

Curs bàsic d'Anàlisi de dades amb Stata

1

Sessió 2

- Estadística descriptiva
 - Estadístics descriptius
 - Mitjanes, Medianes, Intervals de Confiança, Percentils
 - Gràfics descriptius variables quantitatives
 - Taules de contingència
 - Taules epidemiològiques (incidència prevalença)
 - Gràfics descriptius variables qualitatives
- Grandària Mostral
 - Càlcul de la grandària mostral
 - Càlcul del poder
 - Generació de nombres aleatoris
- Exercici Pràctic

2

Estadística Descriptiva

3

Variables cuantitativas

	Paramètrics	No paramètrics
Localización	Media	Moda , Mediana
Dispersión	Varianza Desviación típica	Rango (min/max) Intervalo Interquartílico ($P_{25} - P_{75}$)
Asimetría	Asimètria	--
Gràfics	Media (Intervalo de confianza)	Diagrama de cajas, Diagrama de puntos, Histograma

4

Variables cuantitativas

- Descripción de variables cuantitativas

```
summarize  variables
```

by *vargrupo*, **sort: summarize** *variables*

```

summarize edadsero

  Variable |      Obs      Mean    Std. Dev.      Min      Max
-----+-----
  edadsero |      444    26.50658    5.866372    13.94795    50.27869

by sex, sort:summarize edadsero

-> sex = male
  Variable |      Obs      Mean    Std. Dev.      Min      Max
-----+-----
  edadsero |      350    26.89573    6.025914    13.94795    50.27869

-> sex = female
  Variable |      Obs      Mean    Std. Dev.      Min      Max
-----+-----
  edadsero |       94    25.05758    4.997224    16.68969    38.8356

```

Variables cuantitativas

- Descripción de variables cuantitativas

```
mean    variables, over(vargrupo)
```

```
. mean datapos, over(expcateg)
```

Mean estimation

Number of obs = 417

homo: expcateg = homo
idu: expcateg = idu

	Over	Mean	Std. Err.	[95% Conf. Interval]
datapos				
	homo	1993.679	.3803832	1992.932 1994.427
	idu	1993.352	.1862981	1992.986 1993.718

Variables cuantitativas

- Descripción de variables cuantitativas

ameans *variables*, **add(#)**

. amean **edadsero**

Variable	Type	Obs	Mean	[95% Conf. Interval]	
-----+-----					
edadsero	Arithmetic	444	26.50658	25.95942	7.05374
	Geometric	444	25.90548	25.39713	26.424
	Harmonic	444	25.33755	24.85023	5.84438
-----+-----					

7

Variables cuantitativas

- Calculo de estadísticos de variable en función de otra

tabstat *variables*, **by(vargrupo)** **statistics(lista)**

tabstat **edadsero**, **by(sex)** **statistics(count mean median min max semean)**

Summary for variables: edadsero
by categories of: sex (sexe)

sex	N	mean	p50	min	max	se(mean)
-----+-----						
male	327	26.66295	25.67671	13.94795	50.27869	.3273748
female	88	25.03693	24.51507	16.68969	38.83562	.5338185
-----+-----						
Total	415	26.31815	25.44763	13.94795	50.27869	.2833073
-----+-----						

8

Variables cuantitativas

- Calculo de intervalos de confianza

```
ci variables, normal (poisson) (binomial)
cii denominador numerador, poisson
cii denominador numerador , binomial
cii n media desviacion , normal
```

```
. ci edadsero
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
edadsero	415	26.31815	.2833073	25.76125	26.87505

```
. cii 1000 4, poisson
```

Variable	Exposure	Mean	Std. Err.	-- Poisson Exact -- [95% Conf. Interval]	
	1000	.004	.002	.0010899	.0102416

```
. cii 100 10, binomial
```

Variable	Obs	Mean	Std. Err.	-- Binomial Exact -- [95% Conf. Interval]	
	100	.1	.03	.0490047	.176222

Variables cuantitativas

- Calculo de percentiles

```
centile variables, c(5 10 25 50 75 90 95)
```

```
. centile edadsero, c(5 10 25 50 75 90 95)
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf. Interval]	
edadsero	415	5	18.28884	17.70245	18.96871
		10	19.87397	18.96774	20.64329
		25	22.60274	21.81362	23.32044
		50	25.44763	24.96491	26.14509
		75	28.95628	28.25659	29.87559
		90	34.17821	32.42179	35.35978
		95	37.12159	35.35755	39.54065

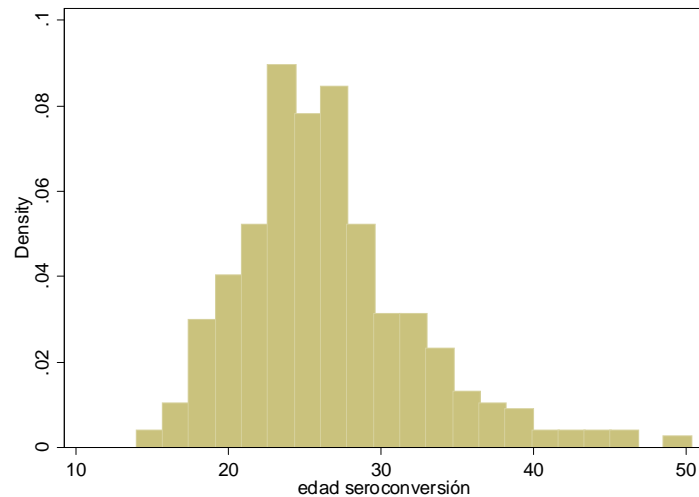
```
.
```

Variables cuantitativas

- Histograma

```
histogram variable, bin(#) width(#) normal  
kdensity by(vargrup)
```

```
histogram edadsero,
```



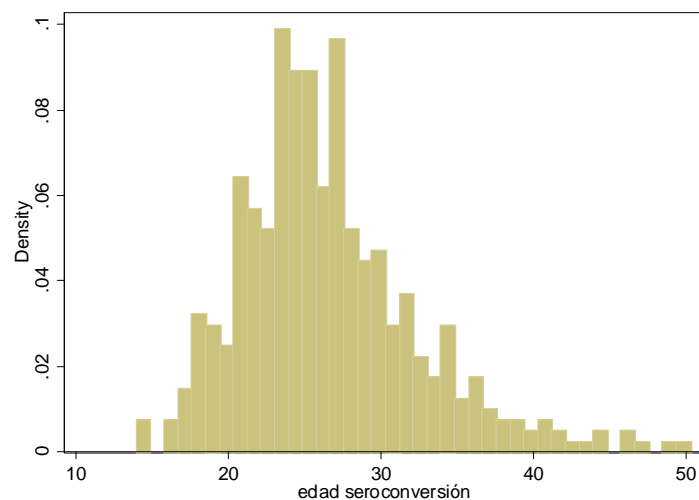
11

Variables cuantitativas

- Histograma

```
histogram variable, bin(#) width(#) normal  
kdensity by(vargrup)
```

```
histogram edadsero, bin(40)
```



