ijl} + \delta_{ij})$
- control arm: $\pi_{ijl} = exp(1 + 1.36i + 0.588x_{ijl} + \delta_{ij})$

Missing model: $logit(R_{ijl} = 0|Y_{ij}, X_{ij}) = -1.8 + X_{ijl} + i$

Independent working correlation matrix

Table 5: The results of unadjusted CRA and adjusted CRA in scenario 3

	mis	Methods	Est	MCSD	SD	coverage	non_con
k=25, m=25							
1	5	UCRA	0.883	0.110	0.134	0.06	0
2	5	CRA	1.320	0.170	0.169	0.95	0
k=25, m=50							
11	5	UCRA	0.881	0.078	0.094	0.00	0
12	5	CRA	1.316	0.119	0.119	0.96	0
k=50, m=25							
21	5	UCRA	0.883	0.077	0.094	0.00	0
22	5	CRA	1.320	0.119	0.119	0.95	0
k=50, m=50							
31	5	UCRA	0.884	0.054	0.067	0.00	0
32	5	CRA	1.321	0.084	0.084	0.95	0

Table 6: The results of IPWs in scenario 3

				Est		MCSD		SD		coverage		non_con	
mis	True	misp	Methods	No	Stab	No	Stab	No	Stab	No	Stab	No	Stab
k=25, m=25													
5	1.32	0.31	IPW1	1.396	1.325	0.305	0.200	0.334	0.194	0.95	0.95	110	0
5			IPW2	1.367	1.325	0.323	0.200	0.341	0.189	0.95	0.94	0	0
5			IPW1_CLU	1.397	1.325	0.306	0.200	0.335	0.194	0.95	0.95	112	0
5			IPW2_CLU	1.368	1.325	0.324	0.200	0.343	0.189	0.95	0.94	0	0
k=25, m=50													
5	1.32	0.308	IPW1	1.357	1.322	0.223	0.140	0.249	0.137	0.96	0.96	86	0
5			IPW2	1.337	1.322	0.237	0.140	0.252	0.133	0.95	0.94	0	0
5			IPW1_CLU	1.359	1.322	0.223	0.140	0.250	0.137	0.96	0.96	90	0
5			IPW2_CLU	1.337	1.322	0.239	0.140	0.253	0.133	0.95	0.94	0	0
k=50, m=25													
5	1.316	0.309	IPW1	1.362	1.321	0.220	0.140	0.251	0.138	0.96	0.95	88	0
5			IPW2	1.345	1.321	0.229	0.140	0.257	0.136	0.96	0.94	0	0
5			IPW1_CLU	1.364	1.321	0.220	0.140	0.251	0.138	0.96	0.95	90	0
5			IPW2_CLU	1.345	1.321	0.230	0.140	0.258	0.136	0.96	0.94	0	0
k=50, m=50													
5	1.321	0.308	IPW1	1.350	1.324	0.158	0.098	0.185	0.098	0.97	0.95	66	0
5			IPW2	1.334	1.324	0.174	0.098	0.189	0.096	0.96	0.94	0	0
5			IPW1_CLU	1.350	1.324	0.158	0.098	0.186	0.098	0.96	0.95	67	0
5			IPW2_CLU	1.334	1.323	0.175	0.098	0.190	0.096	0.96	0.94	0	0

Scenario 4:

Data generation model:

- Intervention arm: $\pi_{ijl} = exp(1 + 1.36i + x_{ijl} + \delta_{ij})$
- control arm: $\pi_{ijl} = exp(1 + 1.36i + 0.588x_{ijl} + \delta_{ij})$

Missing model: (with cluster effects)

$logit(R_{ijl} = 0|Y_{ij}, X_{ij}) = -1.8 + X_{ijl} + i + \theta_{ij}$,$\theta_{ij} \sim N(0, 0.004)$

Exchangeable working correlation matrix

Table 7: The results of unadjusted CRA and adjusted CRA in scenario 4

	mis	Methods	Est	MCSD	SD	coverage	non_con
k=25, m=25							
1	5	UCRA	0.884	0.110	0.146	0.07	0
2	5	CRA	1.321	0.170	0.169	0.94	0
k=25, m=50							
11	5	UCRA	0.881	0.078	0.099	0.00	0
12	5	CRA	1.316	0.119	0.119	0.95	0
k=50, m=25							
21	5	UCRA	0.885	0.076	0.109	0.01	0
22	5	CRA	1.319	0.119	0.119	0.94	0
k=50, m=50							
31	5	UCRA	0.884	0.055	0.078	0.00	0
32	5	CRA	1.321	0.084	0.084	0.95	0

