

The box office of movie sequels

The great variability in the success of movies has encouraged the production of sequels, with the hope being that the existence of the original movie provides a built-in audience for the followup movie. There is no question that sometimes this works out very well — movies like *Terminator 2: Judgment Day*, *Austin Powers: The Spy Who Shagged Me*, *Rambo: First Blood Part II*, and *Lethal Weapon 2* made a tremendous amount of money (far more than the original films did), and several sequels are commonly considered to be at least as good as the original films (for example, *The Empire Strikes Back*, *The Godfather, Part II*, *X2: X-Men United*, *Aliens*, *Mad Max 2: The Road Warrior*, and *Spider-Man 2*). Unfortunately, many sequels are both critical and box office disasters. Is there, in fact, evidence that sequels have typical total grosses that are different of those from nonsequels?

First, here are the data. The movies considered are those that opened into wide release (at least 500 screens the first weekend) in 2004.

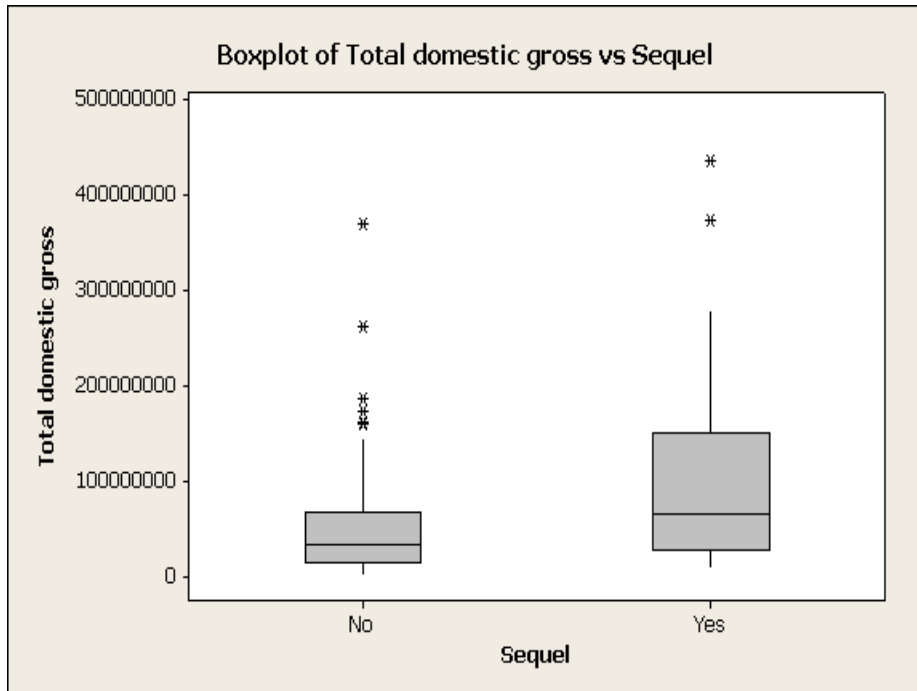
Row	Movie	Sequel	Total domestic gross
1	13 Going on 30	No	56044241
2	50 First Dates	No	120776832
3	A Cinderella Story	No	51431160
4	After the Sunset	No	28328132
5	Against the Ropes	No	5881504
6	Agent Cody Banks 2: Destination London	Yes	23222861
7	Alexander	No	34293771
8	Alfie	No	13395939
9	Alien vs. Predator	Yes	80281096
10	Along Came Polly	No	87856565
11	Anacondas: The Hunt for the Blood Orchid	Yes	31526393
12	Anchorman: The Legend of Ron Burgundy	No	84136909
13	Around the World in 80 Days	No	24004159
14	Barbershop 2: Back in Business	Yes	64955956
15	Benji Off the Leash	No	3785401
16	Birth	No	5005883
17	Blade: Trinity	Yes	52397389
18	Bobby Jones: Stroke of Genius	No	2694071
19	Breakin' All the Rules	No	11827301
20	Bridget Jones: The Edge of Reason	Yes	40203020
21	Broken Lizard's Club Dread	No	4992159
22	Catch That Kid	No	16702864
23	Catwoman	No	40198710

