

## 11.2 NOTE&SYNTAX:

### **Power Analysis: sampsi -- Sample size and power determination**

What is Power?  $\alpha$ ?

Alpha is the probability that I reject  $H_0$  when  $H_0$  is true, for example:

$H_0$ : the color of the slip is pink (actually it is a pink slip)

Smaller better.

Power  $\pi(\theta)$  is the probability that I reject  $H_0$  when  $H_0$  is not true, for example:

$H_0$ : the color of the slip is pink (actually it is a yellow slip)

Larger better.

Two questions we always ask when doing Power Analysis:

- 1) What is the required sample size if I want to compare two groups at this alpha level? And
- 2) How much power I can get if I have this sample size.

### **Example 1: one-sample test of proportion**

In our class, we have 7 female students, and 13 male students. If my null hypothesis is:

$H_0$ : proportion of female students = .5

And we already know that our true proportion of female students is .35.

**1<sup>st</sup> question:** how many sample size I need to have such that my power=.8 (which means the probability that I reject  $H_0$  would be .8 when the true proportion of female students is .35) and alpha=.05?

```
. sampsi .5 .35, power(.8) onesample
```

Estimated sample size for one-sample comparison of proportion to hypothesized value

Test  $H_0$ :  $p = 0.5000$ , where  $p$  is the proportion in the population

Assumptions:

```
alpha = 0.0500 (two-sided)
power = 0.8000
alternative p = 0.3500
```

Estimated required sample size:

```
n = 85
```

**Interpretation:** from above results, we know at least 85 students I need to be able to achieve the goal (with power=.8 and alpha=.05).

**2<sup>nd</sup> question:** now we have sample size ( $n=20$ ), how much power I can get?

```
. sampsi .5 .35,n(20) onesample
```

Estimated power for one-sample comparison of proportion to hypothesized value

Test Ho:  $p = 0.5000$ , where  $p$  is the proportion in the population

Assumptions:

```
alpha = 0.0500 (two-sided)
alternative p = 0.3500
sample size n = 20
```

Estimated power:

```
power = 0.2587
```

**Interpretation:** for this small sample size  $n=20$ , I may only have  $\text{power}=.26$ , which means, if the true proportion of female students is  $.35$ , there is only around  $.26$  probability I will reject the statement saying the proportion of female students is  $.5$  (that would be really poor).

### Example 2: two-sample test of proportion

Suppose using past surveys as a guide, we estimate that 30% male students smoke, and only 25% female students smoke.

**1<sup>st</sup> question:** we want to know the required sample sizes for both male and female students for the test with  $\alpha=.05$ (two-sided) and the power of  $.80$ :

```
. sampsi .3 .2, alpha(.05) power(.8)
```

Estimated sample size for two-sample comparison of proportions

Test Ho:  $p_1 = p_2$ , where  $p_1$  is the proportion in population 1  
and  $p_2$  is the proportion in population 2

Assumptions:

```
alpha = 0.0500 (two-sided)
power = 0.8000
p1 = 0.3000
p2 = 0.2000
n2/n1 = 1.00
```

Estimated required sample sizes:

```
n1 = 313
n2 = 313
```

**2<sup>nd</sup> question:** If we have a sample of male students ( $n=300$ ), and female students ( $n=250$ ), how much power we can get at  $\alpha=.01$ ?

```
. sampsi .3 .2,n1(300) n2(250) alpha(.01)
```

Estimated power for two-sample comparison of proportions

Test Ho:  $p_1 = p_2$ , where  $p_1$  is the proportion in population 1  
and  $p_2$  is the proportion in population 2

Assumptions:

```
alpha = 0.0100 (two-sided)
p1 = 0.3000
p2 = 0.2000
sample size n1 = 300
n2 = 250
n2/n1 = 0.83
```

Estimated power:

power = 0.5027

### Example 3: one-sample test of mean

Suppose we wish to test the effects of a low-fat diet on people's weight. We measure the difference in weight for each observation before and after being on the diet. Our null hypothesis  $H_0$ : the diet does not have effect at all. From past studies, we also estimate that the sd of the difference in weight levels will be about 6 pounds.