Table 8: The results of IPWs in scenario 4

				Est		MCSD		SD		coverage		non_con	
mis	True	misp	Methods	No	Stab	No	Stab	No	Stab	No	Stab	No	Stab
k=25, m=25													
5	1.321	0.31	IPW1	1.397	1.325	0.306	0.200	0.334	0.194	0.94	0.94	110	0
5			IPW2	1.367	1.327	0.313	0.203	0.325	0.190	0.90	0.93	0	0
5			IPW1_CLU	1.398	1.325	0.307	0.200	0.335	0.194	0.94	0.93	112	0
5			IPW2_CLU	1.367	1.327	0.315	0.202	0.327	0.190	0.90	0.93	0	0
k=25, m=50													
5	1.319	0.308	IPW1	1.359	1.322	0.225	0.140	0.249	0.137	0.94	0.94	86	0
5			IPW2	1.343	1.324	0.228	0.141	0.239	0.135	0.84	0.94	0	0
5			IPW1_CLU	1.360	1.322	0.225	0.140	0.250	0.137	0.94	0.94	90	0
5			IPW2_CLU	1.344	1.324	0.230	0.141	0.240	0.135	0.83	0.94	0	0
k=50, m=25													
5	1.316	0.309	IPW1	1.362	1.321	0.220	0.140	0.251	0.138	0.96	0.95	88	0
5			IPW2	1.345	1.322	0.225	0.139	0.251	0.136	0.93	0.94	0	0
5			IPW1_CLU	1.364	1.321	0.220	0.140	0.251	0.138	0.96	0.95	90	0
5			IPW2_CLU	1.345	1.322	0.226	0.139	0.251	0.136	0.93	0.94	0	0
k=50, m=50													
5	1.321	0.308	IPW1	1.350	1.324	0.159	0.098	0.185	0.098	0.96	0.95	66	0
5			IPW2	1.341	1.323	0.164	0.099	0.184	0.097	0.88	0.95	0	0
5			IPW1_CLU	1.351	1.324	0.159	0.098	0.186	0.098	0.96	0.95	67	0
5			IPW2_CLU	1.341	1.323	0.164	0.099	0.185	0.097	0.87	0.95	0	0

Simulation

Recall:

1. Data Generation Function:

$$\pi_{ijl} = exp(\beta_0 + \beta_1 i + f_i(x_{ijl}) + \delta_{ij})$$

- *ijl* means the *ith* intervention group, the *jth* cluster, the *lth* individual.
- i=0 control group while i=1 intervention group.
- β_0 is a constant, β_1 is the true intervention effect. $f_i(x_{ijl})$ is a function of baseline coveariate X in the *ith* intervention group. We set $\beta_0 = 1, \beta_1 = 1.36$, which consistent to Hossain's paper.
- X_{ijl} is generated by using the methods:

  $$X_{ijl} = \alpha_{ij} + u_{ijl}$$

  $$\alpha_{ij} \sim N(0, 0.18), u_{ijl} \sim N(0, 3.37)$$

  where α_{ij} is the (ij)th cluster effect on X and u_{ijl} is the individual-level error on X. Therefore, the variance of x is $0.18 + 3.37 = 3.55$, the ICC is $\rho = 0.18/3.55 = 0.05$
- $\delta_{ij} \sim N(0, \sigma_b^2)$. We set $\sigma_b^2 = 0.2$
- Y_{ijl} is generated as Bernoulli random varaible with parameter π_{ijl}

Data generating scenarios: S1 and S3.

S1:

- Intervention arm and control arm: $\pi_{ijl} = exp(1 + 1.36i + x_{ijl} + \delta_{ij})$

S3:

- Intervention arm: $\pi_{ijl} = exp(1 + 1.36i + x_{ijl} + \delta_{ij})$
- control arm: $\pi_{ijl} = exp(1 + 1.36i + 0.588x_{ijl} + \delta_{ij})$

We tried k=25 and k=50 clusters in each intervention arm, respectively. And we also considered cluster size with mean m=25 and m=50, where the cluster size was chosen from uniform distributions: UNI(20,30) and UNI(45,55).