24	Cellular	No	32003620
25	Chasing Liberty	No	12189514
26	Christmas with the Kranks	No	73701902
27	Collateral	No	100003492
28	Confessions of a Teenage Drama Queen	No	29302097
29	Connie and Carla	No	8054280
30	Darkness	No	22160085
31	Dawn of the Dead	No	58885635
32	Dirty Dancing: Havana Nights	Yes	14140215
33	Disney's Teacher's Pet	No	6491350
34	DodgeBall: A True Underdog Story	No	114324072
35	Ella Enchanted	No	22913677
36	Envy	No	12181484
37	Eternal Sunshine of the Spotless Mind	No	34126138
38	Eurotrip	No	17718223
39	Exorcist: The Beginning	No	41814863
40	Fahrenheit 9/11	No	119194771
41	Fat Albert	No	48114556
42	First Daughter	No	9055010
43	Friday Night Lights	No	61188085
44	Garfield: The Movie	No	75367693
45	Godsend	No	14334645
46	Godzilla(1998)	No	136023813
47	Harold and Kumar Go to White Castle	No	18225165
48	Harry Potter and the Prisoner of Azkaban	Yes	249358727
49	Hellboy	No	59035104
50	Hero	No	53583486
51	Hidalgo	No	67286731
52	Home on the Range	No	50026353
53	I, Robot	No	144795350
54	Jersey Girl	No	25266129
55	Johnson Family Vacation	No	31179516
56	Kill Bill Vol. 2	Yes	66207920
57	King Arthur	No	51877963
58	Ladder 49	No	74541707
59	Laws of Attraction	No	17848322
60	Lemony Snicket's A Series of Unfortunate Events	No	118627117
61	Little Black Book	No	20422207
62	Main Hoon Na	No	1722450
63	Man on Fire	No	77862546
64	Mean Girls	No	86049418
65	Meet the Fockers	Yes	279167575
66	Miracle	No	64371181
67	Mr. 3000	No	21800302

68	My Baby's Daddy	No	17321573
69	National Lampoon's Gold Diggers	No	527000
70	National Treasure	No	172975674
71	Never Die Alone	No	5644575
72	New York Minute	No	14018364
73	Ocean's Twelve	Yes	125531634
74	Paparazzi	No	15712072
75	Raise Your Voice	No	10411980
76	Raising Helen	No	37486138
77	Ray	No	75305995
78	Resident Evil: Apocalypse	Yes	50740078
79	Saw	No	55153403
80	Scooby-Doo 2: Monsters Unleashed	Yes	84185387
81	Secret Window	No	47781388
82	Seed of Chucky	Yes	17016190
83	Shall We Dance	No	57887882
84	Shark Tale	No	160762022
85	Shaun of the Dead	No	13464388
86	Shrek 2	Yes	436471036
87	Sky Captain and the World of Tomorrow	No	37760080
88	Sleepover	No	8070311
89	Soul Plane	No	13922211
90	Spanglish	No	42044321
91	Spartan	No	4434432
92	Spider-Man 2	Yes	373377893
93	Starsky & Hutch	No	88200225
94	Super Babies: Baby Geniuses 2	Yes	9109322
95	Surviving Christmas	No	11198345
96	Suspect Zero	No	8712564
97	Taking Lives	No	32682342
98	Taxi	No	36609966
99	Team America: World Police	No	32774834
100	The Alamo	No	22406362
101	The Big Bounce	No	6471394
102	The Bourne Supremacy	Yes	176049130
103	The Butterfly Effect	No	57650876
104	The Chronicles of Riddick	Yes	57637485
105	The Cookout	No	11540112
106	The Day After Tomorrow	No	186739919
107	The Flight of the Phoenix	No	21009180
108	The Forgotten	No	66641205
109	The Girl Next Door	No	14589444
110	The Grudge	No	110175871
111	The Incredibles	No	261437578

