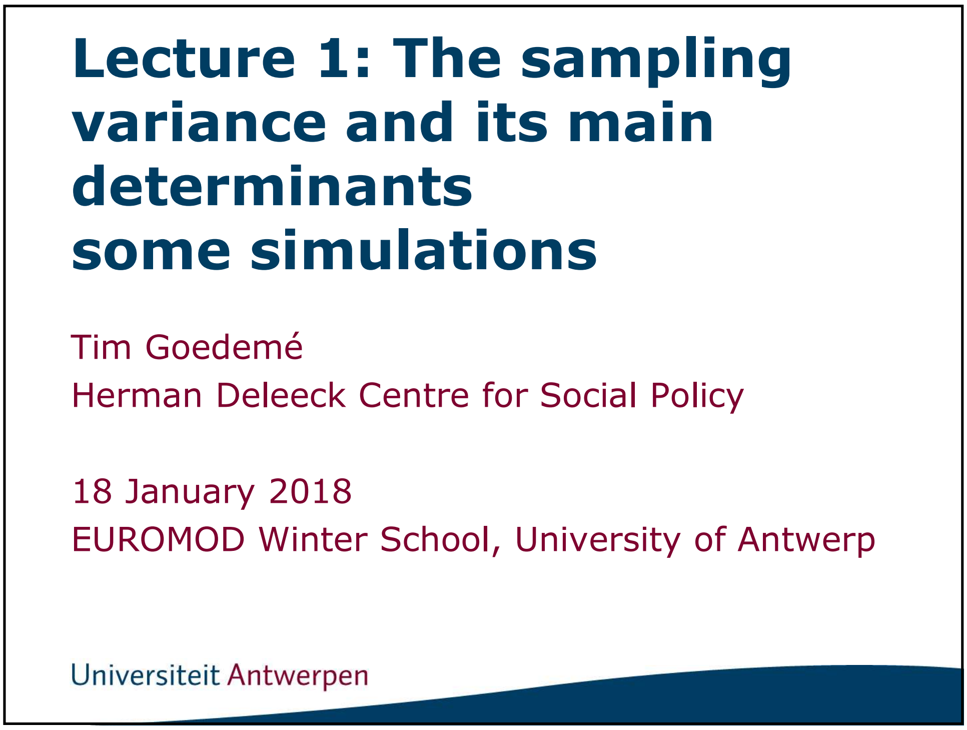
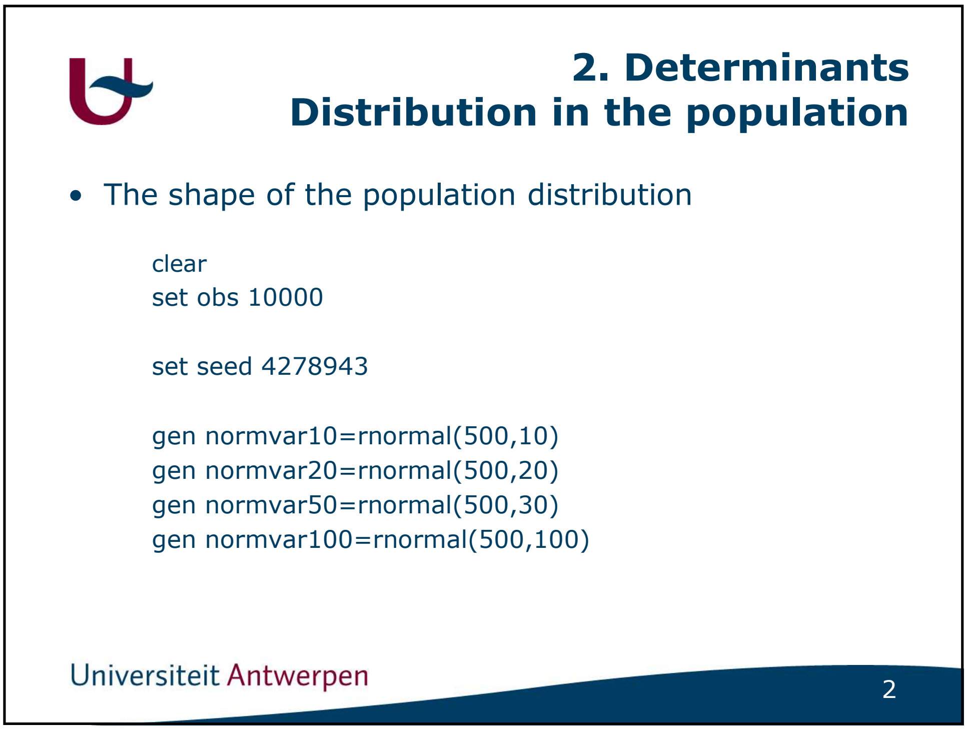


Lecture 1: The sampling variance and its main determinants some simulations

Tim Goedemé
Herman Deleeck Centre for Social Policy

18 January 2018
EUROMOD Winter School, University of Antwerp

Universiteit Antwerpen

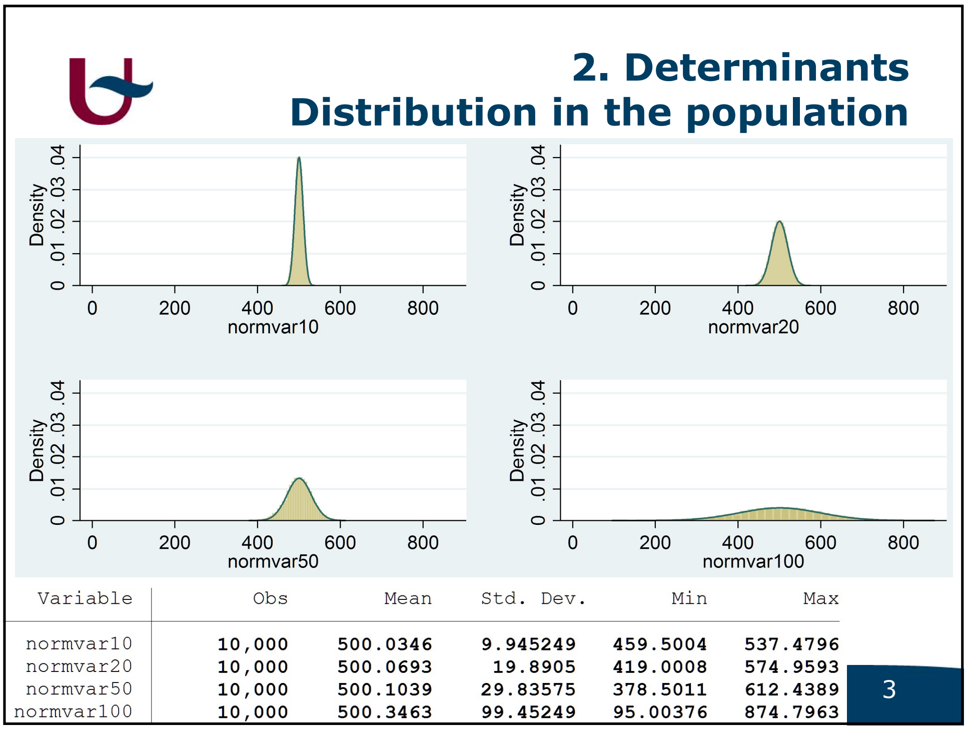


2. Determinants Distribution in the population

- The shape of the population distribution

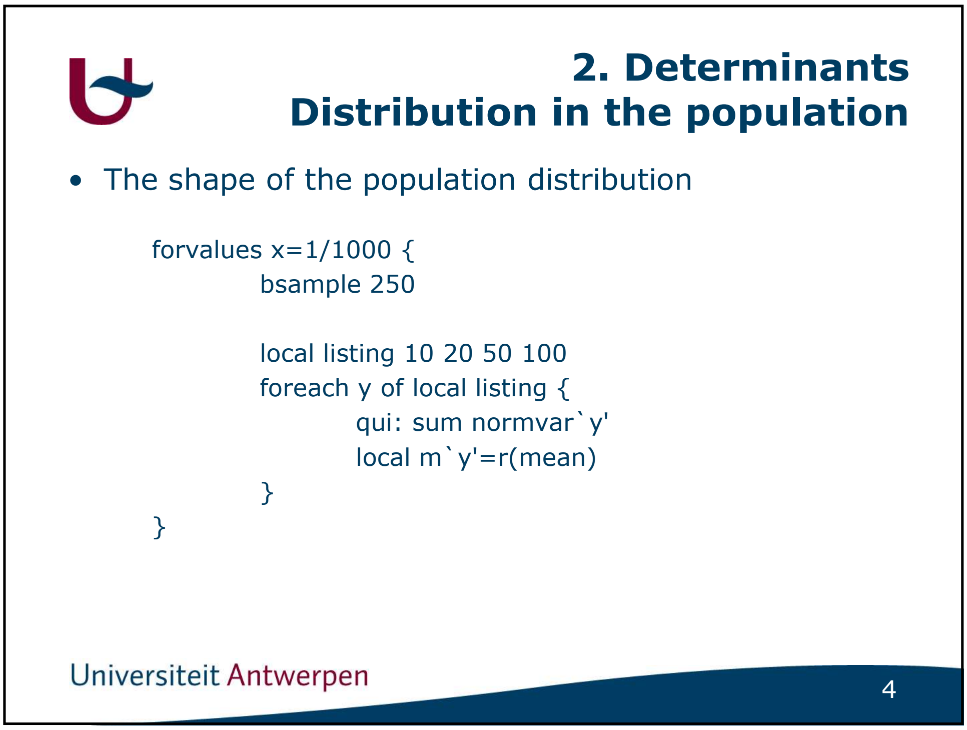
```
clear
set obs 10000

set seed 4278943

gen normvar10=rnormal(500,10)
gen normvar20=rnormal(500,20)
gen normvar50=rnormal(500,30)
gen normvar100=rnormal(500,100)
```