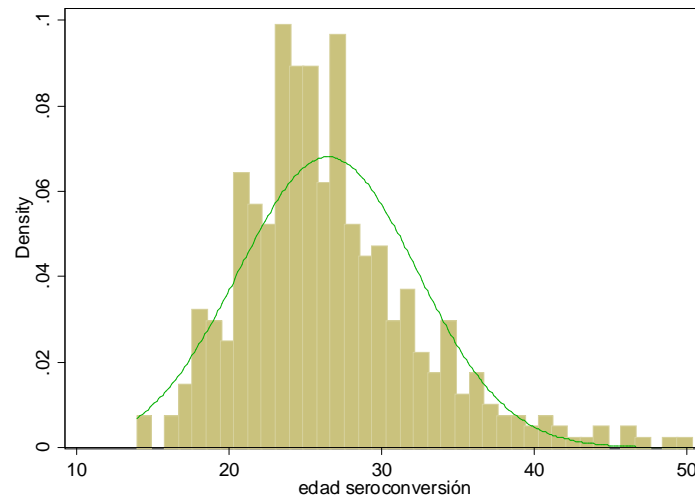
12

Variables cuantitativas

- Histograma

```
histogram variable, bin(#) width(#) normal  
kdensity by(vargrup)
```

```
histogram edadsero, bin(40)normal
```



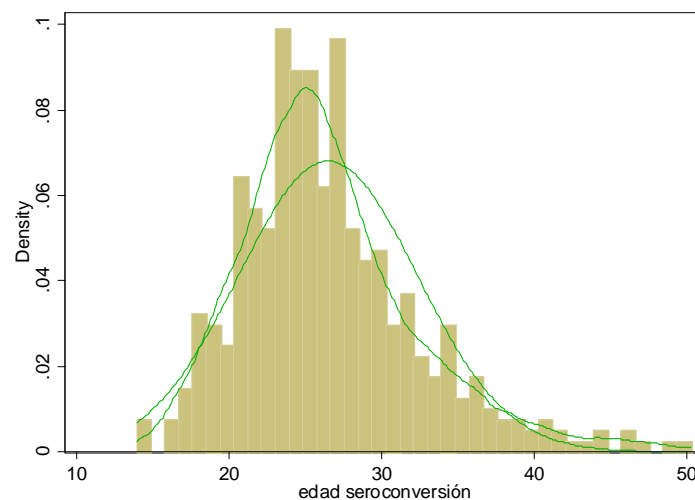
13

Variables cuantitativas

- Histograma

```
histogram variable, bin(#) width(#) normal  
kdensity by(vargrup)
```

```
histogram edadsero, bin(40)normal kdensity
```



14

Variables cuantitativas

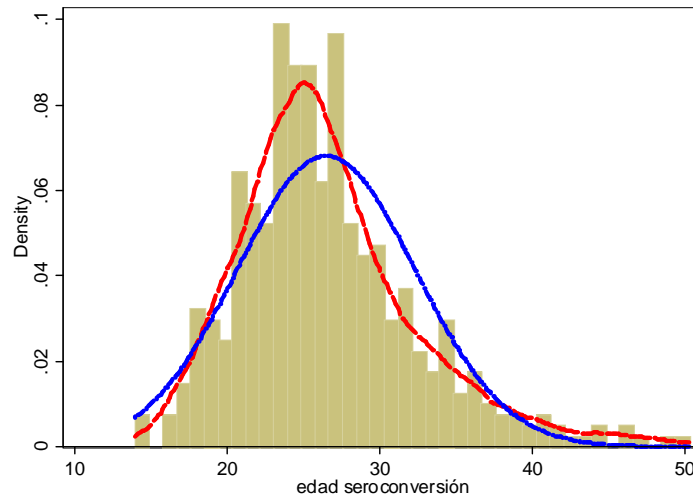
- Histograma

```

histogram variable, bin(#) width(#) normal
kdensity by(vargrup)
    
```

```

histogram edadsero, bin(40)normal kdensity
kdenopts(lpattern("_") lcolor(red) lwidth(thick))
normopts(lpattern(".-") lcolor(blue) lwidth(thick))
    
```



15

Variables cuantitativas

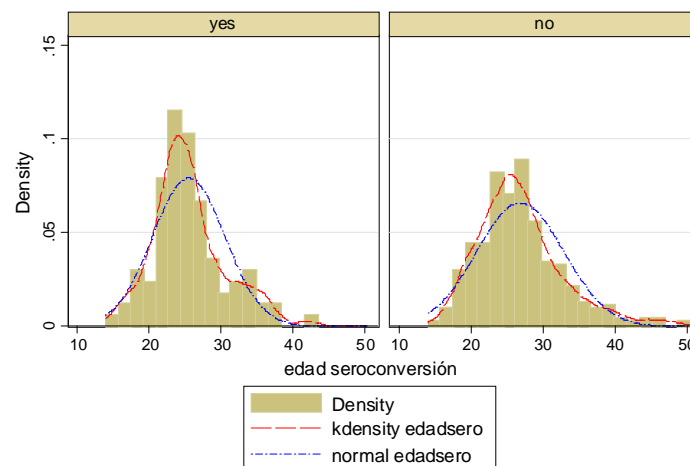
- Histograma

```

histogram variable, bin(#) width(#) normal
kdensity by(vargrup)
    
```

```

histogram edadsero, normal normopts(lpattern(".-")
lcolor(blue)) kdensity kdenopts(lpattern("_")
lcolor(red)) by(aids)
    
```



Graphs by aids diagnosis

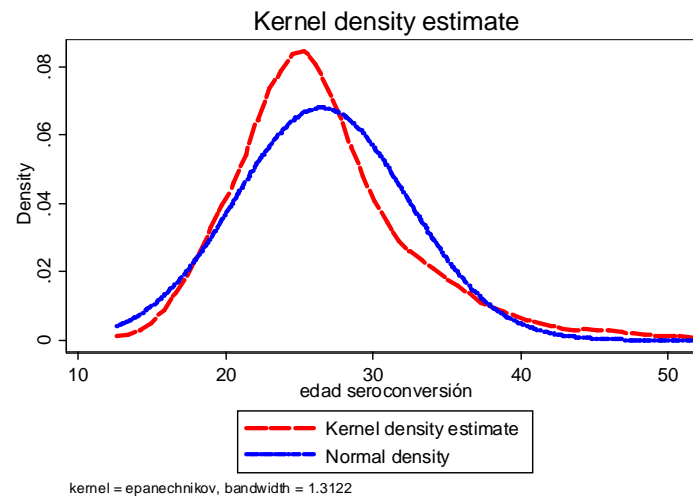
16

Variables cuantitativas

- Gràfico densidad

kdensity *variable*, **normal**

```
kdensity edadsero, lpattern("_") lcolor(red) lwidth(thick))  
normal normopts(lpattern(".-") lcolor(blue) lwidth(thick))
```