112	The Ladykillers	No	39692139
113	The Manchurian Candidate	No	65948711
114	The Notebook	No	81001787
115	The Passion of the Christ	No	370614210
116	The Perfect Score	No	10387706
117	The Phantom of the Opera	No	51225796
118	The Polar Express	No	162753127
119	The Prince & Me	No	28165882
120	The Princess Diaries 2	Yes	95149435
121	The Punisher	No	33682273
122	The SpongeBob SquarePants Movie	No	85416609
123	The Stepford Wives	No	59475623
124	The Terminal	No	77032279
125	The Village	No	114195633
126	The Whole Ten Yards	Yes	16323969
127	Thunderbirds	No	6768055
128	Torque	No	21176322
129	Travellers and Magicians	No	505295
130	Troy	No	133228348
131	Twisted	No	25195050
132	Two Brothers	No	18947630
133	Van Helsing	No	120025245
134	Vanity Fair	No	16123851
135	Walking Tall	No	45860039
136	Welcome to Mooseport	No	14469428
137	White Chicks	No	69148997
138	Wicker Park	No	12831121
139	Wimbledon	No	16831505
140	Win a Date with Tad Hamilton!	No	16964743
141	Without a Paddle	No	58156435
142	Wooly Boys	No	335726
143	You Got Served	No	40066497
144	Yu-Gi-Oh!	No	19762690

The first thing to do, of course, is to look at the data. Here are side-by-side boxplots of the total domestic gross separated by whether the movie is a sequel or not:



We immediately see a problem. The two-sample t -test assumes that each of the populations are normally distributed with equal variance, and it certainly seems that neither of these of these assumptions hold here.

Let's ignore that for the moment and look at the output from the t -test.

Two-sample T for Total domestic gross

Sequel	N	Mean	StDev	SE Mean
No	123	50410882	54172705	4884588
Yes	21	111573939	122306822	26689537

Difference = μ (No) - μ (Yes)

Estimate for difference: -61163057

95% CI for difference: (-92916602, -29409512)

T-Test of difference = 0 (vs not =): T-Value = -3.81 P-Value = 0.000

DF = 142

Both use Pooled StDev = 68031184.4643

The t -test is significant at a .001 level (the p -value is less than .001), indicating a highly significant difference in total grosses. The average gross for sequels is more than \$61 million higher than that for nonsequels, which is certainly quite a bit of money! The

confidence interval of $(-92916602, -29409512)$ for the true average difference in grosses reinforces the significant difference by type of movie, as it does not include zero.

We've noted that this test is equivalent to the test for the significance of an indicator variable for sequel in a regression model, and the following output confirms that:

Regression Analysis: Total domestic gross versus Sequel?

The regression equation is

$$\text{Total domestic gross} = 50410882 + 61163057 \text{ Sequel?}$$

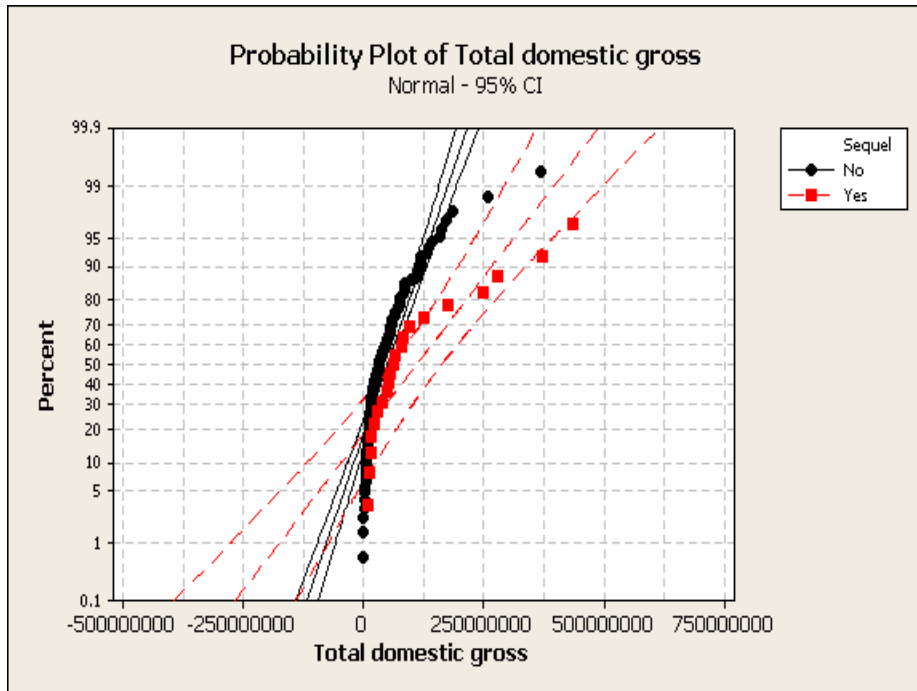
Predictor	Coef	SE Coef	T	P
Constant	50410882	6134165	8.22	0.000
Sequel?	61163057	16063015	3.81	0.000