Universiteit Antwerpen



2. Determinants Distribution in the population

```
    Variable         Obs        Mean    Std. Dev.       Min        Max

   normvar10      10,000    500.0346    9.945249   459.5004   537.4796
   normvar20      10,000    500.0693     19.8905   419.0008   574.9593
   normvar50      10,000    500.1039    29.83575   378.5011   612.4389
  normvar100      10,000    500.3463    99.45249   95.00376   874.7963
```



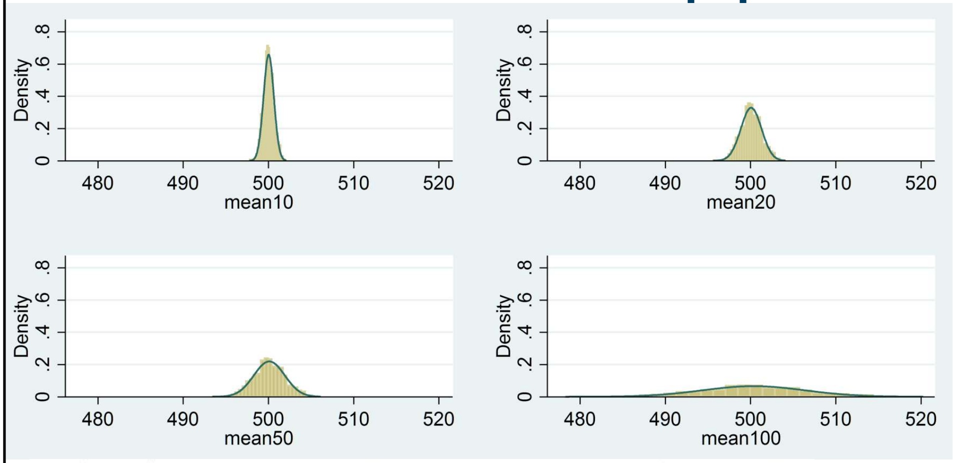
2. Determinants Distribution in the population

- The shape of the population distribution

```
forvalues x=1/1000 {
        bsample 250

        local listing 10 20 50 100
        foreach y of local listing {
                qui: sum normvar`y'
                local m`y'=r(mean)
        }
}
```

Universiteit Antwerpen

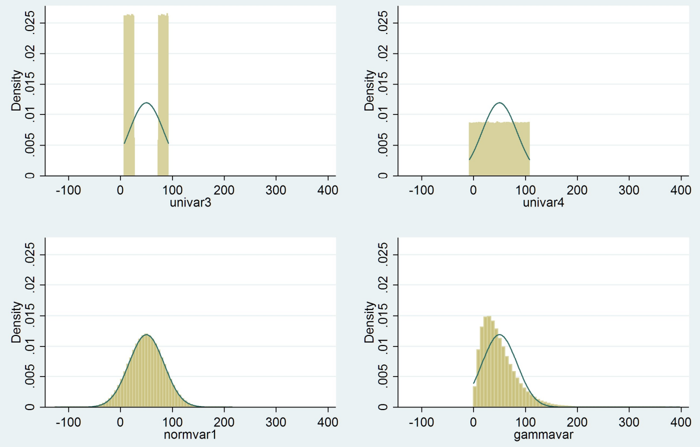
2. Determinants
Distribution in the population



```
    Variable        Obs        Mean    Std. Dev.       Min        Max

      mean10      1,000    500.0369    .6060214    497.831   502.0189
      mean20      1,000    500.0738    1.212043   495.6619   504.0377
      mean50      1,000    500.1108    1.818064   493.4929   506.0566
     mean100      1,000    500.3692    6.060213   478.3096   520.1886
```

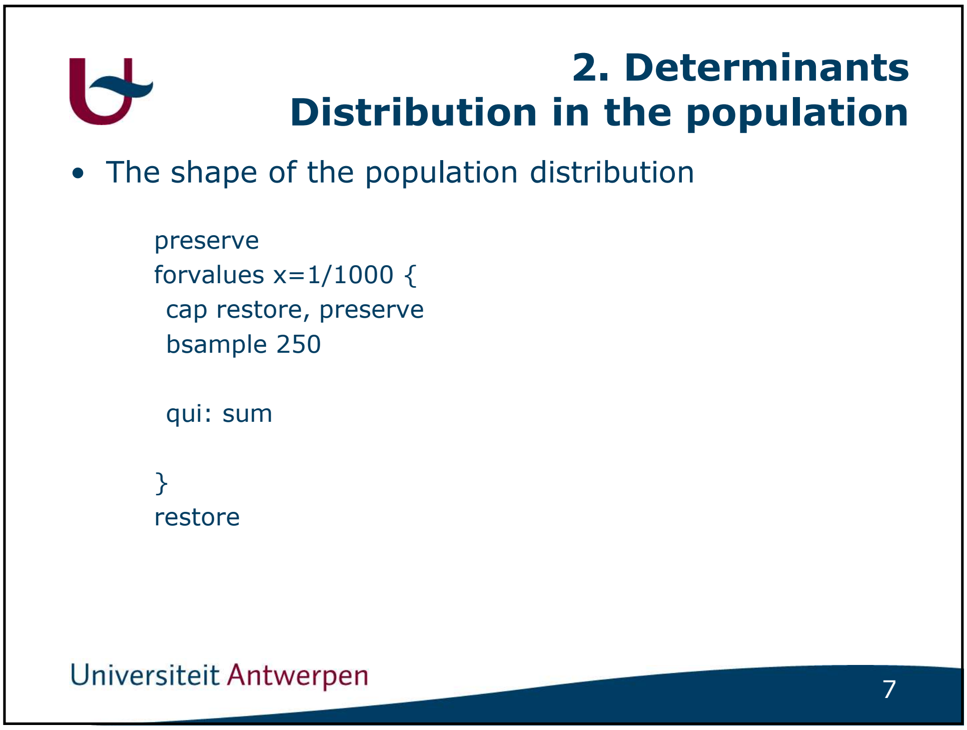
5

2. Determinants
Distribution in the population



```
    Variable        Obs        Mean    Std. Dev.       Min        Max

    gammavar  1,000,000    50.00715    33.50935   .0217693   397.5127
     univar3  1,000,000    50.00809    33.51855   7.377846   92.61644
     univar4  1,000,000    50.00306    33.51426  -8.102166   108.0365
    normvar1  1,000,000    49.97284    33.52434  -127.1313   215.4159
```

6



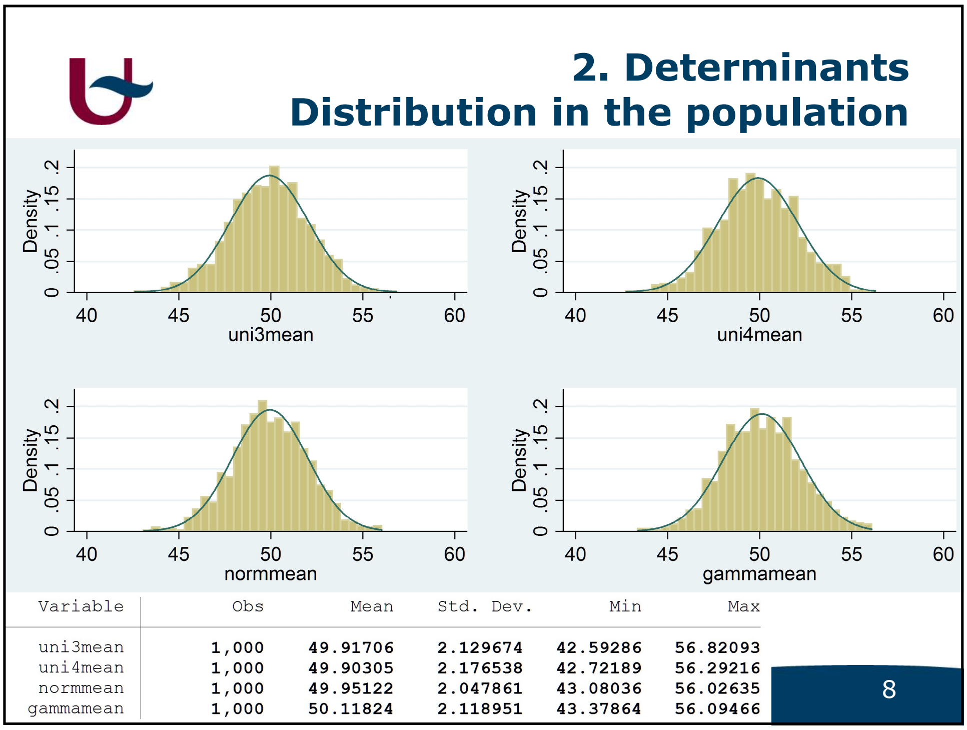
2. Determinants
Distribution in the population

- The shape of the population distribution

```
preserve
forvalues x=1/1000 {
 cap restore, preserve
 bsample 250

 qui: sum

}
restore
```