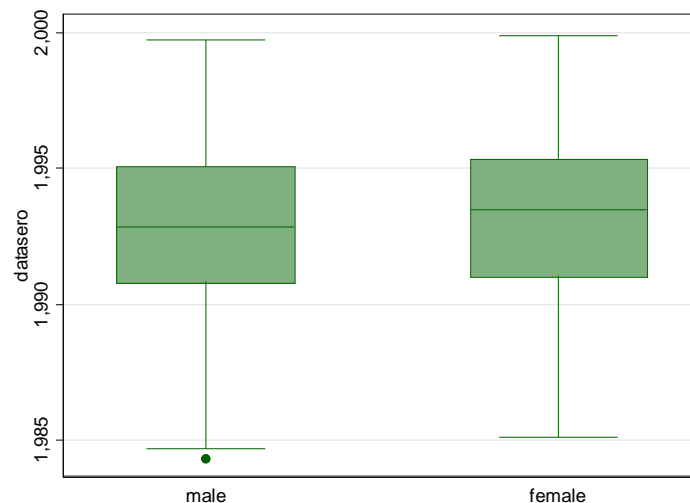
17

Variables cuantitativas

- Diagrama de cajas

graph box *variable*, **over**(*vargrup*)

```
graph box datasero, over(sex)
```



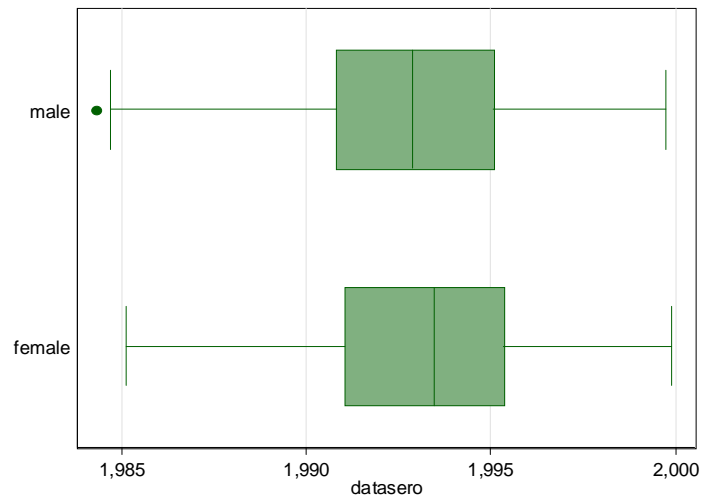
18

Variables cuantitativas

- Diagrama de cajas

graph hbox variable, **over**(vargrup)

graph hbox datasero, over(sex)



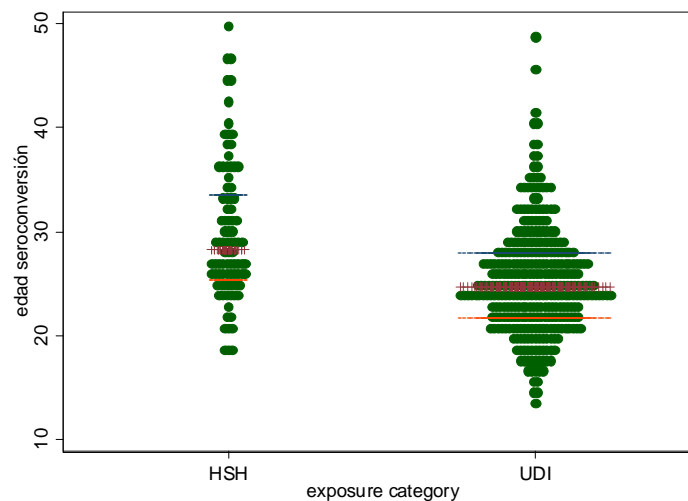
19

Variables cuantitativas

- Diagrama de puntos

dotplot var, **over**(vargrup) **center median bar**

dotplot edadsero , over(expcateg) center median bar



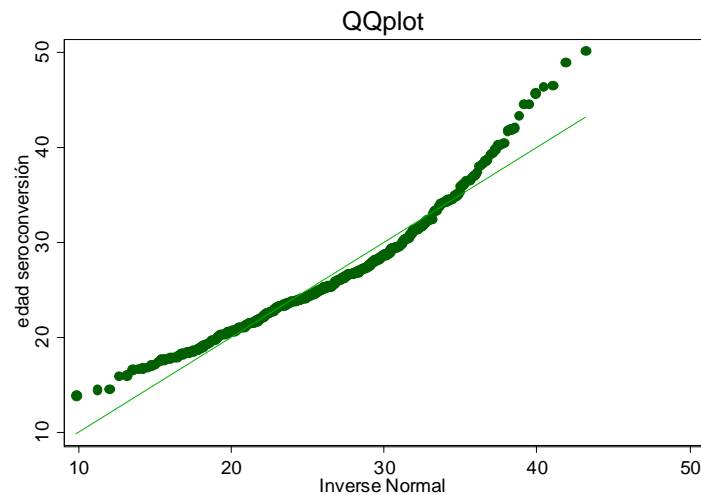
20

Variables cuantitativas

- QQplot (valor obs vs valor esperado percentil en la normal)

qnorm *var, opciones gráfico*

```
qnorm edadsero , title(QQplot)
```



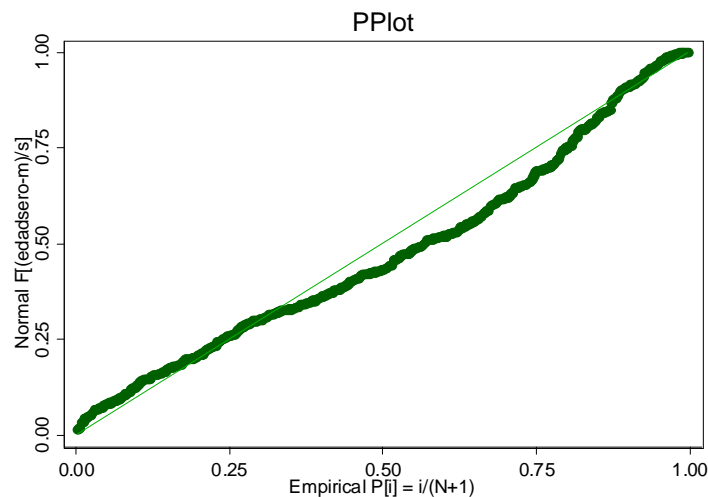
21

Variables cuantitativas

- PPplot (percentil obs vs percentil esperado en la normal)

pnorm *var, opciones gráfico*

```
pnorm edadsero , title(PPplot)
```



22

Variables cualitativas

	Medidas	Gràficos
Univariante	Tablas de Frecuencia	Pastel, Gràfico barras
Bivariante	Tablas de contingencia	Gràfico de barras , Gràfico de barras acumulado
Estudio cohorte	Tasa de incidencia Incidencia acumulada Riesgo Relativ	--
Estudio de casos- control	Odds Odds ratio	--

23

Variables cualitativas

- Calculo de tablas de frecuencia

tab1 variables, plot

tabulate variable

```
. tab1 sex expcateg
-> tabulation of sex
```

sexe	Freq.	Percent	Cum.
male	329	78.90	78.90
female	88	21.10	100.00
Total	417	100.00	

```
-> tabulation of expcateg
```

exposure category	Freq.	Percent	Cum.
homo	83	19.90	19.90
idu	334	80.10	100.00
Total	417	100.00	

24

Variables cualitativas

- Tablas de contingencia

tabulate variable1 variable2 ,row col chi exact

```
. tabulate expcateg aids, row col chi exact
```

Key			
frequency			
row percentage			
column percentage			

exposure	aids diagnosis		
category	yes	no	Total
homo	20	63	83
	24.10	75.90	100.00
	21.28	19.50	19.90
idu	74	260	334
	22.16	77.84	100.00
	78.72	80.50	80.10
Total	94	323	417
	22.54	77.46	100.00
	100.00	100.00	100.00

Pearson chi2(1) = 0.1434 Pr = 0.705
 Fisher's exact = 0.769
 1-sided Fisher's exact = 0.402

25

Variables cualitativas

- Tablas de contingencia

table var1 var2 var3,row col c(estad var)

```
. table expcateg aids sex, row
```

sexe and aids diagnosis				
exposure	-- male --		- female -	
category	yes	no	yes	no
homo	20	63		
idu	57	189	17	71
Total	77	252	17	71