**1<sup>st</sup> question:** Now if the low-fat diet does have effect as much as will reduce weight by 10 lbs. We want to know the required sample size for the test with  $\alpha=.01$ , one-sided test, with  $\text{power}=.90$ :

```
. sampsi 0 -10,sd(6) a(.01) onesided p(.95) onesample
```

Estimated sample size for one-sample comparison of mean  
to hypothesized value

Test  $H_0$ :  $m = 0$ , where  $m$  is the mean in the population

Assumptions:

```
alpha = 0.0100 (one-sided)
power = 0.9500
alternative m = -10
sd = 6
```

Estimated required sample size:

```
n = 6
```

**2<sup>nd</sup> question:** Now if we decide to conduct the study with  $n=10$ , what would be the power we can get at a one-sided significance level of  $\alpha=.01$ ?

```
. sampsi 0 -10,sd(6) a(.01) onesided n(10) onesample
```

Estimated power for one-sample comparison of mean  
to hypothesized value

Test  $H_0$ :  $m = 0$ , where  $m$  is the mean in the population

Assumptions:

```
alpha = 0.0100 (one-sided)
alternative m = -10
sd = 6
sample size n = 10
```

Estimated power:

```
power = 0.9984
```

### Example 4: two-sample test of mean

```
. sysuse auto,clear
(1978 Automobile Data)
```

```
. bysort foreign:sum weight
```

```
-> foreign = Domestic
```

Variable	Obs	Mean	Std. Dev.	Min	Max
weight	52	3317.115	695.3637	1800	4840

```
-> foreign = Foreign
```

Variable	Obs	Mean	Std. Dev.	Min	Max
weight	22	2315.909	433.0035	1760	3420

Suppose, from pilot study (such as from 1978 Auto dataset), it was reported that the mean and standard deviation of weight of foreign vehicles were 2300 and 433 lbs, respectively. The mean and sd of weight of domestic vehicles were 3300 and 700 lbs, respectively. Since it is easier to find domestic vehicles than imported vehicles, we decide that  $n_2$ , the sample size of domestic cars, should be twice  $n_1$ , the size of the sample of foreign cars; that is,  $r=n_2/n_1=2$ .

**1<sup>st</sup> question:** to compute the sample sizes for  $\alpha=.05$ (two-sided) and the power of .99, we use:

```
. sampsi 2300 3300,sd1(430) sd2(700) r(2) p(.99)
```

Estimated sample size for two-sample comparison of means

Test Ho:  $m_1 = m_2$ , where  $m_1$  is the mean in population 1  
and  $m_2$  is the mean in population 2

Assumptions:

```
alpha = 0.0500 (two-sided)
power = 0.9900
m1 = 2300
m2 = 3300
sd1 = 430
sd2 = 700
n2/n1 = 2.00
```

Estimated required sample sizes:

```
n1 = 8
n2 = 16
```

**2<sup>nd</sup> question:** if we now have sample size  $n_1=n_2=10$ , what would be the power we can get?

```
. sampsi 2300 3300,sd1(430) sd2(700) n1(10) n2(10)
```

Estimated power for two-sample comparison of means

Test Ho:  $m_1 = m_2$ , where  $m_1$  is the mean in population 1  
and  $m_2$  is the mean in population 2

Assumptions:

```
alpha = 0.0500 (two-sided)
m1 = 2300
m2 = 3300
sd1 = 430
```

```

          sd2 =          700
sample size n1 =          10
          n2 =          10
          n2/n1 =         1.00

```

Estimated power:

```

power =    0.9706

```

### Chi-square test:

In Stata, chi-square tests are options for the command tabulate. One application is Chi-square test of Independence. It is used when we want to know whether there is an association between two categorical variables (typically on 2\*2 tables, but they can also work on 2\*k tables). The null hypothesis that the two variables are independent.

```

. tab var1 var2,chi2

```

For example, if we replace mpg1=1 if mpg>mean (21.3), and mpg1=0 if mpg<=21.3, and we want to know whether mpg is associated with foreign:

```

. sysuse auto,clear
(1978 Automobile Data)

. gen mpg1=mpg>21.3

. tab mpg1 foreign,chi2

```

mpg1	Car type		Total
	Domestic	Foreign	
0	36	7	43
1	16	15	31
Total	52	22	74

```

Pearson chi2(1) =    8.8892    Pr = 0.003

```

These results indicate that there is statistically significant relationship between mpg and foreign (chi-square with one degree of freedom=8.889, p=.003)

**\*\*Note:** Chi-square test assumes the expected value of each cell is 5 or higher, if this assumption is violated, you then need to use Fisher's exact test below:  
option: exact\*\*

```

. tab var1 var2,chi2 exact

```

For our particular dataset, this assumption is met (all cell>5).