S = 68031184 R-Sq = 9.3% R-Sq(adj) = 8.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	6.71027E+16	6.71027E+16	14.50	0.000
Residual Error	142	6.57210E+17	4.62824E+15		
Total	143	7.24313E+17			

The t -test given above has three important assumptions. First, the data have to constitute a random sample from some population. These movies are all of the wide release movies of 2004 (so it's not a sample of 2004 movies), but these movies could be viewed as a sample from the ongoing process of the production of Hollywood movies. Second, the populations of grosses for the two groups must each be (roughly) Gaussian. The side-by-side boxplots indicate that grosses are right-tailed, and normal plots do also:



Note, however, that it is clear that the sequel distribution is generally shifted upwards from the nonsequel distribution (to the right in the plot), reinforcing that sequels are apparently doing better than nonsequels.

The third assumption of the two-sample *t*-test is that the variances must be the same in the two populations. That also seems suspect here, but the *t*-test that does not assume constant variances still indicates a significantly higher average gross for sequels than for nonsequels, although the evidence is much weaker:

Two-sample T for Total domestic gross

Sequel	N	Mean	StDev	SE Mean
No	123	50410882	54172705	4884588
Yes	21	111573939	122306822	26689537

Difference = mu (No) - mu (Yes)

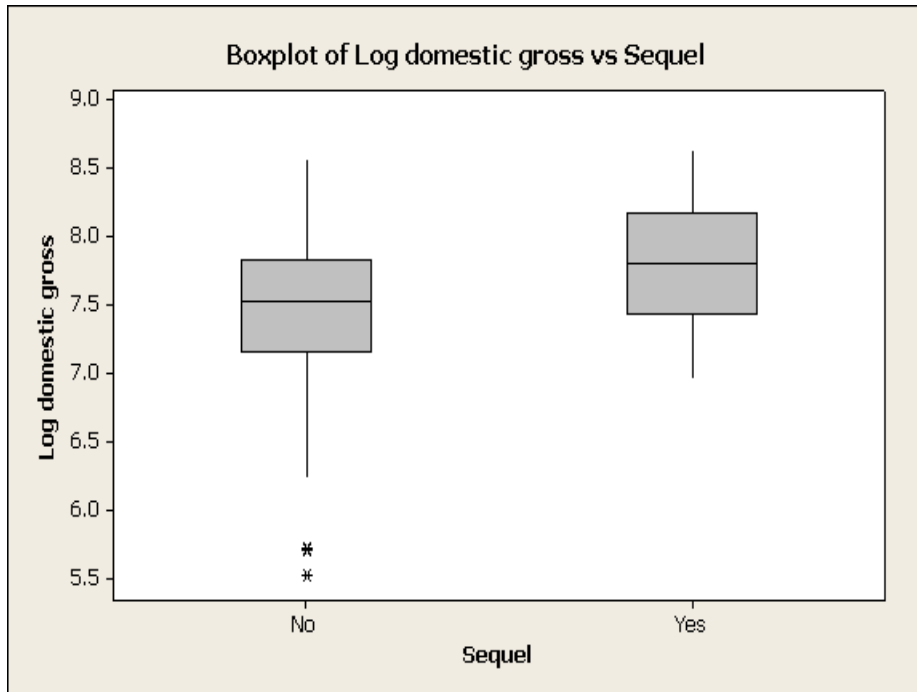
Estimate for difference: -61163057

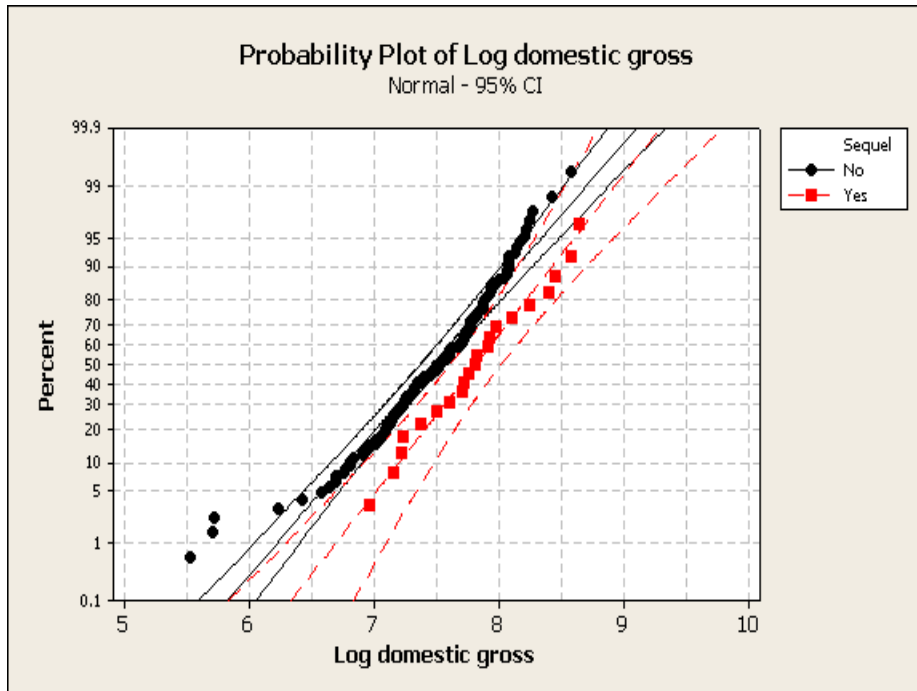
95% CI for difference: (-117588870, -4737244)