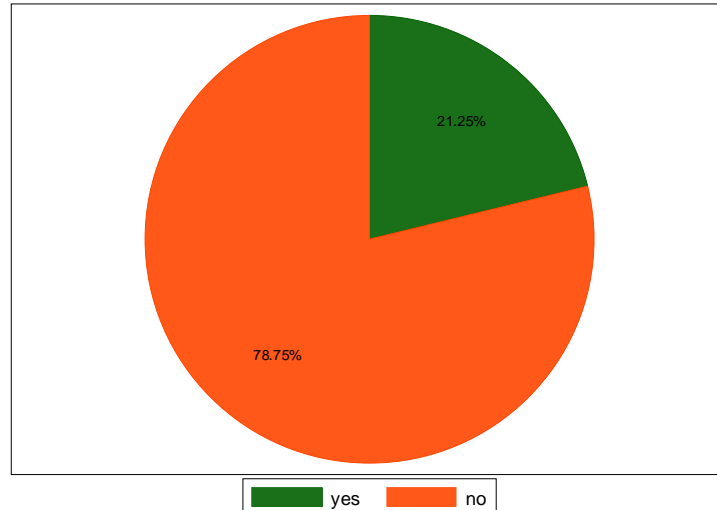
```
. table expcateg aids sex, row col c(mean edadsero)
```

sexe and aids diagnosis						
exposure	male			female		
category	yes	no	Total	yes	no	Total
homo	29.46427	30.4994	30.24381			
idu	24.80268	25.68932	25.48388	22.4016	25.66792	25.03693
Total	26.01348	26.86298	26.66295	22.4016	25.66792	25.03693

Variables cualitativas

- Diagrama de sectores

```
graph pie ,over(var1) plabel(_all percent)
graph pie ,over(aids) plabel(_all percent)
```

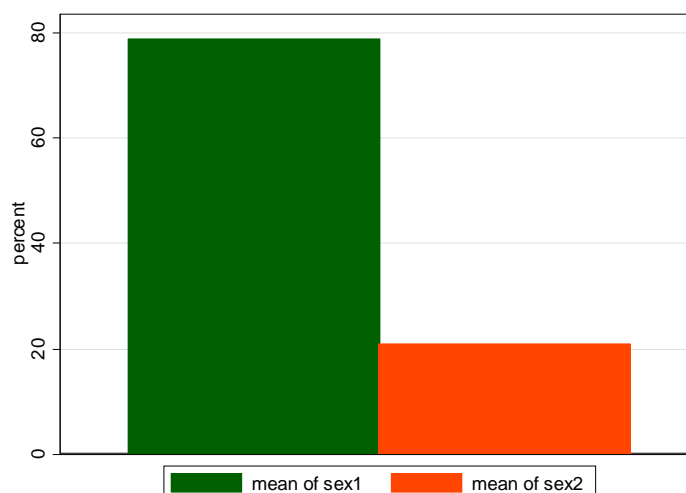


27

Variables cualitativas

- Diagrama de barras

```
graph bar (stat) vars ,over(vargrupo)
tabulate sex, generate (sex)
graph bar sex1 sex2, percent
```

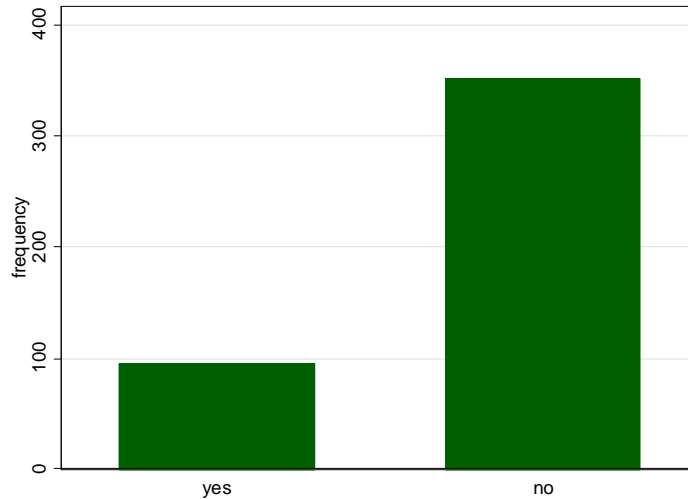


28

Variables cualitativas

- Catplot

```
catplot {bar|hbar} catvar1 , by(var2)  
    {percent|percent(var2)} asyvars stack  
catplot bar aids
```

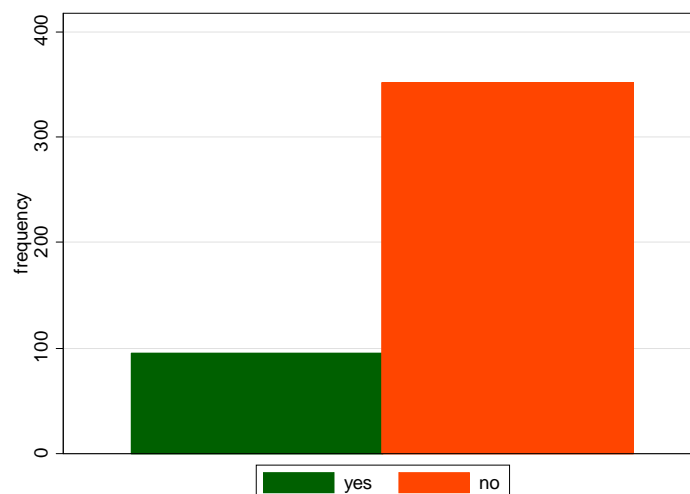


29

Variables cualitativas

- Catplot

```
catplot {bar|hbar} catvar1 , by(var2)  
    {percent|percent(var2)} asyvars stack  
catplot bar aids, asyvars
```

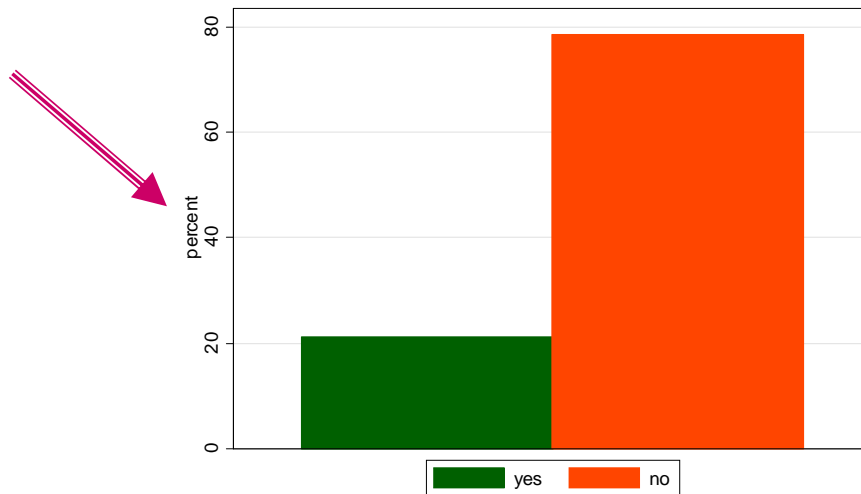


30

Variables cualitativas

- Catplot

```
catplot {bar|hbar} catvar1 , by(var2)  
    {percent|percent(var2)} asyvars stack  
catplot bar aids, asyvars percent
```