### \*\*\*ANOVA\*\*\*

#### **Situation 1: only one independent variable**

```

. /***one-way ANOVA***/
. sysuse auto,clear
(1978 Automobile Data)

. oneway mpg foreign

```

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	378.153515	1	378.153515	13.18	0.0005
Within groups	2065.30594	72	28.6848048		
Total	2443.45946	73	33.4720474		

Bartlett's test for equal variances:  $\chi^2(1) = 3.4818$  Prob> $\chi^2 = 0.062$

. anova mpg foreign

Number of obs = 74 R-squared = 0.1548  
 Root MSE = 5.35582 Adj R-squared = 0.1430

Source	Partial SS	df	MS	F	Prob > F
Model	378.153515	1	378.153515	13.18	0.0005
foreign	378.153515	1	378.153515	13.18	0.0005
Residual	2065.30594	72	28.6848048		
Total	2443.45946	73	33.4720474		

Two tables present almost the same information. They both tell you how much variance of "mpg" can be explained by factor "foreign".

**Situation 2: more than one independent variables.**

. anova mpg foreign price weight

Number of obs = 74 R-squared = 1.0000  
 Root MSE = 0 Adj R-squared =

Source	Partial SS	df	MS	F	Prob > F
Model	2443.45946	73	33.4720474		
foreign	24.5	1	24.5		
price	2065.30594	72	28.6848048		
weight	0	0			
Residual	0	0			
Total	2443.45946	73	33.4720474		

If you have continuous independent variables, you need to tell Stata by using option "cont(var1 var2)", otherwise you might get wrong results:

. anova mpg foreign price weight,cont(price weight)

Number of obs = 74 R-squared = 0.6631  
 Root MSE = 3.42919 Adj R-squared = 0.6487

Source	Partial SS	df	MS	F	Prob > F
Model	1620.30716	3	540.102388	45.93	0.0000
foreign	24.3746721	1	24.3746721	2.07	0.1544
price	1.01946674	1	1.01946674	0.09	0.7693
weight	659.44086	1	659.44086	56.08	0.0000

Residual		823.152295	70	11.7593185
-----				
Total		2443.45946	73	33.4720474

. anova mpg foreign price weight,cont(price weight) reg

Source	SS	df	MS	Number of obs =	74
Model	1620.30716	3	540.102388	F( 3, 70) =	45.93
Residual	823.152295	70	11.7593185	Prob > F =	0.0000
				R-squared =	0.6631
				Adj R-squared =	0.6487
Total	2443.45946	73	33.4720474	Root MSE =	3.4292

mpg	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_cons	40.10359	1.674213	23.95	0.000	36.76447 43.4427
foreign					
1	1.855891	1.289063	1.44	0.154	-.7150641 4.426846
2	(dropped)				
price	.0000566	.0001922	0.29	0.769	-.0003268 .00044
weight	-.0067758	.0009048	-7.49	0.000	-.0085805 -.0049712

. reg mpg foreign price weight

Source	SS	df	MS	Number of obs =	74
Model	1620.30716	3	540.102388	F( 3, 70) =	45.93
Residual	823.152295	70	11.7593185	Prob > F =	0.0000
				R-squared =	0.6631
				Adj R-squared =	0.6487
Total	2443.45946	73	33.4720474	Root MSE =	3.4292

mpg	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
foreign	-1.855891	1.289063	-1.44	0.154	-4.426846 .7150641
price	.0000566	.0001922	0.29	0.769	-.0003268 .00044
weight	-.0067758	.0009048	-7.49	0.000	-.0085805 -.0049712
_cons	41.95948	2.377726	17.65	0.000	37.21725 46.7017

. \*\*Note: the above two commands produce almost a same result\*\*

Question 1: How can I plot ANOVA cell means in Stata?