Universiteit Antwerpen

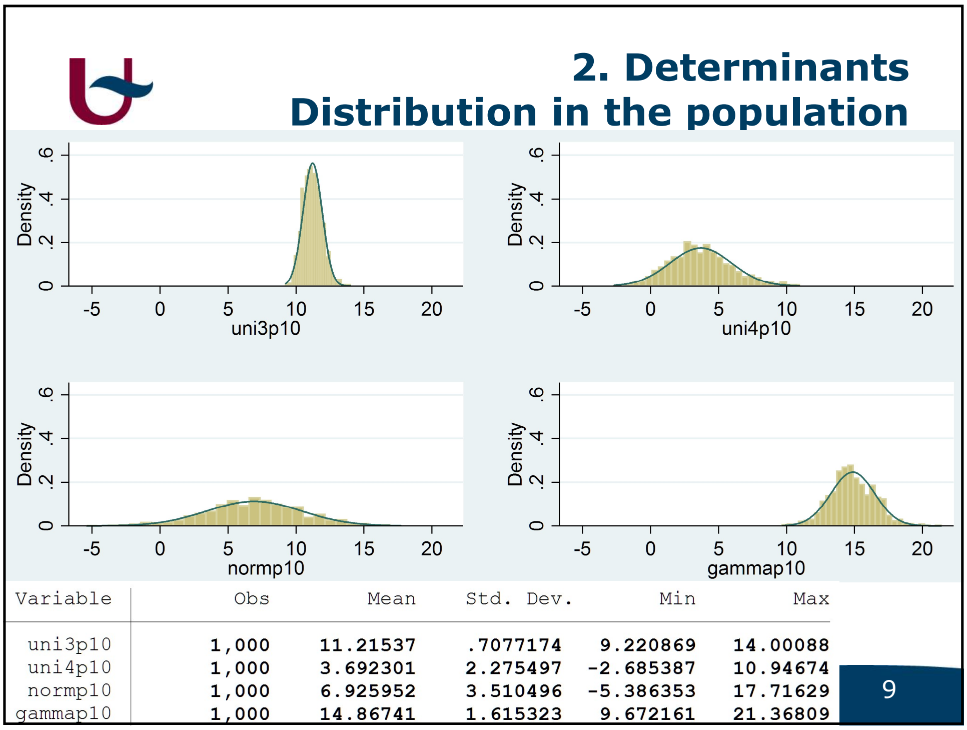
7



2. Determinants
Distribution in the population

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    uni3mean |      1,000    49.91706    2.129674   42.59286   56.82093
    uni4mean |      1,000    49.90305    2.176538   42.72189   56.29216
    normmean |      1,000    49.95122    2.047861   43.08036   56.02635
   gammamean |      1,000    50.11824    2.118951   43.37864   56.09466
```

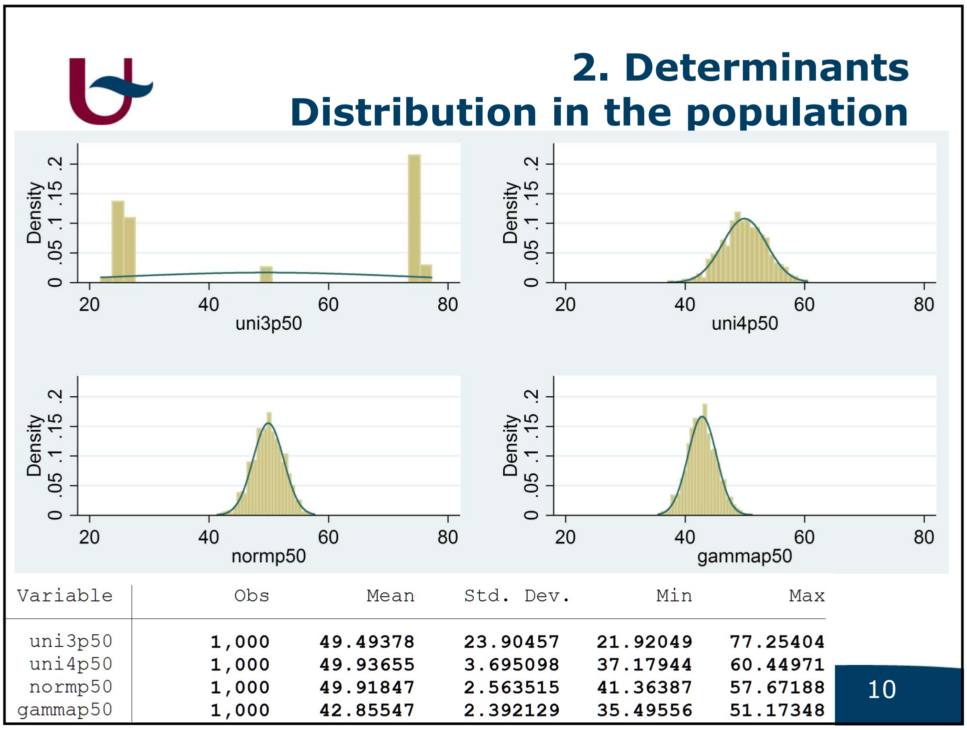
8



2. Determinants
Distribution in the population

```
    Variable         Obs        Mean    Std. Dev.       Min        Max

     uni3p10       1,000    11.21537    .7077174   9.220869   14.00088
     uni4p10       1,000    3.692301    2.275497  -2.685387   10.94674
     normp10       1,000    6.925952    3.510496  -5.386353   17.71629
   gammap10        1,000    14.86741    1.615323   9.672161   21.36809
```

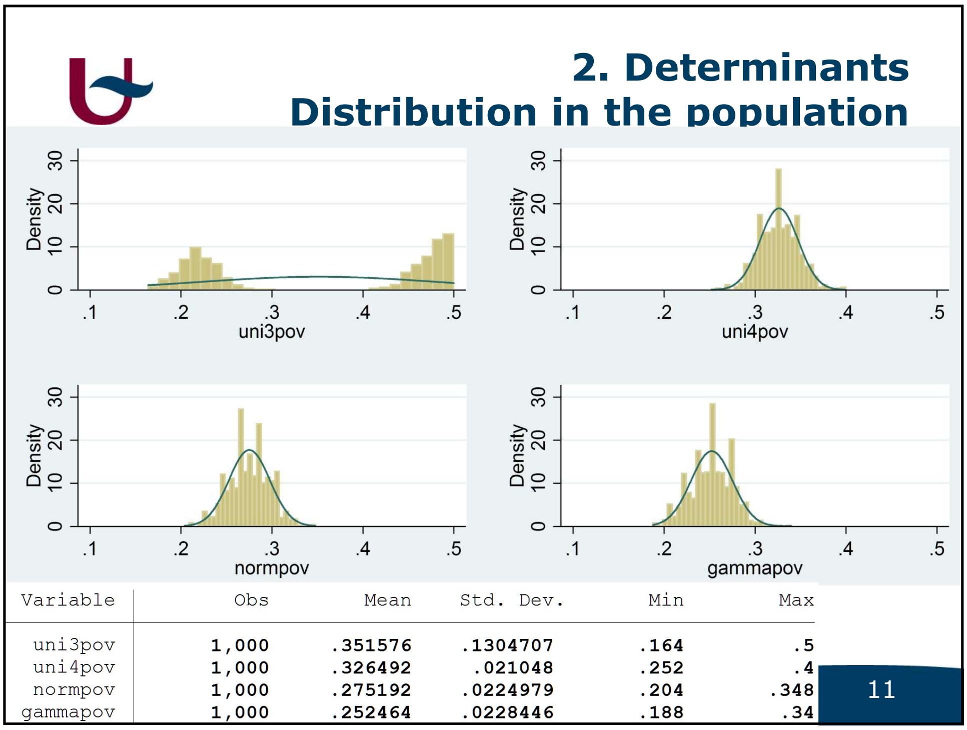
9



2. Determinants
Distribution in the population

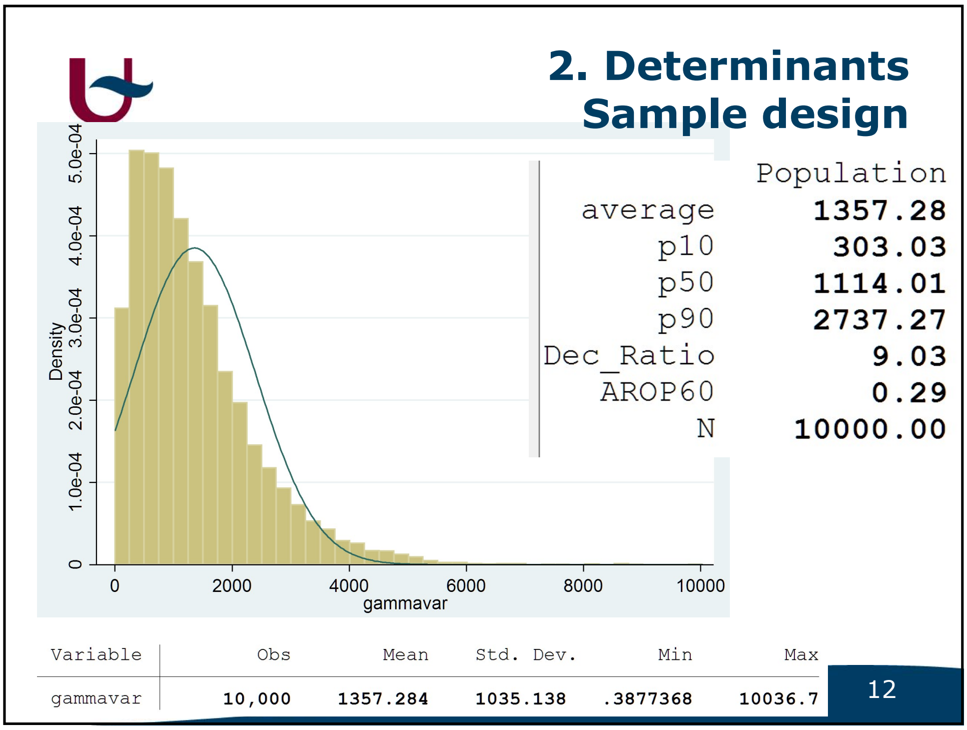
```
    Variable         Obs        Mean    Std. Dev.       Min        Max