Since we are also considering to mimic the HALI data, the simulation should have the similar CV of cluster size as HALI data, which is about 0.1. And the UNI(20,30) is appropriate for mimicking the HALI cluster size with a similar CV. The other cluster size distribution is revised as UNI(40,60) so that the CV is around 0.1. That is:

- With mean=25, m $\sim$ UNI(20,30)
- With mean=50, m $\sim$ UNI(40,60)

2. Missingness generation:

We assume the missing mechanism is covariate dependent missingness (CDM).

The missingness is generated by the logistic regression model (the original method in Hossain's paper):

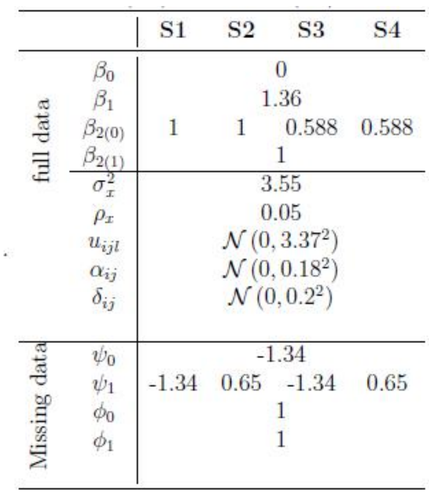
$$logit(R_{ijl} = 0 | Y_{ij}, X_{ij}) = \psi_i + \phi_i X_{ijl}$$

With values:

$$logit(R_{ijl} = 0|Y_{ij}, X_{ij}) = -1.34 + X_{ijl}$$

The following table shows the parameters used in Hossain's paper. And we used the same parameters as Hossain's.

Parameter Table in Hossain's Paper



		S1	S2	S3	S4
full data	β_0	0			
	β_1	1.36			
	$\beta_{2(0)}$	1	1	0.588	0.588
	$\beta_{2(1)}$	1			
	σ_x^2	3.55			
	ρ_x	0.05			
	u_{ijl}	$\mathcal{N}(0, 3.37^2)$			
	α_{ij}	$\mathcal{N}(0, 0.18^2)$			
	δ_{ij}	$\mathcal{N}(0, 0.2^2)$			
Missing data	ψ_0	-1.34			
	ψ_1	-1.34	0.65	-1.34	0.65
	ϕ_0	1			
	ϕ_1	1			

This time we consider two scenarios in Hossain's paper: scenario 1 and scenario 3.

In scenario 1, covariate of X in each intervention group is the same while in scenario 3 is different. However, the missing generation methods are the same in these two scenarios since the generation is not associated with intervention arm and the parameters are the same.

We then tried to consider more scenarios in missingness generation.

3. More missingness generation models based on previous results

We chose missing model 2 and 5 from those 6 models that we built last time.

We changed the value of intercepts to make the missing percentage around 30 %. I generated some values and found -1.8 is suitable.

Missingness Model 2. x+arm

A convariate representing intervention arm is added. In Hossain's paper, S2 and S4 have different missing generation models since:

Control arm: $logit(R_{ijl} = 0|Y_{ij}, X_{ij}) = -1.34 + X_{ijl}$

Intervention arm: $logit(R_{ijl} = 0|Y_{ij}, X_{ij}) = 0.65 + X_{ijl}$

which means the coefficient for arm covariate is 1.99. However, I did not choose this coefficient since it increases the missing percentage from 30 % to 60 %, which may make these scenarios uncomparable. So I just choose 1 as the coefficient and the model is specified as:

$$logit(R_{ijl} = 0|Y_{ij}, X_{ij}) = -1.8 + X_{ijl} + i$$

Missingness Model 5. x+arm+θ_{ij}

This model is same as model 2, while cluster effects were considered.

$$logit(R_{ijl} = 0|Y_{ij}, X_{ij}) = -1.8 + X_{ijl} + i + \theta_{ij}$$

$$\theta_{ij} \sim N(0, 0.003936118)$$

The random cluster effects added in model 4-6 is chosen as $\theta_{ij} \sim N(0, 0.003936118)$. This can make the missing ICC as 0.05.

The calculation methods of missing ICC is from Fan and Dr. Turner's paper [1] :'An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes'.

$$\text{ICC}=\sigma_\theta^2/(\sigma_\theta^2 + \pi^2/3)$$

However, with a missing icc as 0.05, the variance becomes quite small.

4. Check the distribution of x_{ijl}

Check the values of x_{ijl} to see whether extreme values will be generated.