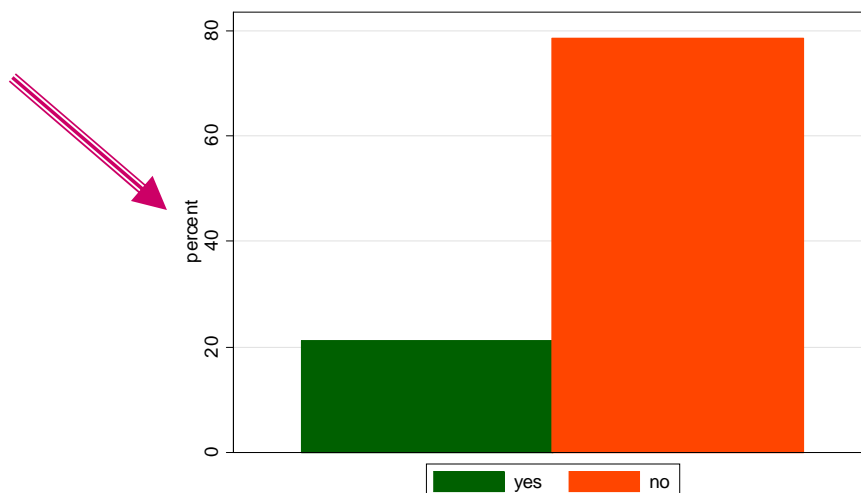


31

Variables cualitativas

- Catplot

```
catplot {bar|hbar} catvar1 , by(var2)  
    {percent|percent(var2)} asyvars stack  
catplot bar aids, asyvars percent
```

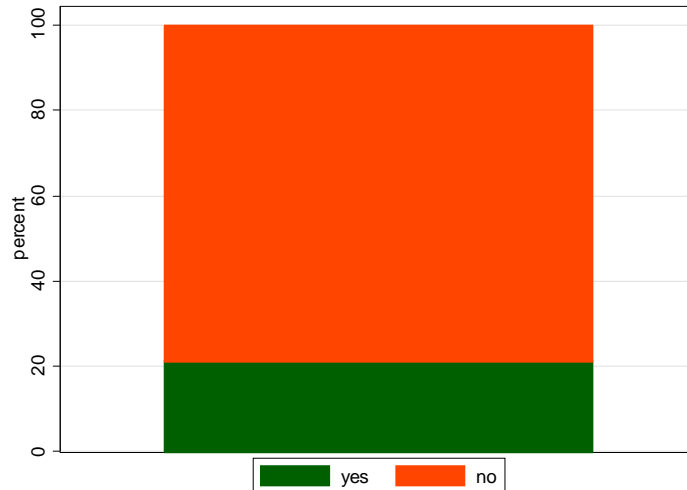


32

Variables cualitativas

- Catplot

```
catplot {bar|hbar} catvar1 , by(var2)
        {percent|percent(var2)} asyvars stack
catplot bar aids, asyvars percent stack
```

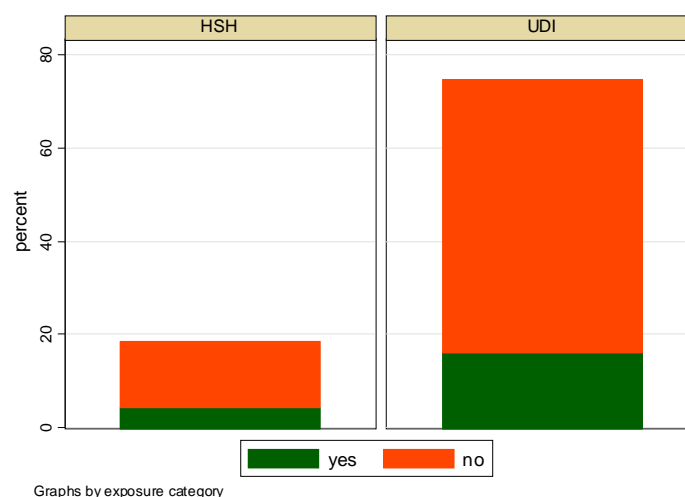


33

Variables cualitativas

- Catplot

```
catplot {bar|hbar} catvar1 , by(var2)
        {percent|percent(var2)} asyvars stack
catplot bar aids, asyvars percent stack by(expcateg)
```



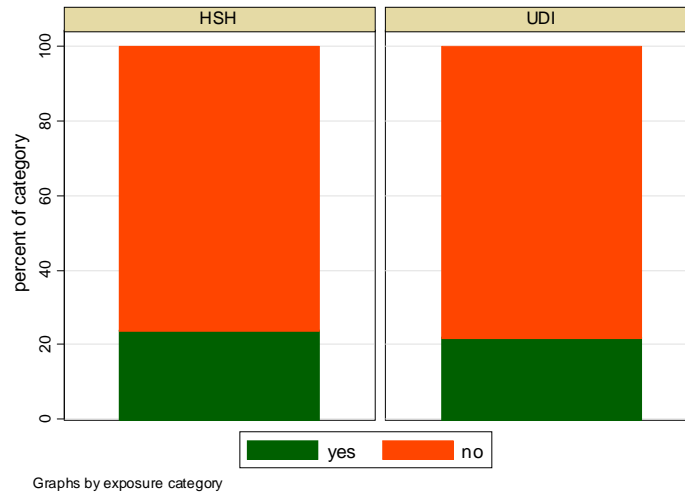
Graphs by exposure category

34

Variables cualitativas

- Catplot

```
catplot {bar|hbar} catvar1 , by(var2)
      {percent|percent(var2)} asyvars stack
catplot bar aids, asyvars percent(expcateg) stack
      by(expcateg)
```



35

Tablas epidemiológicas

- Estudio de cohortes. Calculo de tasas de incidencia

```
ir varcaso varexp vartemprisk, options
```

```
ir aids HSH tsida
```

	HSH			
	Exposed	Unexposed	Total	
SIDA	20	73	93	
tsida	386.2433	1549.648	1935.892	
Incidence rate	.0517808	.0471075	.0480399	
	Point estimate		[95% Conf. Interval]	
Inc. rate diff.	.0046734		-.0197577	.0291044 (tb)
Inc. rate ratio	1.099207		.670391	1.802314 (tb)
Attr. frac. ex.	.0902529		-.4916669	.4451578 (tb)
Attr. frac. pop	.0194092			
	(midp)	Pr(k>=20) =	0.3470 (exact)	
	(midp)	2*Pr(k>=20) =	0.6940 (exact)	

36

Tablas epidemiológicas

- Estudio de cohortes. Calculo de tasas de incidencia
`iri casos_a temps_a casos_b temps_b, tb`

<code>iri 20 386.2433 73 1549.648</code>				
	SHS			
	Exposed	Unexposed		
-----	-----	-----		
SIDA	20	73		
tsida	386.2433	1549.648		
-----	-----	-----		
Incidence rate	.0517808	.0471075		
	Point estimate		[95% Conf. Interval]	
-----	-----	-----		
Inc. rate diff.	.0046734		-.0197577	.0291044 (tb)
Inc. rate ratio	1.099207		.670391	1.802314 (tb)
Attr. frac. ex.	.0902529		-.4916669	.4451578 (tb)
Attr. frac. pop	.0194092			
-----	-----	-----		
	(midp)	Pr(k>=20) =		0.3470 (exact)
	(midp)	2*Pr(k>=20) =		0.6940 (exact)