     uni3p50       1,000    49.49378    23.90457   21.92049   77.25404
     uni4p50       1,000    49.93655    3.695098   37.17944   60.44971
     normp50       1,000    49.91847    2.563515   41.36387   57.67188
    gammap50       1,000    42.85547    2.392129   35.49556   51.17348
```

10



2. Determinants
Distribution in the population

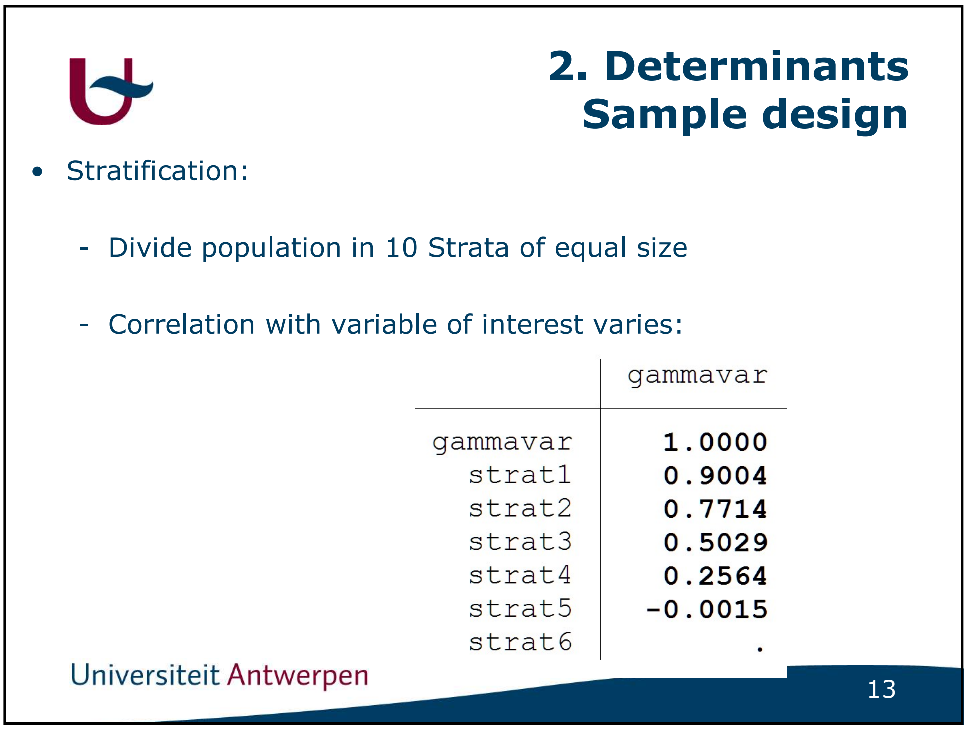
```
Variable         Obs        Mean    Std. Dev.       Min        Max

 uni3pov       1,000     .351576    .1304707       .164         .5
 uni4pov       1,000     .326492     .021048       .252         .4
 normpov       1,000     .275192    .0224979       .204       .348
gammapov       1,000     .252464    .0228446       .188        .34
```



2. Determinants
Sample design

	Population
average	1357.28
p10	303.03
p50	1114.01
p90	2737.27
Dec_Ratio	9.03
AROP60	0.29
N	10000.00

```
Variable         Obs        Mean    Std. Dev.       Min        Max

gammavar      10,000    1357.284    1035.138    .3877368    10036.7
```



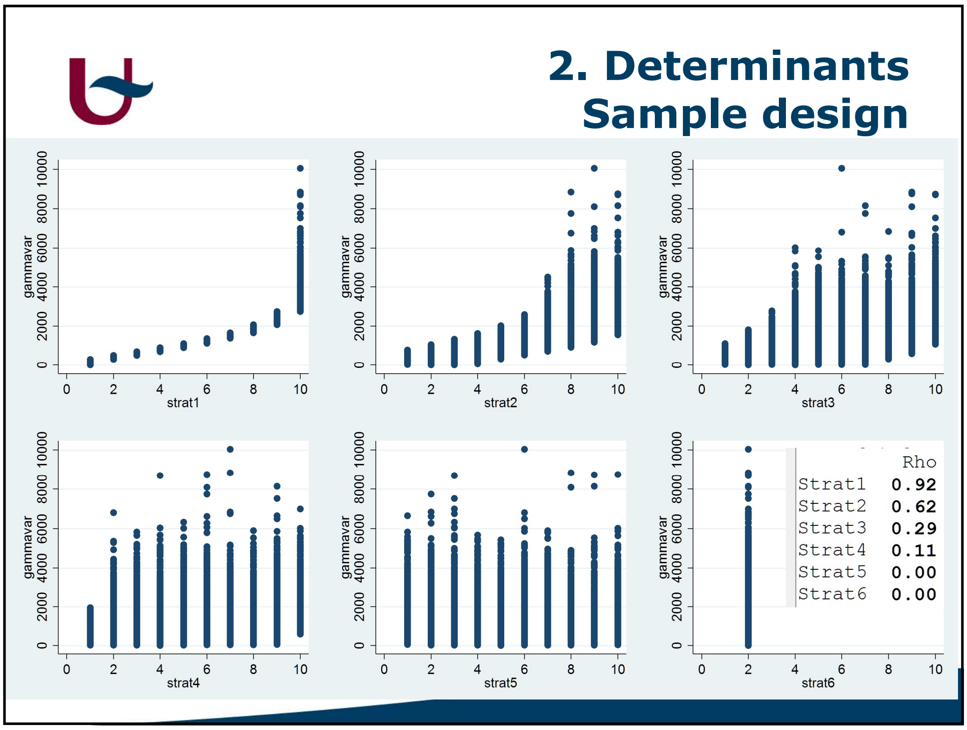
2. Determinants Sample design

- Stratification:
  - Divide population in 10 Strata of equal size
  - Correlation with variable of interest varies:

```
             |  gammavar
-------------+-----------
    gammavar |    1.0000
      strat1 |    0.9004
      strat2 |    0.7714
      strat3 |    0.5029
      strat4 |    0.2564
      strat5 |   -0.0015
      strat6 |         .
```

Universiteit Antwerpen

13



2. Determinants Sample design

	Rho
Strat1	0.92
Strat2	0.62
Strat3	0.29
Strat4	0.11
Strat5	0.00
Strat6	0.00

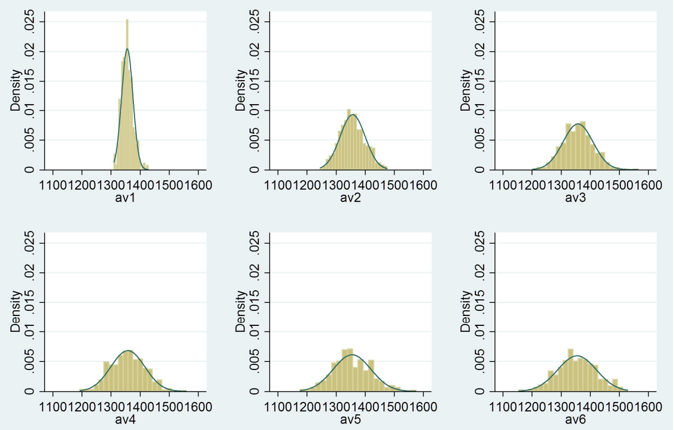
2. Determinants Sample design

- Proportional stratified random sample (with replacement):

```
if `strat'<6 {
     bsample 25, strata(strat`strat')
}
else bsample 250, strata(strat`strat')
```

- 500 samples each

2. Determinants Sample design

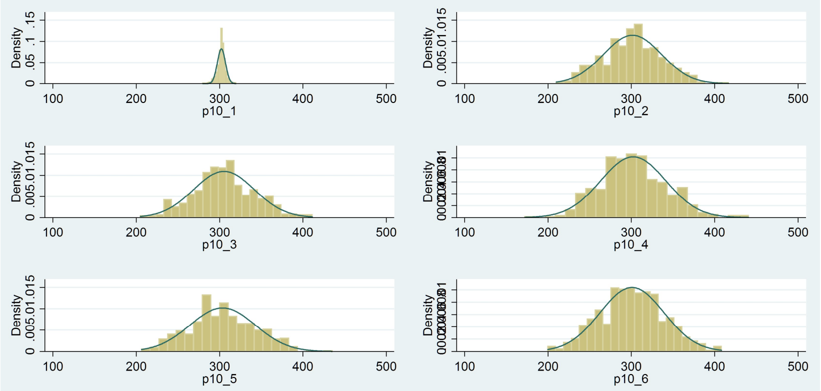