T-Test of difference = 0 (vs not =): T-Value = -2.25 P-Value = 0.035

DF = 21

The real problem here, of course, is that the domestic gross variable is long right-tailed. This naturally suggests analyzing it in the log scale, which clears up the problems quite dramatically.





Although normality is a bit questionable for the nonsequel log grosses, the assumptions are certainly much closer to being satisfied. Here is the output for the standard (equal variance) *t*-test:

Two-sample T for Log domestic gross

Sequel	N	Mean	StDev	SE Mean
No	123	7.461	0.530	0.048
Yes	21	7.812	0.477	0.10

Difference = mu (No) - mu (Yes)

Estimate for difference: -0.351272

95% CI for difference: (-0.595353, -0.107190)

T-Test of difference = 0 (vs not =): T-Value = -2.84 P-Value = 0.005

DF = 142

Both use Pooled StDev = 0.5229

The observed difference in average logged grosses of .351 is significant at a .005 level. Since we are dealing with a logged variable here, we need to think in terms of multiplicative differences; that is, sequels have expected gross that is roughly 2.25 times that of nonsequels ($10^{.351} = 2.243$).

Since the observed standard deviations in the logged scale are similar to each other, it is not surprising that the t -test that does not assume constant variances gives similar results:

Two-sample T for Log domestic gross

Sequel	N	Mean	StDev	SE Mean
No	123	7.461	0.530	0.048
Yes	21	7.812	0.477	0.10

Difference = μ (No) - μ (Yes)

Estimate for difference: -0.351272

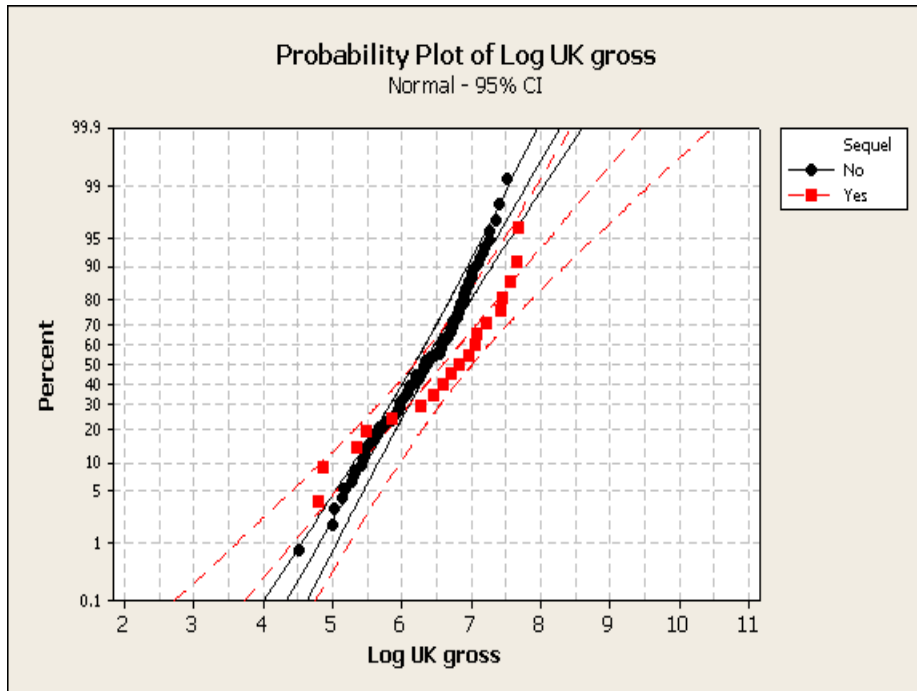
95% CI for difference: (-0.585532, -0.117012)