. use <http://www.ats.ucla.edu/stat/stata/faq/crf24>,clear  
(CRF Example - Kirk, 1st Edition)

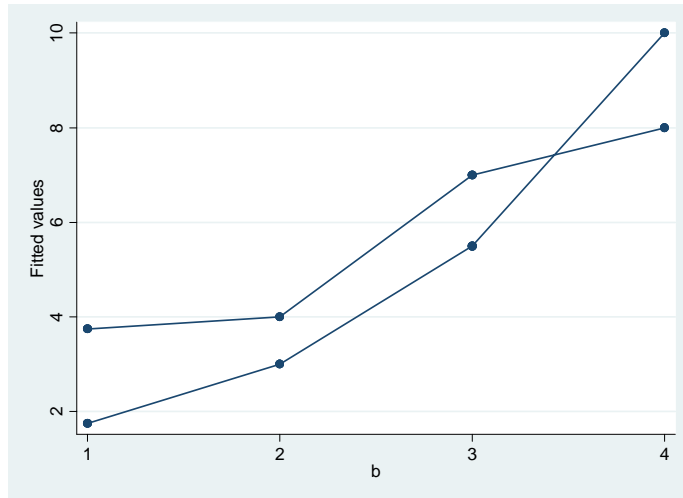
. anova y a b a\*b

Source	Partial SS	df	MS	F	Prob > F
Model	217	7	31	40.22	0.0000
a	3.125	1	3.125	4.05	0.0554
b	194.5	3	64.8333333	84.11	0.0000
a*b	19.375	3	6.45833333	8.38	0.0006

Number of obs = 32      R-squared = 0.9214  
Root MSE = .877971      Adj R-squared = 0.8985

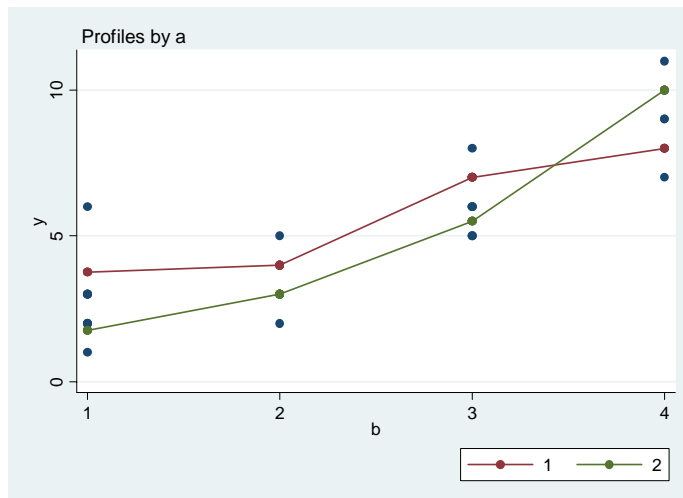
Residual	18.5	24	.770833333
Total	235.5	31	7.59677419

```
. predict yhat
(option xb assumed; fitted values)
. sort a b
. graph twoway scatter yhat b, connect(L)
```

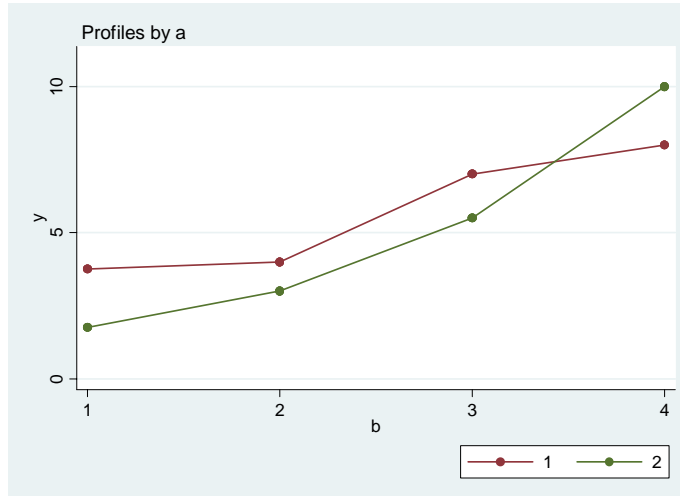


Or, you may use Stata add-on program "anovaplot":

```
. findit gr0009.pkg /*and install this program*/
. anovaplot b a
```



```
. anovaplot b a, scatter(msym(i)) /**use option "scatter(ms(i))" to suppress
plotting of response data***/
```