```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         av1 |        500    1356.058    19.50588   1310.552   1428.615
         av2 |        500    1356.953     42.9279   1245.744   1475.519
         av3 |        500    1358.104    51.57698   1201.169   1566.287
         av4 |        500    1357.965      58.432   1193.144   1558.686
         av5 |        500    1355.083    65.05759   1176.831   1575.077
-------------+---------------------------------------------------------
         av6 |        500    1354.736    66.99023   1154.819   1528.803
```

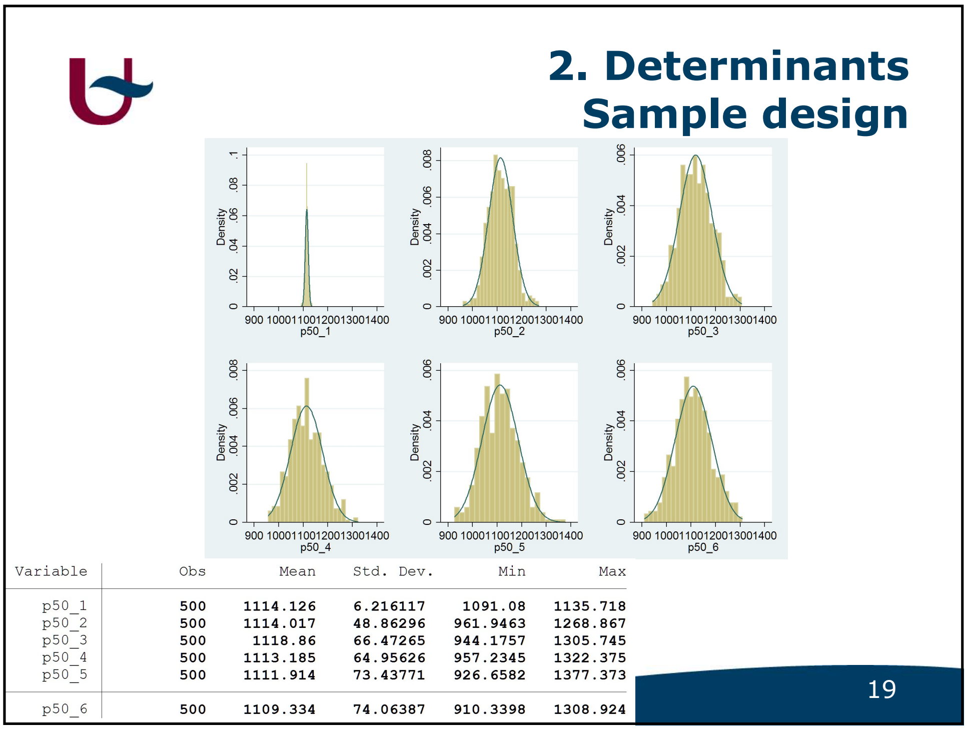
2. Determinants Sample design

- Stratification:
  - Effect depends also on statistic of interest

2. Determinants Sample design



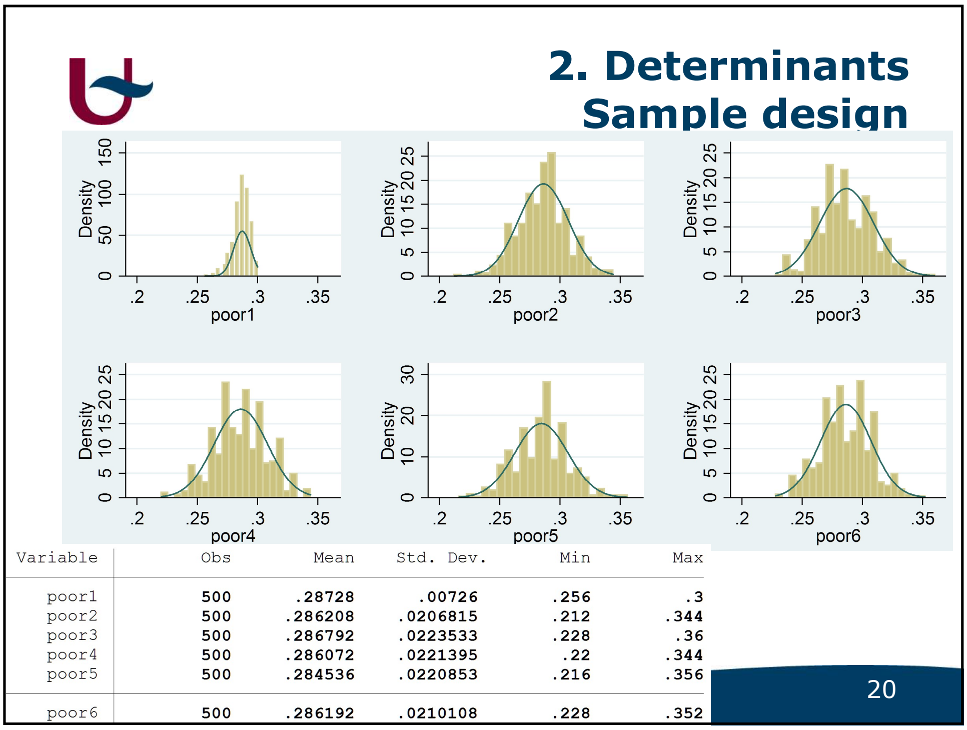
```
Variable          Obs        Mean    Std. Dev.       Min        Max

   p10_1          500    302.1921    4.832357   279.8178   319.5673
   p10_2          500    301.4527    34.81528   209.8498   416.5609
   p10_3          500    305.3007    36.54603   204.9383   411.3605
   p10_4          500    302.3185    39.22937   172.1416   440.6783
   p10_5          500    303.7435    39.07421   206.4495   435.1558

   p10_6          500    300.8718    38.36178   199.7205   408.2195
```



2. Determinants Sample design

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       p50_1 |        500    1114.126    6.216117    1091.08   1135.718
       p50_2 |        500    1114.017    48.86296   961.9463   1268.867
       p50_3 |        500     1118.86    66.47265   944.1757   1305.745
       p50_4 |        500    1113.185    64.95626   957.2345   1322.375
       p50_5 |        500    1111.914    73.43771   926.6582   1377.373
-------------+---------------------------------------------------------
       p50_6 |        500    1109.334    74.06387   910.3398   1308.924
```

19

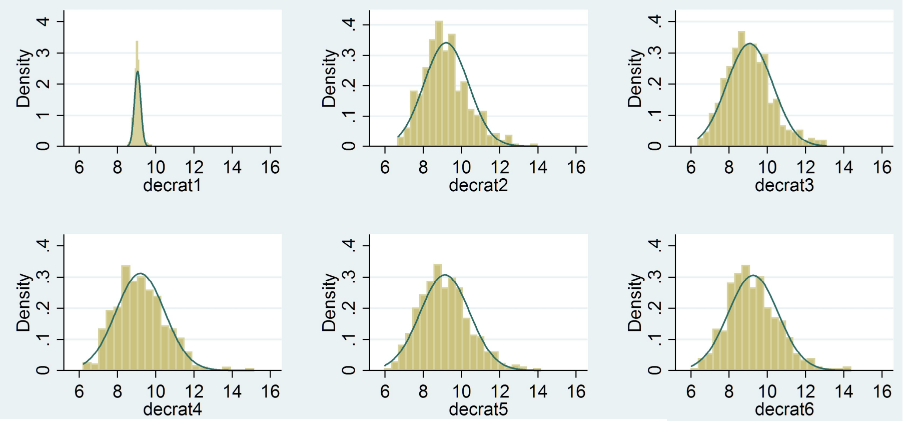


2. Determinants Sample design

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       poor1 |        500      .28728      .00726       .256         .3
       poor2 |        500     .286208    .0206815       .212       .344
       poor3 |        500     .286792    .0223533       .228        .36
       poor4 |        500     .286072    .0221395        .22       .344
       poor5 |        500     .284536    .0220853       .216       .356
-------------+---------------------------------------------------------
       poor6 |        500     .286192    .0210108       .228       .352
```

20

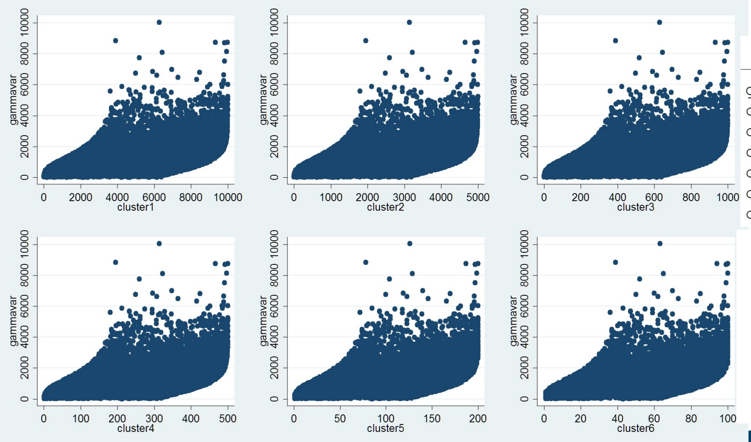
2. Determinants Sample design



```
Variable        Obs        Mean    Std. Dev.       Min        Max

 decrat1        500    9.052211    .1657317   8.537816   9.793636
 decrat2        500    9.196112    1.167475   6.674008   13.96403
 decrat3        500    9.075966    1.207079   6.370455   13.08185
 decrat4        500    9.190802    1.278179   6.210597   15.13555
 decrat5        500    9.128898    1.297242   5.974261   14.14456

 decrat6        500    9.257492    1.304941   6.026394   14.37358
```

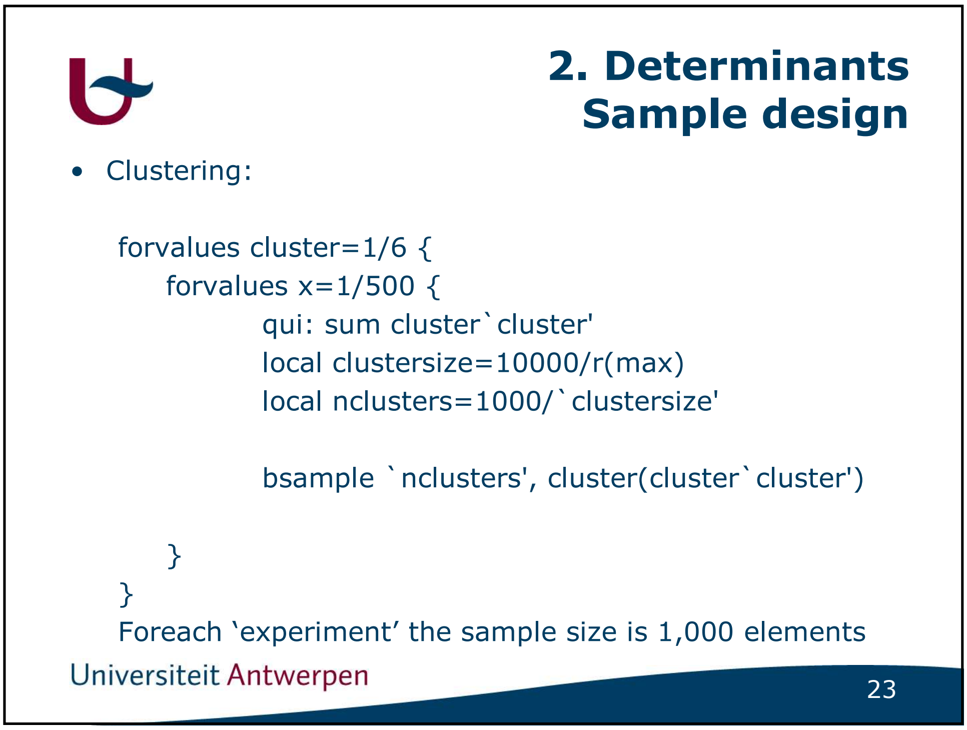
2. Determinants Sample design

- Clusters with similar correlation structure
- Range from 1 to 100 elements in size



```
            gammavar

gammavar      1.0000
cluster1      0.5068
cluster2      0.5068
cluster3      0.5068
cluster4      0.5067
cluster5      0.5067
cluster6      0.5066
```