Generate 1000,000 samples

We can firstly generate 1000,000 samples and calculate the max weigth

```
try_u=rnorm(1000000,0,sqrt(3.37))
try_a=rnorm(1000,0,sqrt(0.18))
try_a2=rep(try_a,1000)
try_x=try_u+try_a2

# The max weight
1/expit(-1.8+min(try_x))
```

```
## [1] 53416.62
```

check the distriubtion of x in simulation data set:

1. Generate one data set:

```
check_x1=one_group3(mis=1,s=1,i=1,k=25,mm=25,seed=123)
```

2. Check the mean value and the sd of the Xs as well as X's distribution

```
mean(check_x1$x)
```

```
## [1] -0.009427745
```

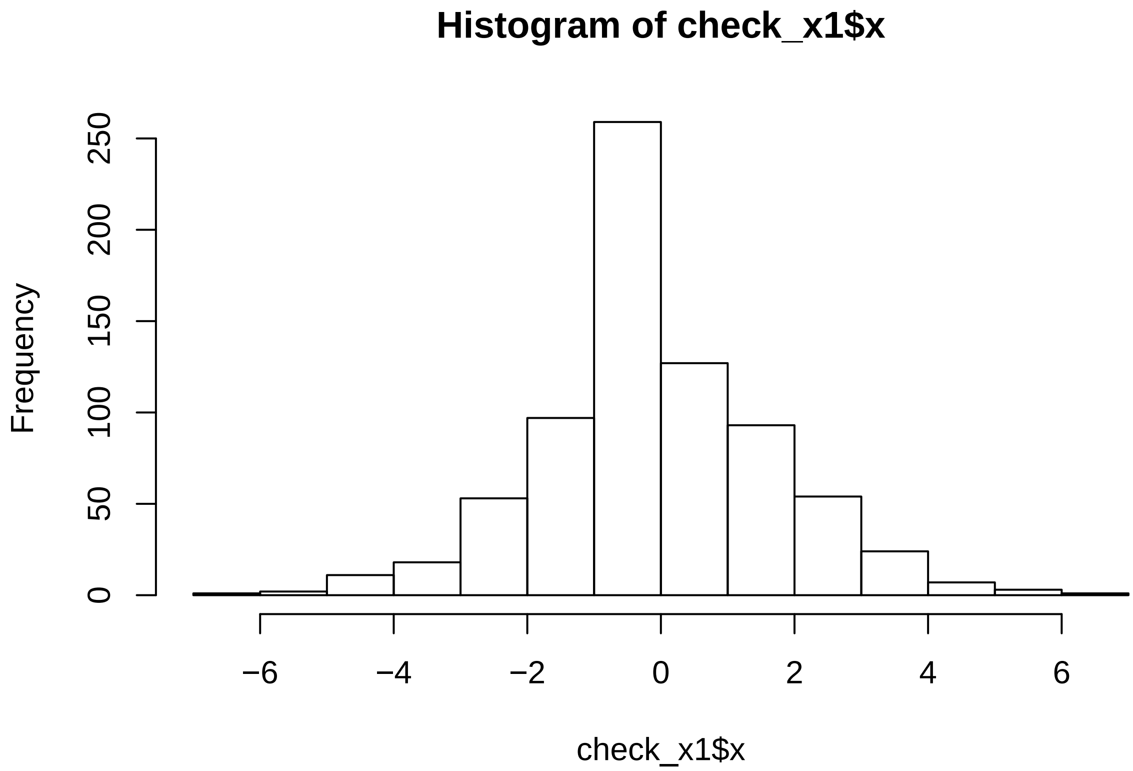
```
sd(check_x1$x)
```

```
## [1] 1.732406
```

```
range(check_x1$x)
```

```
## [1] -6.022233  6.329850
```

```
hist(check_x1$x)
```



```
## max weight
1/expit(-1.8+min(check_x1$x))
```

```
## [1] 2496.472
```

The data generation process can gain some weights bigger than 1000, which may be relatively large. Large weights can cause nonconvergences. Then several methods are used to handle large weights.