Tablas epidemiológicas

- Estudio de cohortes. Calculo de incidencia acumulada
`cs varcaso varexp, options`

<code>cs aids HSH</code>				
	SHS			
	Exposed	Unexposed		Total
-----	-----	-----		-----
Cases	20	74		94
Noncases	63	260		323
-----	-----	-----		-----
Total	83	334		417
Risk	.2409639	.2215569		.2254197
	Point estimate		[95% Conf. Interval]	
-----	-----	-----		-----
Risk difference	.019407		-.0828121	.1216261
Risk ratio	1.087594		.7064223	1.674437
Attr. frac. ex.	.0805389		-.4155838	.4027844
Attr. frac. pop	.0171359			
-----	-----	-----		-----
	chi2(1) =	0.14	Pr>chi2 =	0.7049

Tablas epidemiológicas

- Calculo de incidencia acumulada

csi *casos_a casos_b nocasos_c nocasos_d*

csi 20 74 63 260			
	HSH		
	Exposed	Unexposed	
Cases	20	74	
Noncases	63	260	
Total	83	334	
Risk	.2409639	.2215569	
	Point estimate		
Risk difference	.019407		
Risk ratio	1.087594		
Attr. frac. ex.	.0805389		
Attr. frac. pop	.0171359		
+-----			
	chi2(1) =	0.14	Pr>chi2 = 0.7049

csi - Cohort studies

Exposed

Unexposed

Cases

20

74

Noncases

63

260

☒ Report odds ratio
 ☐ Woolf approximation
 ☐ Test-based confidence intervals
 ☐ Fisher's exact p

95 Confidence level

OK

Cancel

Submit

Tablas epidemiológicas

- Estudio de casos-contróles. Calculo de incidencia acumulada

cc *varcaso varexp, options*

cs aids HSH					
	. cc aids HSH				
	Exposed	Unexposed	Total	Proportion	
				Exposed	
Cases	20	74	94	0.2128	
Controls	63	260	323	0.1950	
Total	83	334	417	0.1990	
	Point estimate		[95% Conf. Interval]		
Odds ratio	1.115401		.5986311	2.013842	(exact)
Attr. frac. ex.	.1034615		-.6704779	.5034366	(exact)
Attr. frac. pop	.0220131				
+-----					
	chi2(1) =	0.14	Pr>chi2 =	0.7049	

Tablas epidemiológicas

- Calculo de incidencia acumulada

cci *casos_a casos_b nocasos_c nocasos_d*

cci 20 74 63 260

. cc aids HSH

	Exposed	Unexposed
Cases	20	74
Controls	63	260
Total	83	334
Point estimate		
Odds ratio	1.115401	
Attr. frac. ex.	.1034615	
Attr. frac. pop	.0220131	

chi2(1) = 0.14 Pr>chi2 = 0.7049

cci - Case-control studies

	Exposed	Unexposed
Cases	20	74
Controls	63	260

☒ Exact confidence intervals
☐ Cornfield approximation
☐ Woolf approximation
☐ Test-based confidence intervals

☐ Fisher's exact p

95 Confidence level

? R

OK Cancel Submit

1.115401
- .6704779 .5034366 (exact)

Cálculo de Tamaño muestral

Cálculo tamaño muestral

- Es mejor ir a por menús

The screenshot shows the Stata software interface. The 'Statistics' menu is open, and 'Classical tests of hypotheses' is selected. A submenu is displayed with various tests, including 'One-sample mean-comparison test', 'Two-sample mean-comparison test', 'Binomial probability test', 'One-sample proportion test', 'Two-sample proportion test', 'One-sample variance-comparison test', 'Two-sample variance-comparison test', 'Robust equal variance test', and 'Sample size and power determination' (highlighted). In the background, a table of results is visible, showing counts for 'Exposed' and 'Unexposed' groups, and a chi-squared test result: $\chi^2(1) = 0.14$, $Pr > \chi^2 = 0.7049$.

	Exposed	Unexposed	Total
Cases	20	74	94
Controls	63	260	323
Total	83	334	417

	Point estimate	95% CI
Ratio	1.115401	.59863
ex.	.1034615	-.67047
pop	.0220131	

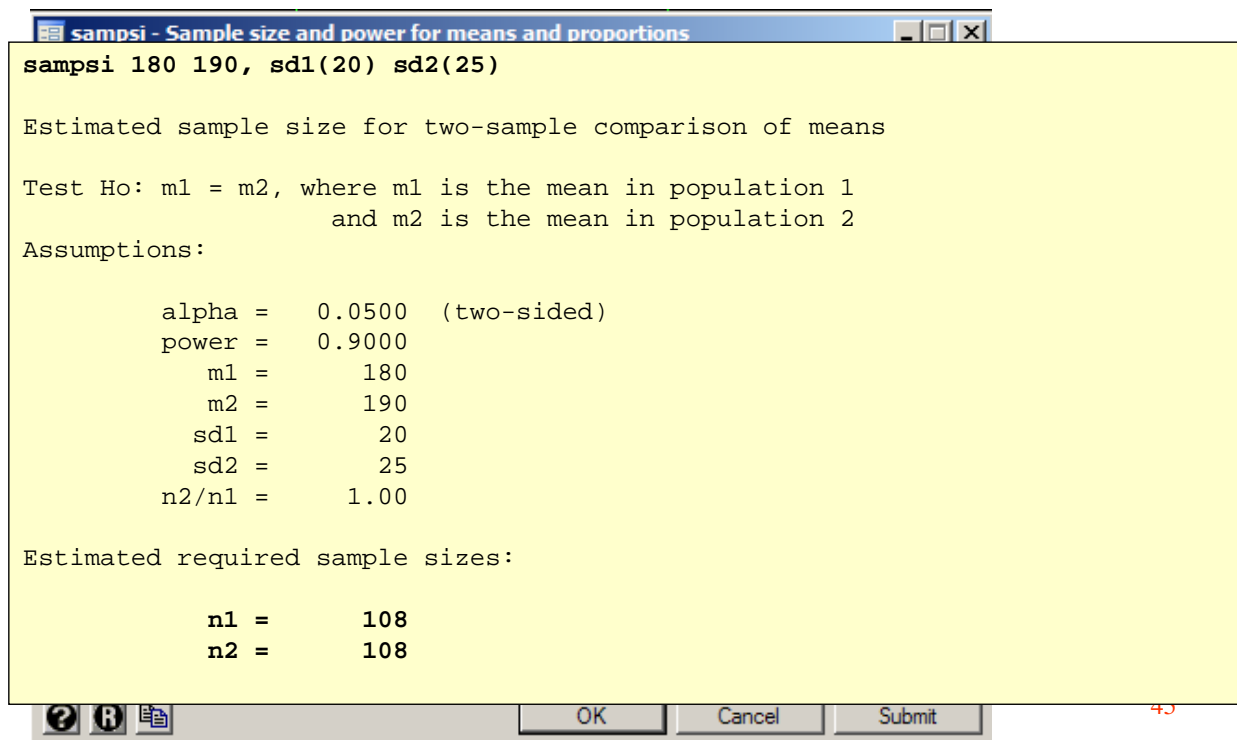
Cálculo tamaño muestral

- Es mejor ir a por menús

The screenshot shows the 'sampsi - Sample size and power for means and proportions' dialog box. The 'Main' tab is selected. Under 'Input', the 'Two-sample comparison of means' option is chosen. The 'Mean one' and 'Mean two' fields are empty, and the 'Std. deviation one' and 'Std. deviation two' fields are both set to 1. The 'One-sample comparison of mean to hypothesized value' option is also selected, with 'Hypothesized' and 'Postulated' fields empty. The 'Two-sample comparison of proportions (values in [0,1])' option is selected, with 'Proportion one' and 'Proportion two' fields empty. The 'One-sample comparison of proportions to hypothesized values (in [0,1])' option is also selected, with 'Hypothesized' and 'Postulated' fields empty. The 'OK' button is highlighted. In the bottom right corner, there is a smaller window showing the 'Options' tab. Under 'Output', the 'Compute sample size' option is chosen. The 'Significance level (alpha)' field is set to .05, and the 'Power of the test' field is set to .90. Under 'Sample-based calculations', the 'Sample one size' field is set to 100, the 'Sample two size' field is set to 100, and the 'Ratio of sample sizes' field is set to 1. The 'Type of test' dropdown is set to 'Two-sided test', and the 'Do not use continuity correction' checkbox is unchecked.