```
 Rho   nClusters
1.00   10000.00
0.67    5000.00
0.38    1000.00
0.35     500.00
0.32     200.00
0.31     100.00
```



2. Determinants Sample design

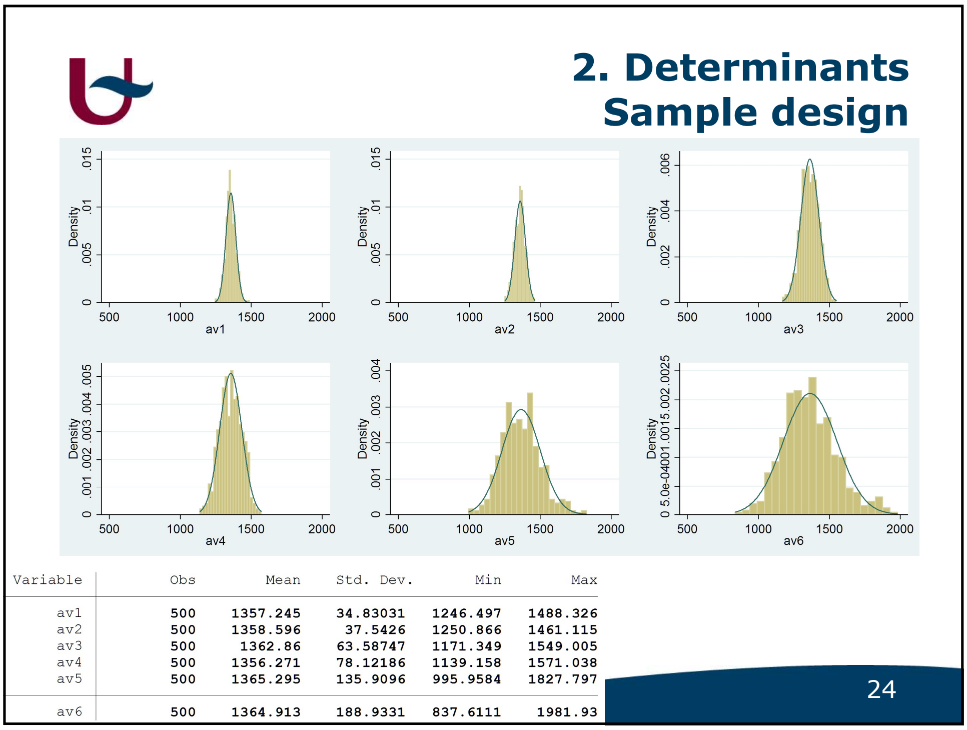
- Clustering:

```
forvalues cluster=1/6 {
    forvalues x=1/500 {
            qui: sum cluster`cluster'
            local clustersize=10000/r(max)
            local nclusters=1000/`clustersize'

            bsample `nclusters', cluster(cluster`cluster')

    }
}
```

Foreach ‘experiment’ the sample size is 1,000 elements

Universiteit Antwerpen

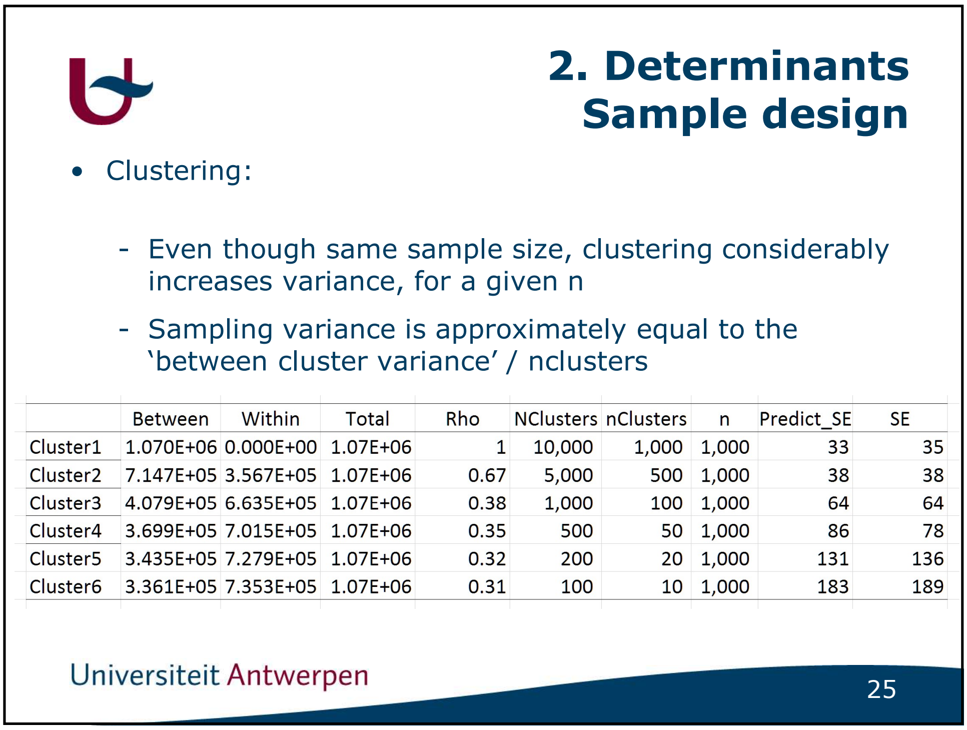
23



2. Determinants Sample design