5. Weights

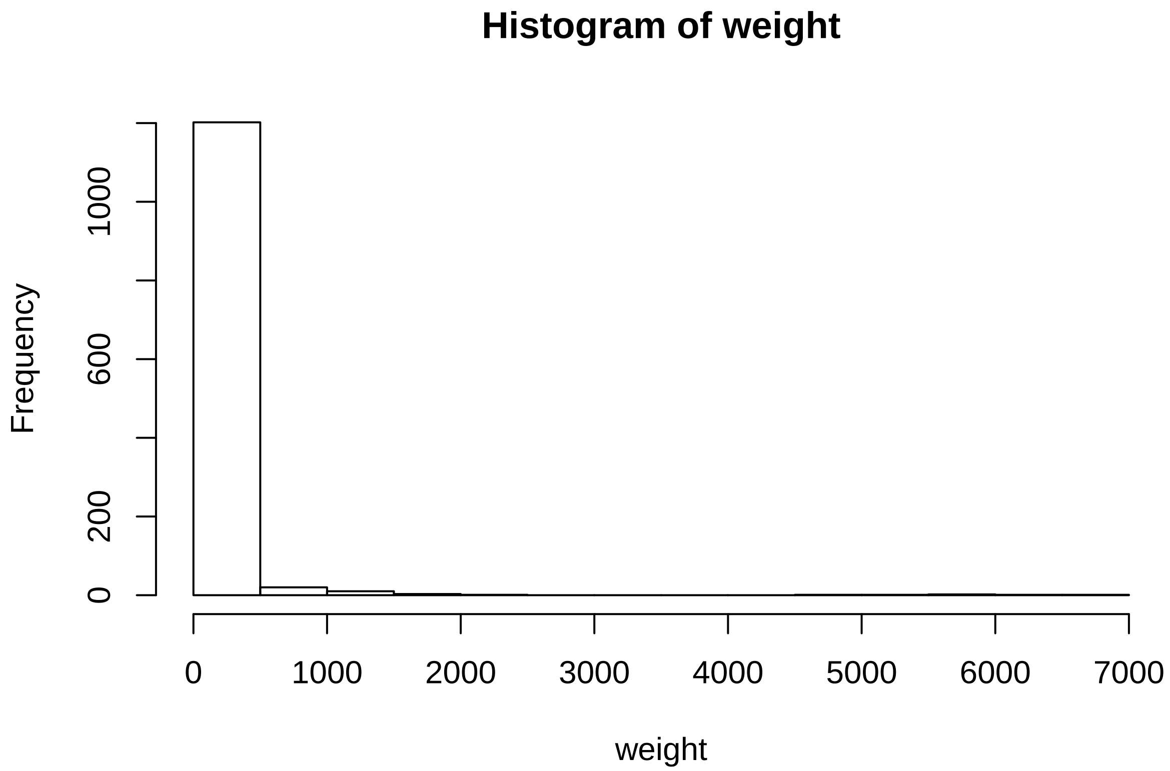
We compared the results of whether deal with the large weigths with stabilization methods or not.

Check the distribution of weights:

```
mis=1
d1=data_gene3(k=25,m=25,s=1,mis=1)
d2=d1
d2$y=ifelse(d2$R==1,NA,d2$y)
d3=data.frame(y=d2$y,x=d2$x,cluster=d2$cluster,arm=d2$arm,missing=d2$R)

logs=glm(missing ~ x+arm, data = d3,
               family = binomial(link='logit'))
weight=1/expit(predict(logs))
```

```r
hist(weight)
```



```r
summary(weight)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##    1.007    3.046    9.261   84.571   35.301 6597.437 
```

```r
sum(weight>100)
```

```
## [1] 146
```

Stablization method:

Stabilized inverse probability weights wecan be defined as[2]:

$$sw = \frac{f_X(X;\mu_1,\sigma_1^2)}{f_{X|C}(X|C=c;\mu_2,\sigma_2^2)}$$

where $f.(.)$ denotes the probability density function with mean μ and variance σ^2, and C is the set of confounders.

```r
#the numerator for stablilized weights
num0 = predict(glm(missing ~ 1,data = d3,
                           family = binomial(link='logit')),type="response")

#the propensity score
ps=expit(predict(logs))
```

```r
sw=ifelse(d3$missing==1, num0/ps, (1-num0)/(1-ps))

d3$sw=sw

range(d3$sw)
```

```
## [1]  0.2189606 41.7656069
```

Summary

- The results of IPW1 and IPW2 look good.
- IPW2 are closer to the true value than IPW1
- Without dealing with weights, there are about 10% of the nonconvergence time. and the results are over-estimated
- With weight stabilization, the results are much closer to the true value and the coverage is about 95 %
- Whether considering the cluster effects or not do not have a big difference in the results. This may be caused by the fact that the cluster effects are relatively small.