T-Test of difference = 0 (vs not =): T-Value = -3.07 P-Value = 0.005

DF = 29

You might wonder if sequels are similarly successful overseas. The following graphs and output refer to logged UK grosses, and they show that the evidence is much weaker in the United Kingdom that sequels will be more successful than nonsequels.





Here is output for the standard *t*-test:

Two-sample T for Log UK gross

Sequel	N	Mean	StDev	SE Mean
No	92	6.302	0.638	0.067
Yes	19	6.592	0.923	0.21

Difference = μ (No) - μ (Yes)

Estimate for difference: -0.28942

95% CI for difference: (-0.636305, 0.056422)

T-Test of difference = 0 (vs not =): T-Value = -1.66 P-Value = 0.100

DF = 109

Both use Pooled StDev = 0.6935

Here is output for the test that does not assume equal variances:

Two-sample T for Log UK gross

Sequel	N	Mean	StDev	SE Mean
No	92	6.302	0.638	0.067
Yes	19	6.592	0.923	0.21

Difference = mu (No) - mu (Yes)

Estimate for difference: -0.289942

95% CI for difference: (-0.751688, 0.171805)

T-Test of difference = 0 (vs not =): T-Value = -1.31 P-Value = 0.206
DF = 21

Minitab commands

Two-sample *t*-tests are obtained by clicking on **Stat** → **Basic Statistics** → **2-Sample t**. There are two possible forms for the data: with the variable in one column, with a second column containing codes for the two groups (the *stacked* form), or with the variable separated into two columns, one for each group (the *unstacked* form. If the data are in stacked form, enter the variable name under **Samples:**, and the variable that defines the groups under **Subscripts:**. The subscript variable can be either numerical or text. If the data are in unstacked form, click the radio button next to **Samples in different columns**, and enter the two variables in the boxes next to **First:** and **Second:**, respectively. If you want the *t*-test that assumes equal variances in the two groups, click the box next to **Assume equal variances**.

You can convert from stacked to unstacked form, and vice versa. To convert from stacked to unstacked, click on **Data** → **Unstack Columns**. Enter the variable(s) to be split up in the box next to **Unstack the data in:**. Enter the variable that defines the groups in the box next to **Using subscripts in:**. You can then choose where to put the new variables, and whether Minitab should name them for you. To convert from unstacked to stacked, click on **Data** → **Stack** → **Columns**. Enter the variables to be combined under **Stack the following columns:**. You can then choose where to put the stacked variable and associated variable of subscripts, and whether you want the subscripts to be the names of the variables (if you uncheck that box, the subscripts are the integers 1, 2, etc.).

Normal plots of more than one variable on the same plot are obtained by clicking on **Graph** → **Probability Plot**, and entering the variable names under **Variables:**. Note that in this two-group situation, you can only get this picture by using the data in *unstacked* form.

Tests of homogeneity of variance, and confidence intervals for standard deviations, are obtained by clicking on **Stat** → **ANOVA** → **Test for Equal Variances**. Enter the variable of interest in the box next to **Response:**, and the variable that defines the groups under **Factors:**.

The median test is obtained by clicking on **Stat** → **Nonparametrics** → **Mood's Median Test**. The data can only be treated if they are in *stacked* form. Enter the variable of interest in the box next to **Response:**, and the variable that defines the groups in the box next to **Factor:**.

To obtain a Mann-Whitney test, click on **Stat** → **Nonparametrics** → **Mann-Whitney**. The data can only be treated if they are in *unstacked* form. Enter the two variables that have the observations for the two groups in the boxes next to **First Sample:** and **Second Sample:**, respectively.