```
Variable |        Obs        Mean    Std. Dev.       Min        Max
---------+---------------------------------------------------------
     av1 |        500    1357.245    34.83031   1246.497   1488.326
     av2 |        500    1358.596     37.5426   1250.866   1461.115
     av3 |        500     1362.86    63.58747   1171.349   1549.005
     av4 |        500    1356.271    78.12186   1139.158   1571.038
     av5 |        500    1365.295    135.9096   995.9584   1827.797
---------+---------------------------------------------------------
     av6 |        500    1364.913    188.9331   837.6111    1981.93
```

24



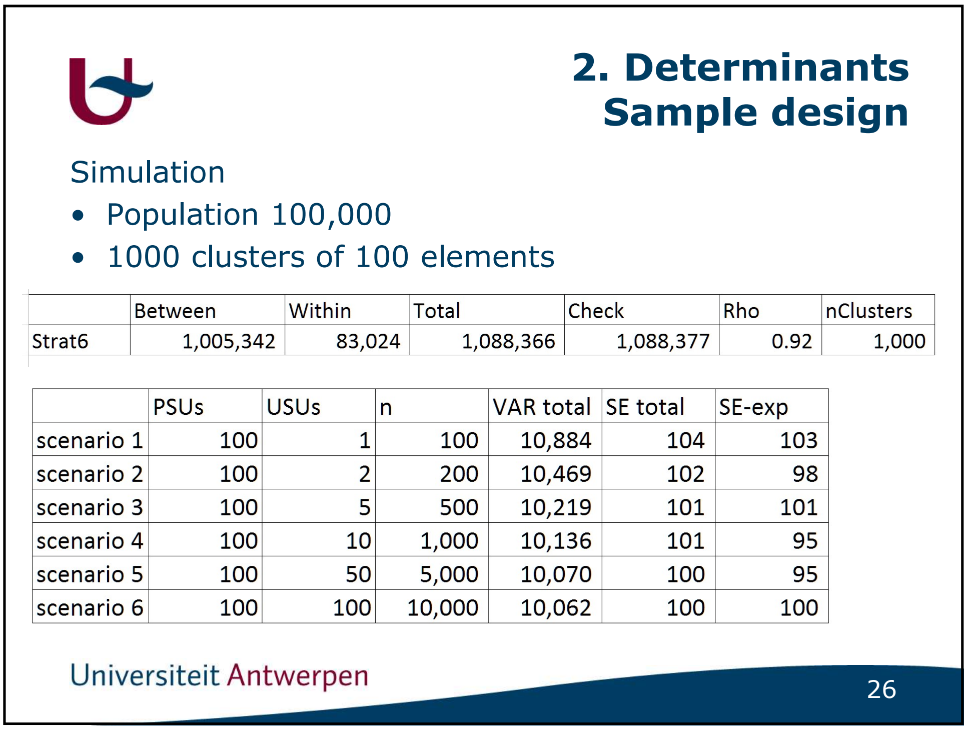
2. Determinants Sample design

- Clustering:
  - Even though same sample size, clustering considerably increases variance, for a given n
  - Sampling variance is approximately equal to the 'between cluster variance' / nclusters

	Between	Within	Total	Rho	NClusters	nClusters	n	Predict_SE	SE
Cluster1	1.070E+06	0.000E+00	1.07E+06	1	10,000	1,000	1,000	33	35
Cluster2	7.147E+05	3.567E+05	1.07E+06	0.67	5,000	500	1,000	38	38
Cluster3	4.079E+05	6.635E+05	1.07E+06	0.38	1,000	100	1,000	64	64
Cluster4	3.699E+05	7.015E+05	1.07E+06	0.35	500	50	1,000	86	78
Cluster5	3.435E+05	7.279E+05	1.07E+06	0.32	200	20	1,000	131	136
Cluster6	3.361E+05	7.353E+05	1.07E+06	0.31	100	10	1,000	183	189

Universiteit Antwerpen

25



2. Determinants Sample design

Simulation

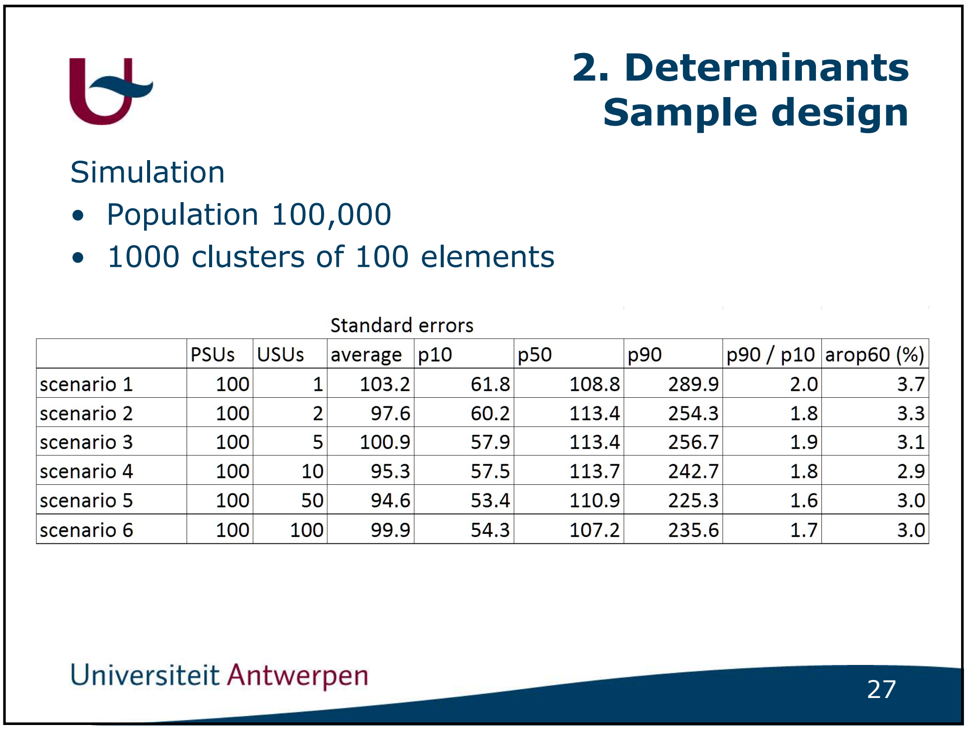
- Population 100,000
- 1000 clusters of 100 elements

	Between	Within	Total	Check	Rho	nClusters
Strat6	1,005,342	83,024	1,088,366	1,088,377	0.92	1,000

	PSUs	USUs	n	VAR total	SE total	SE-exp
scenario 1	100	1	100	10,884	104	103
scenario 2	100	2	200	10,469	102	98
scenario 3	100	5	500	10,219	101	101
scenario 4	100	10	1,000	10,136	101	95
scenario 5	100	50	5,000	10,070	100	95
scenario 6	100	100	10,000	10,062	100	100

Universiteit Antwerpen

26



2. Determinants Sample design

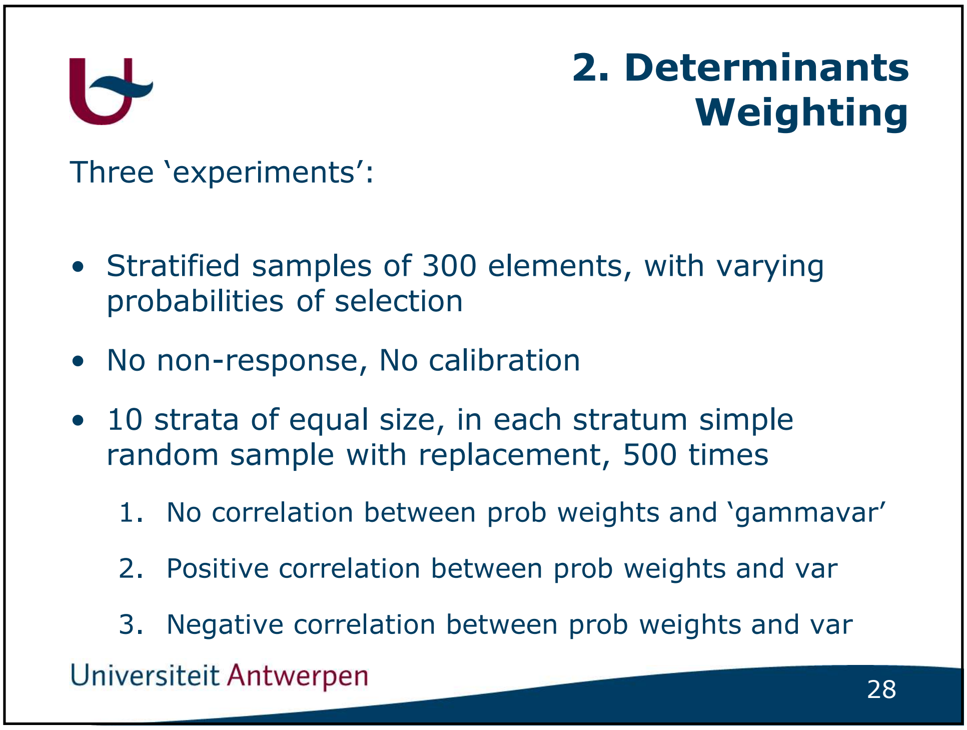
Simulation

- Population 100,000
- 1000 clusters of 100 elements

			Standard errors					
	PSUs	USUs	average	p10	p50	p90	p90 / p10	arop60 (%)
scenario 1	100	1	103.2	61.8	108.8	289.9	2.0	3.7
scenario 2	100	2	97.6	60.2	113.4	254.3	1.8	3.3
scenario 3	100	5	100.9	57.9	113.4	256.7	1.9	3.1
scenario 4	100	10	95.3	57.5	113.7	242.7	1.8	2.9
scenario 5	100	50	94.6	53.4	110.9	225.3	1.6	3.0
scenario 6	100	100	99.9	54.3	107.2	235.6	1.7	3.0

Universiteit Antwerpen

27



2. Determinants Weighting

Three ‘experiments’:

- Stratified samples of 300 elements, with varying probabilities of selection
- No non-response, No calibration
- 10 strata of equal size, in each stratum simple random sample with replacement, 500 times

  1. No correlation between prob weights and ‘gammavar’
  2. Positive correlation between prob weights and var
  3. Negative correlation between prob weights and var