**Question 2: How to do ANOVA contrasts in Stata?**

```
. sysuse auto,clear
. tab rep78
```

Repair Record 1978	Freq.	Percent	Cum.
3	30	50.85	50.85
4	18	30.51	81.36
5	11	18.64	100.00
Total	59	100.00	

```
. anova mpg rep78
```

Number of obs = 59      R-squared = 0.2322  
 Root MSE = 5.46729      Adj R-squared = 0.2048

Source	Partial SS	df	MS	F	Prob > F
Model	506.325167	2	253.162583	8.47	0.0006
rep78	506.325167	2	253.162583	8.47	0.0006
Residual	1673.91212	56	29.8912879		
Total	2180.23729	58	37.5902981		

From above results, Stata only told you that rep78 has significant overall effect on vehicle's mpg, but it didn't tell you the effect contrasts between rep78 categories (group 1: rep78=3; group 2: rep78=4; and group 3: rep78=5). Stata does not have built-in contrast command, but there is an add-on command "anovacontrast" developed by UCLA Academic Technology Services:

```
. net install anovacontrast.pkg
```

```
. keep if rep78>2&rep78<.
(15 observations deleted)
```

NOTE: anovacontrast gives incorrect and misleading results if there are missing or empty cells.

**. anova mpg rep78**

Number of obs = 59 R-squared = 0.2322  
 Root MSE = 5.46729 Adj R-squared = 0.2048

Source	Partial SS	df	MS	F	Prob > F
Model	506.325167	2	253.162583	8.47	0.0006
rep78	506.325167	2	253.162583	8.47	0.0006
Residual	1673.91212	56	29.8912879		
Total	2180.23729	58	37.5902981		

**. table rep78,contents(mean mpg)**

Repair Record	mean(mpg)
1978	
3	19.4333
4	21.6667
5	27.3636

If we want to do group 1 and group 3 contrast:

**. anovacontrast rep78, values(-1 0 1)**

/\*\*group 1 versus group 3\*\*values(numlist) specifies the weights for the  
 > contrast and is not optional\*\*\*/

Contrast variable: rep78 (-1 0 1) Dep Var: mpg					
source	SS	df	MS	Contrast =	7.93
contrast	506.185509	1	506.1855	N of obs =	59
error	1673.91212	56	29.8913	F =	16.93
				Prob > F =	0.0001
				t =	4.12

p-value for above contrast is .0001, which means group 1 vs. group 3 of rep78 has significant different effect on vehicle's mpg.

Now, if we want to do group 2 and group 3 contrast:

**. anovacontrast rep78, values(0 -1 1)**

Contrast variable: rep78 (0 -1 1) Dep Var: mpg					
source	SS	df	MS	Contrast =	5.70
contrast	221.592499	1	221.5925	N of obs =	59
error	1673.91212	56	29.8913	F =	7.41
				Prob > F =	0.0086
				t =	2.72

The difference between group 2 and group 3 is also very significant, p-value=.0086

Finally we are interested in the contrast between group 1 and group 2, but before to do that, from below table, we know the means difference between group 1 and 2 is very small (21.67-19.43=2.24), thus we expect that the contrast between group 1 and 2 may be not significant.

```
. table rep78,contents(mean mpg)
```

Repair Record 1978	mean(mpg)
3	19.4333
4	21.6667
5	27.3636

```
. anovacontrast rep78, values(-1 1 0)
```

```
Contrast variable: rep78 (-1 1 0) Dep Var: mpg
source          SS          df          MS          Contrast =      2.23
-----+-----
contrast | 56.1125128          1          56.1125          F          =      1.88
error    | 1673.91212          56          29.8913          Prob > F    =    0.1761
-----+-----
t          =      1.37
```

From above results (p-value=.176), as we expected, the contrast between group 1 and 2 is not significant.