Ejemplos calculo muestral

- Diferencia de medias



The screenshot shows a window titled "sampsi - Sample size and power for means and proportions". The command entered is `sampsi 180 190, sd1(20) sd2(25)`. The output indicates an estimated sample size for a two-sample comparison of means. The test is $H_0: \mu_1 = \mu_2$, where μ_1 is the mean in population 1 and μ_2 is the mean in population 2. The assumptions are: $\alpha = 0.0500$ (two-sided), power = 0.9000, $\mu_1 = 180$, $\mu_2 = 190$, $\sigma_1 = 20$, $\sigma_2 = 25$, and $n_2/n_1 = 1.00$. The estimated required sample sizes are $n_1 = 108$ and $n_2 = 108$. The window has buttons for "?", "R", a document icon, "OK", "Cancel", and "Submit".

```
sampsi 180 190, sd1(20) sd2(25)
```

Estimated sample size for two-sample comparison of means

Test $H_0: \mu_1 = \mu_2$, where μ_1 is the mean in population 1
and μ_2 is the mean in population 2

Assumptions:

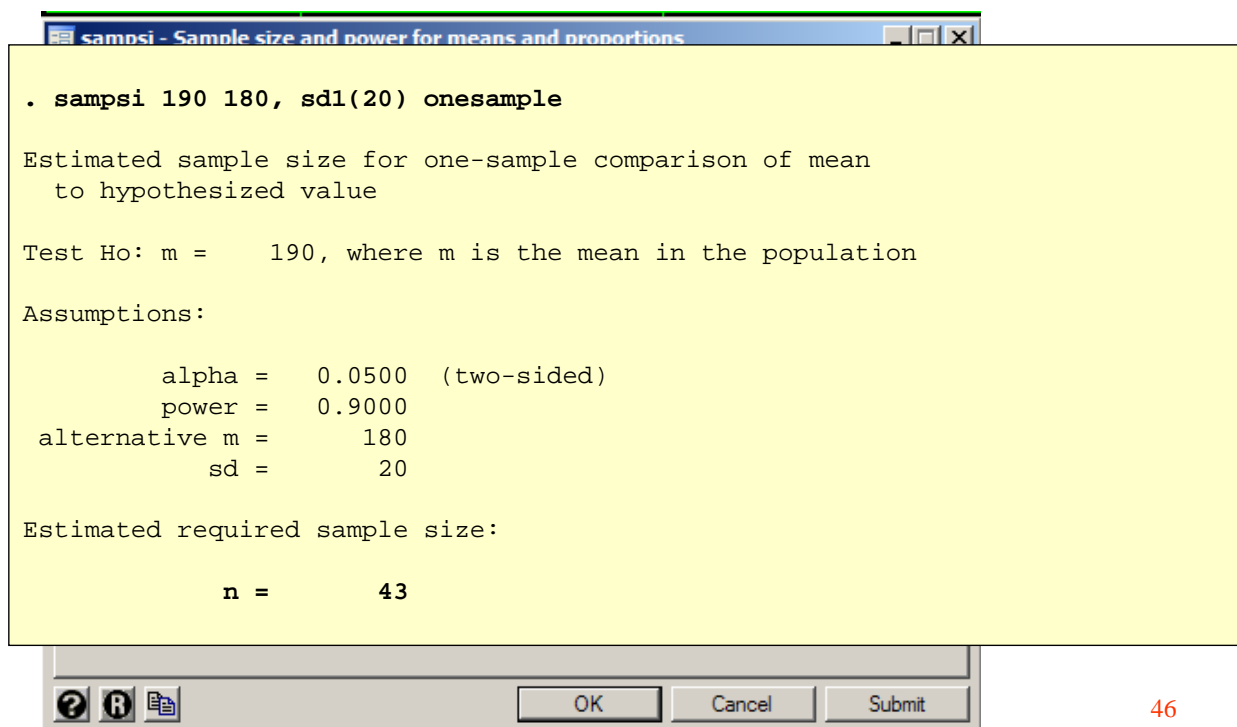
alpha =	0.0500	(two-sided)
power =	0.9000	
μ_1 =	180	
μ_2 =	190	
σ_1 =	20	
σ_2 =	25	
n_2/n_1 =	1.00	

Estimated required sample sizes:

n_1 =	108
n_2 =	108

Ejemplos calculo muestral

- Diferencia de media frente a un valor estándar



The screenshot shows a window titled "sampsi - Sample size and power for means and proportions". The command entered is `. sampsi 190 180, sd1(20) onesample`. The output indicates an estimated sample size for a one-sample comparison of mean to a hypothesized value. The test is $H_0: \mu = 190$, where μ is the mean in the population. The assumptions are: $\alpha = 0.0500$ (two-sided), power = 0.9000, alternative $\mu = 180$, and $\sigma = 20$. The estimated required sample size is $n = 43$. The window has buttons for "?", "R", a document icon, "OK", "Cancel", and "Submit".

```
. sampsi 190 180, sd1(20) onesample
```

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

Test $H_0: \mu = 190$, where μ is the mean in the population

Assumptions:

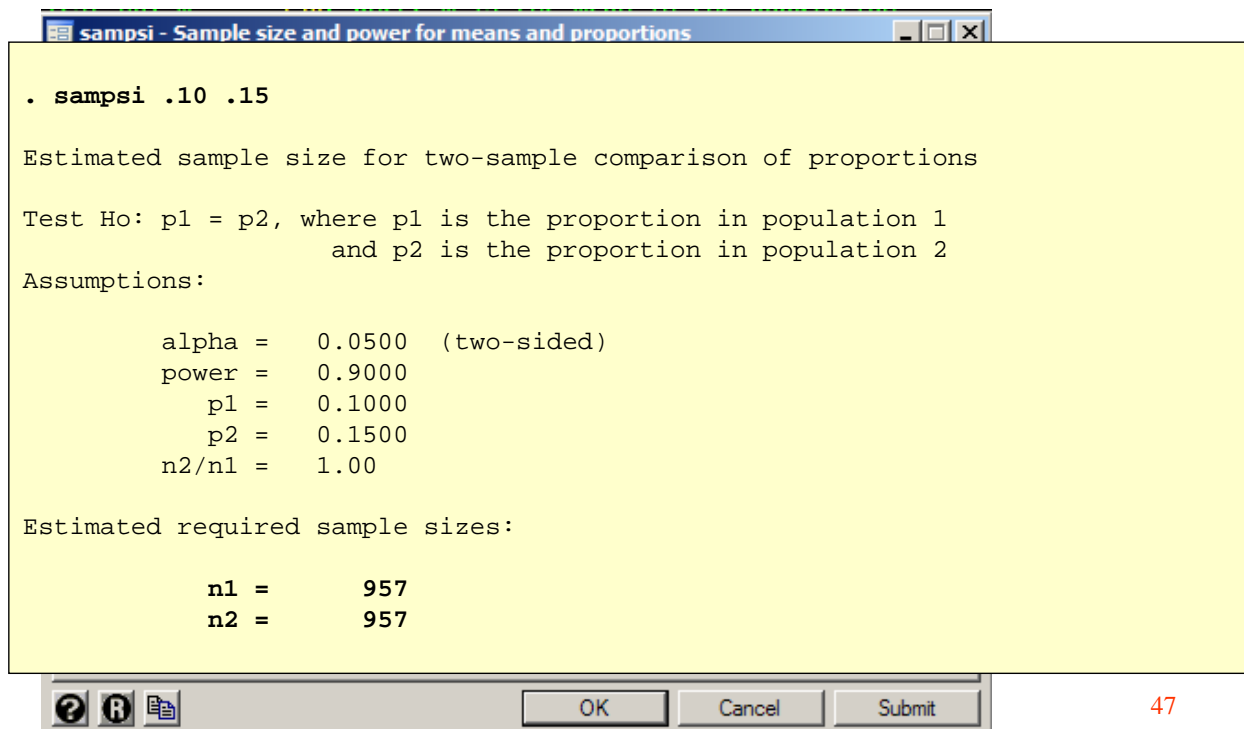
alpha =	0.0500	(two-sided)
power =	0.9000	
alternative μ =	180	
σ =	20	

Estimated required sample size:

n =	43
-------	----

Ejemplos calculo muestral

- Diferencia de proporciones



samps1 - Sample size and power for means and proportions

`. samps1 .10 .15`

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 =	0.9000	
p1 =	0.1000	
p2 =	0.1500	
n2/n1 =	1.00	

Estimated required sample sizes:

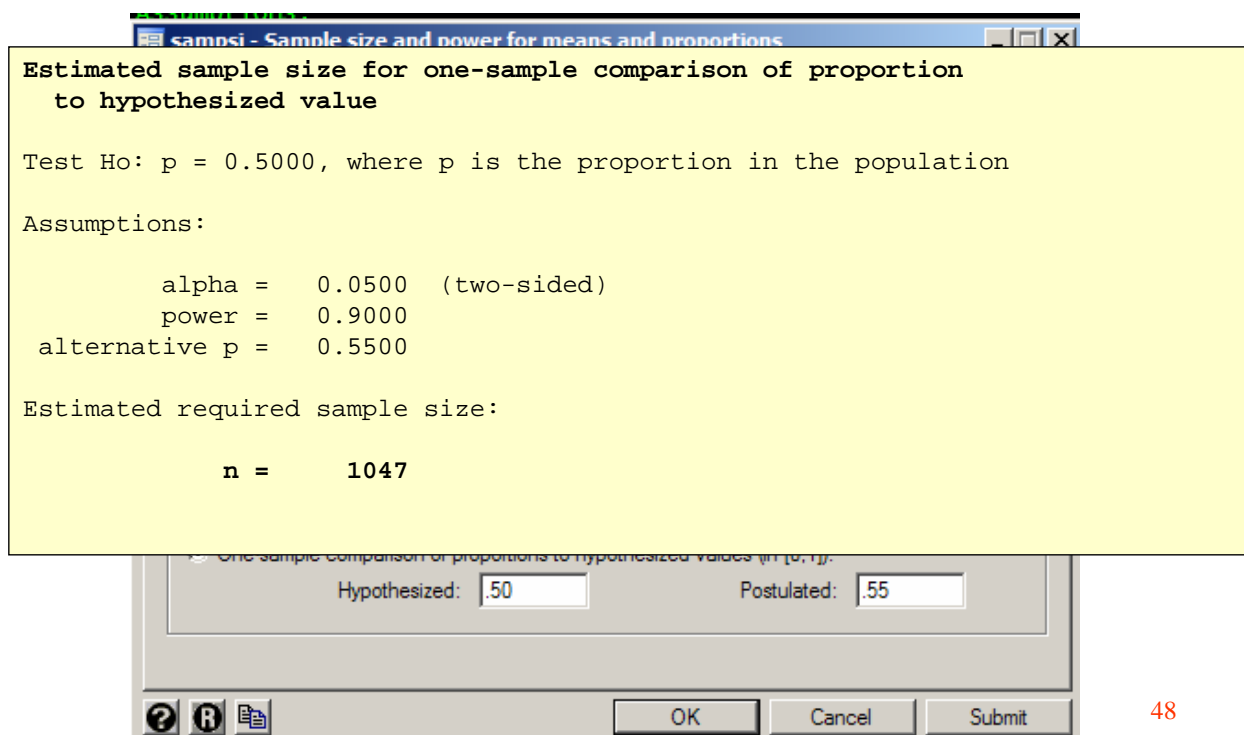
n1 =	957
n2 =	957

Buttons: ? R [icon] OK Cancel Submit

47

Ejemplos calculo muestral

- Diferencia de una proporción frente a un estándar



samps1 - Sample size and power for means and proportions

Estimated sample size 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)
power =	0.9000	
alternative p =	0.5500	

Estimated required sample size:

n =	1047
-----	------

Buttons: ? R [icon] OK Cancel Submit

Below the dialog box, a separate window shows:
Hypothesized: Postulated:

48

Generación de valores aleatorios

- Crea secuencias de números aleatorios a partir de una semilla que es fija al abrir Stata pero se puede cambiar
`set seed 339487731`
- Crea una variable U que tiene una secuencia de números aleatorios entre [0,1)
`generate u = runiform()`
- Crea una variable Z a partir de una variable normal con media 0 y desviación típica 1
`generate z = rnormal()`
- Crea una variable N a partir de una variable normal con media m y desviación típica s
`generate n = rnormal(m,s)`
- Crea una variable B a partir de una variable binomial con n observaciones y un probabilidad p de éxito
`generate b = rbinomial(n,p)`
- Crea una variable P a partir de una variable Poisson con un número de casos promedio de m
`generate r = rpoisson(m)`

49

Generación secuencia numeros aleatorios

- Se utiliza el comando `ralloc`
`ralloc bloc size treat, nsubj(387) osize(3) eq ntreat(2)`
`sav(mywide)`

	StratID	bloc	size	SeqInBlk	treat
1	1	1	2	1	B
2	1	1	2	2	A
3	1	2	4	1	B
4	1	2	4	2	A
5	1	2	4	3	A
6	1	2	4	4	B
7	1	3	4	1	A
8	1	3	4	2	B
9	1	3	4	3	A
10	1	3	4	4	B
11	1	4	4	1	A
12	1	4	4	2	A
13	1	4	4	3	B
14	1	4	4	4	B
15	1	5	6	1	B
16	1	5	6	2	A
17	1	5	6	3	A
18	1	5	6	4	A
19	1	5	6	5	B
20	1	5	6	6	B
21	1	6	2	1	B
22	1	6	2	2	A

Secuencia de Tratamiento

Orden dentro del bloque

Número de bloque

Tamaño del bloque

50