Universiteit Antwerpen

28

2. Determinants Weighting

No correlation between probability of selection and variable of interest.

5 scenarios:

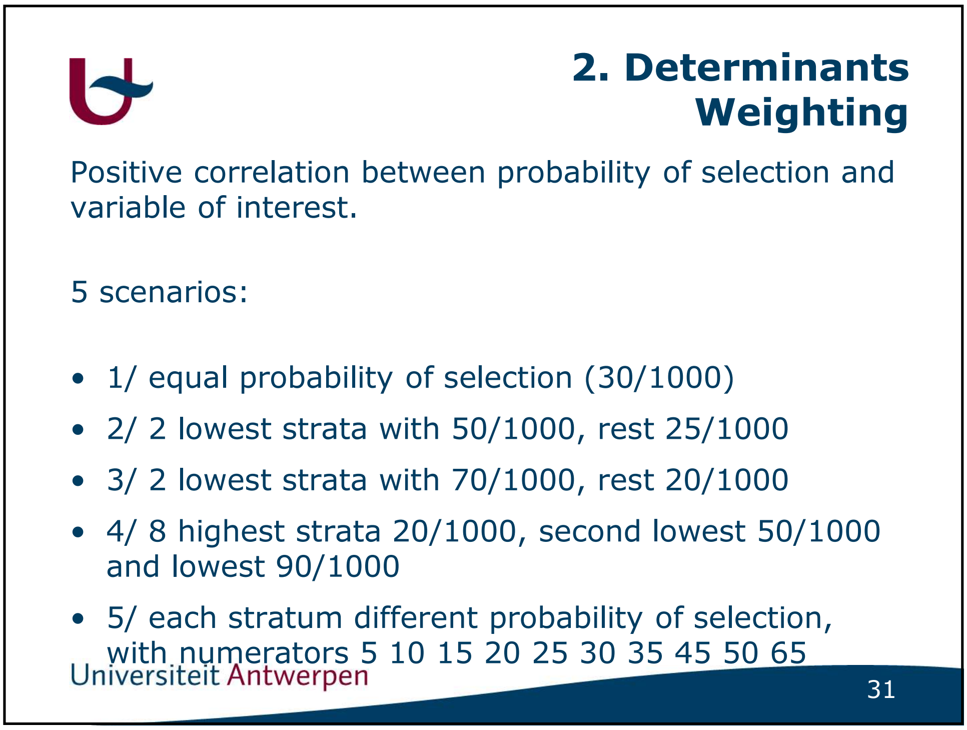
- 1/ equal probability of selection (30/1000)
- 2/ 2 strata with 50/1000, rest 25/1000
- 3/ 2 strata with 70/1000, rest 20/1000
- 4/ 8 strata 20/1000, one 50/1000 and one 90/1000
- 5/ each stratum different probability of selection, with numerators 5 10 15 20 25 30 35 45 50 65

2. Determinants Weighting

No systematic correlation between probability of selection and variable of interest.

Standard errors

	var(weights)	average	p10	p50	p90	p90 / p10	arop60 (%)
scenario 1	0.0	58.8	35.5	65.6	164.9	1.2	2.0
scenario 2	89.2	62.1	37.6	71.1	168.5	1.2	2.0
scenario 3	318.5	66.5	40.9	77.8	183.7	1.4	2.3
scenario 4	327.0	69.9	41.2	74.6	187.6	1.3	2.5
scenario 5	812.2	81.6	46.9	93.2	220.9	1.5	2.6

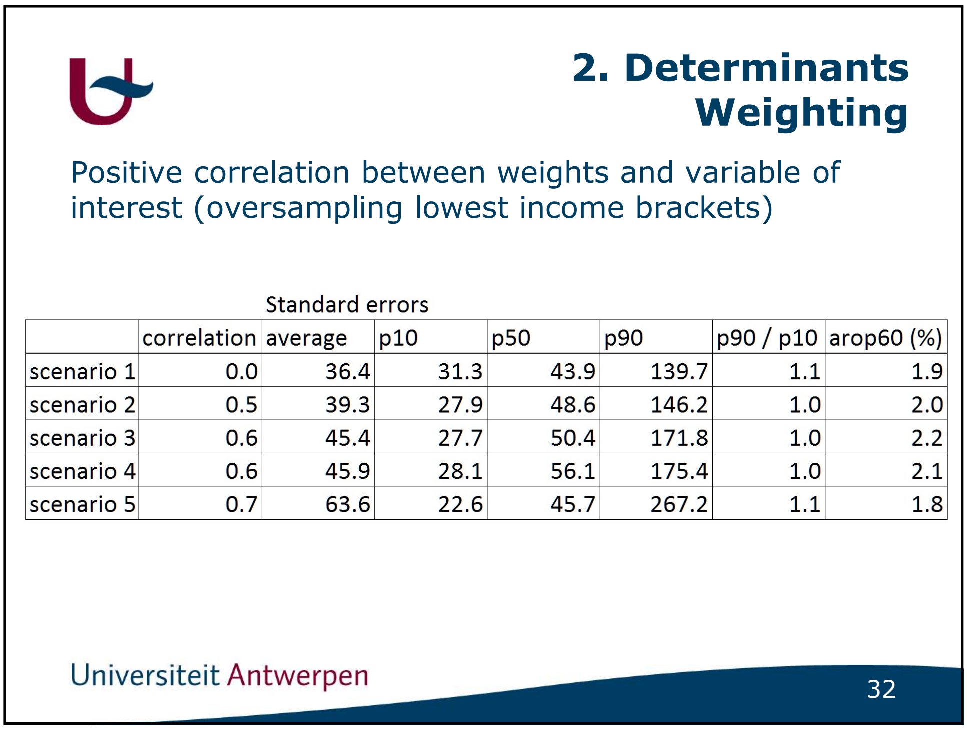


2. Determinants Weighting

Positive correlation between probability of selection and variable of interest.

5 scenarios:

- 1/ equal probability of selection (30/1000)
- 2/ 2 lowest strata with 50/1000, rest 25/1000
- 3/ 2 lowest strata with 70/1000, rest 20/1000
- 4/ 8 highest strata 20/1000, second lowest 50/1000 and lowest 90/1000
- 5/ each stratum different probability of selection, with numerators 5 10 15 20 25 30 35 45 50 65

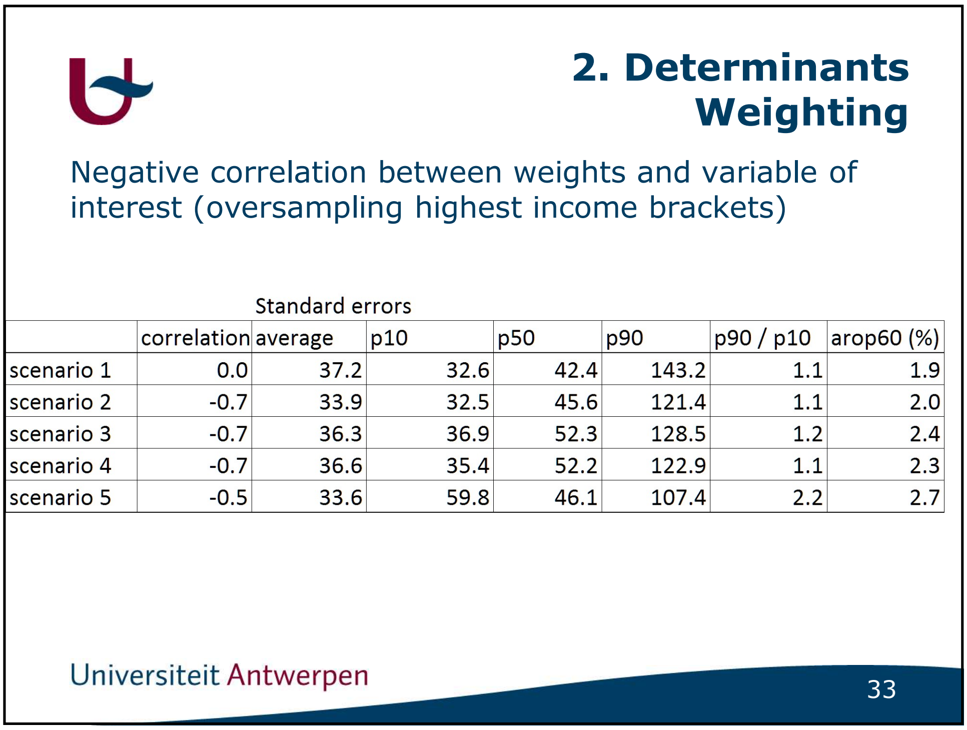


2. Determinants Weighting

Positive correlation between weights and variable of interest (oversampling lowest income brackets)

Standard errors

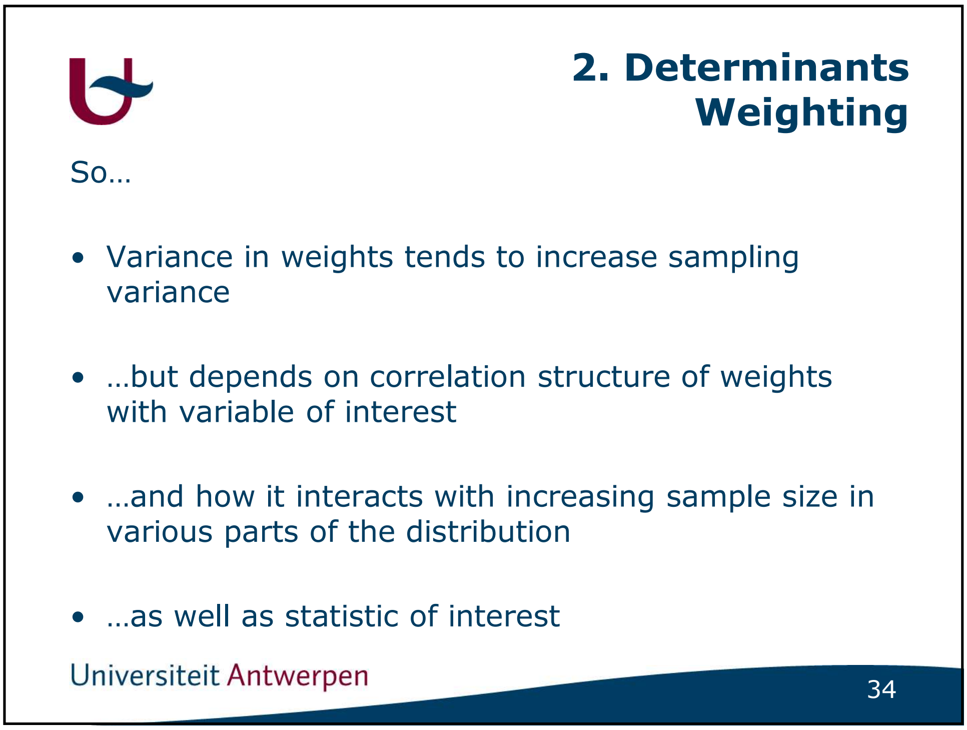
	correlation	average	p10	p50	p90	p90 / p10	arop60 (%)
scenario 1	0.0	36.4	31.3	43.9	139.7	1.1	1.9
scenario 2	0.5	39.3	27.9	48.6	146.2	1.0	2.0
scenario 3	0.6	45.4	27.7	50.4	171.8	1.0	2.2
scenario 4	0.6	45.9	28.1	56.1	175.4	1.0	2.1
scenario 5	0.7	63.6	22.6	45.7	267.2	1.1	1.8



2. Determinants Weighting

Negative correlation between weights and variable of interest (oversampling highest income brackets)

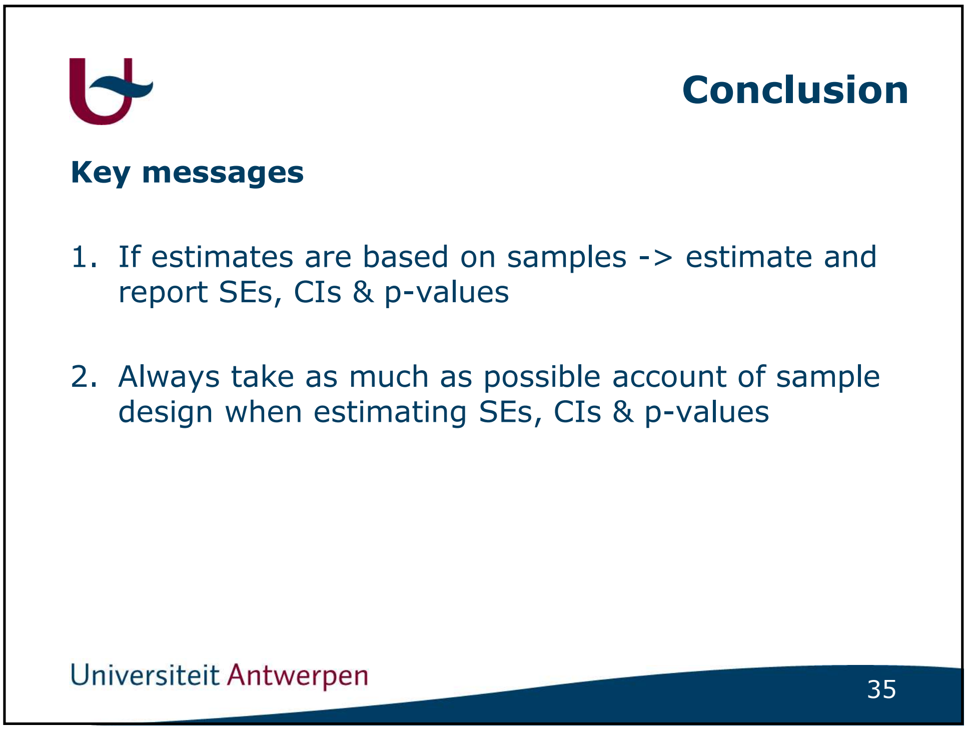
		Standard errors					
	correlation	average	p10	p50	p90	p90 / p10	arop60 (%)
scenario 1	0.0	37.2	32.6	42.4	143.2	1.1	1.9
scenario 2	-0.7	33.9	32.5	45.6	121.4	1.1	2.0
scenario 3	-0.7	36.3	36.9	52.3	128.5	1.2	2.4
scenario 4	-0.7	36.6	35.4	52.2	122.9	1.1	2.3
scenario 5	-0.5	33.6	59.8	46.1	107.4	2.2	2.7



2. Determinants Weighting

So…

- Variance in weights tends to increase sampling variance
- …but depends on correlation structure of weights with variable of interest
- …and how it interacts with increasing sample size in various parts of the distribution
- …as well as statistic of interest



Conclusion

Key messages

1. If estimates are based on samples -> estimate and report SEs, CIs & p-values

2. Always take as much as possible account of sample design when estimating SEs, CIs & p-values