How to Generate Data Set for Statistical Analysis Plan

การสร้างชุดข้อมูลสำหรับการวางแผนวิเคราะห์ทางสถิติ

Kavin Thinkhrmrop

M.P.H. and Dr.P.H., Biostatistics

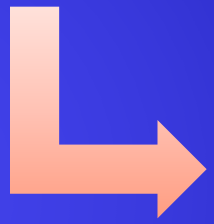
Categorical data (1)

Initial command

```
set obs n    --> n = records number
gen id = _n --> running number
```

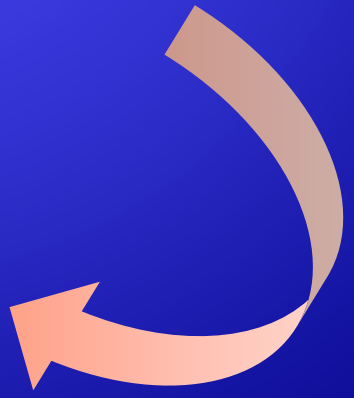
Categorical data (2)

```
gen vname1 = uniform()
```



vname1
.3091256
.0403894
.7973289
.5346149

```
gen vname2 = round(uniform())
```



vname1	vname2
.3091256	0
.0403894	0
.7973289	1
.5346149	1

Categorical data (3)

```
gen vname3 = 1+int(5*uniform())
```

```
vname3
   2
   1
   4
   3
   5
```

Continuous data (1)

```
gen vname = a+(b-a)*uniform()
```

Data will be between a – b

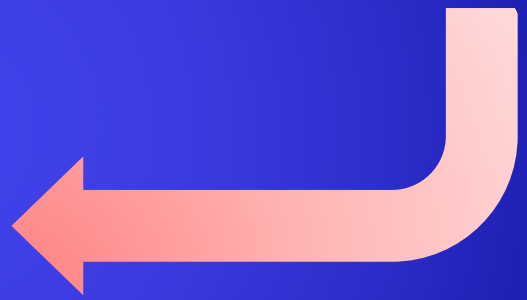
```
gen age1 = 20+(60-20)*uniform()
```

age1
32.36502
21.61558
51.89316
41.38459
55.39351

Continuous data (2)

```
gen age2 = 20+int((60-20+1)*uniform())
```

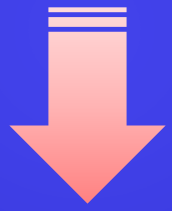
age1	age2
32.36502	32
21.61558	21
51.89316	52
41.38459	41
55.39351	56



Continuous data (3)

```
gen vname = int(mu+sigma*invnorm(uniform()))

gen age = int(28+5*invnorm(uniform()))
```



```
Variable |       Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
     age |      1000      27.376    4.742058        13         42
```

Continuous data (4)

```
gen weight = round(65+10*invnorm(uniform()),.1)
```

```
weight
 69.6
 75.2
 68.8
 71.6
 65.5
```

```
Variable |       Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
  weight |      1000     65.0579    9.862147      32.8      102.2
```

Generate data set (1)

Data dictionary

No.	V name	V description	Code and values of variables
1.	id	Running number	1 – 1,000
2.	gender	Gender of sample	1 = Male; 2 = Female
3.	age	Age of sample in year	Number of age between 20-60 years
4.	weight	Body weight in kilograms	Number of weight with 1 decimal digit
5.	mstatus	Marital status	1 = Single; 2 = Married; 3 = Withdraw
6.	occ	Main occupation at present	0 = Unemployed; 1 = Famer; 2 = Own business; 3 = Gov. official/State ent.; 4 = Others
7.	ov	Infected by liver fluke	0 = Negative; 1 = Positive
8.	cca	Bile duct cancer diagnosed	0 = Negative; 1 = Positive

Generate data set (2)

Data structure

id	gender	age	weight	mstatus	occ	ov	cca
1	1	32	54.5	1	1	0	1
2	1	68	79.3	2	3	1	1
3	2	47	85.7	2	0	1	0
4	1	41	49.8	2	2	0	1
5	2	38	63	3	1	1	0
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1000	2	59	98.2	1	4	0	1

Generate data set (3)

STATA Command

```
(1) set obs 1000
(2) gen id = _n
(3) gen gender = 1+int(2*uniform())
(4) gen age = 20+int((60-20+1)*uniform())
(5) gen weight = round(65+10*invnorm(uniform()),.1)
(6) gen mstatus = 1+int(3*uniform())
(7) gen occ = 0+int(5*uniform())
(8) gen ov = round(uniform())
(9) gen cca = round(uniform())
```

Generate data set (4)

List of variables

id	gender	age	weight	mstatus	occ	ov	cca
1	1	56	70.4	3	4	0	1
2	2	37	51.4	2	1	1	1
3	2	60	45.1	2	0	1	0
4	2	60	59.5	3	3	0	1
5	2	44	71.7	1	4	0	1
6	1	39	50.0	3	2	1	1
7	2	24	77.0	1	1	1	1
8	1	28	67.5	1	4	1	0
9	2	25	73.6	3	2	0	1
10	2	40	75.7	1	0	0	0

Generate data set (5)

Summarize

Variable	Obs	Mean	Std. Dev.	Min	Max
id	1000	500.5	288.8194	1	1000
gender	1000	1.514	.5000541	1	2
age	1000	40.27	11.76042	20	60
weight	1000	65.3848	9.849329	38.4	97.6
mstatus	1000	1.964	.8120119	1	3
occ	1000	3.016	1.399894	0	4
ov	1000	.493	.5002012	0	1
cca	1000	.509	.5001691	0	1

Thanks