Applications of loops - bootstrapping

Computationally intensive methods of statistical analysis

Using the computer to make hairy statistics easy

Confidence intervals

- If you want to know how tall people are:
  - Collect a data set and calculate a mean ($\bar{x}$)
  - We know that a different set of data will give us a different mean – the current mean is probably not exactly equal to the mean for all people
- Confidence intervals give us a way of expressing what we would expect the mean to be if we collected another sample

The usual method of calculating confidence intervals

- From a sample of data, calculate the mean, standard deviation, and standard error
- Calculate a (usually 95%) confidence interval with:

$$\bar{x} \pm t\left(s_{\bar{x}}\right)$$

- Works great for a wide range of conditions
- Sometimes it doesn't work well, sometimes it's not possible to use it at all

When the simple method won't work, resample

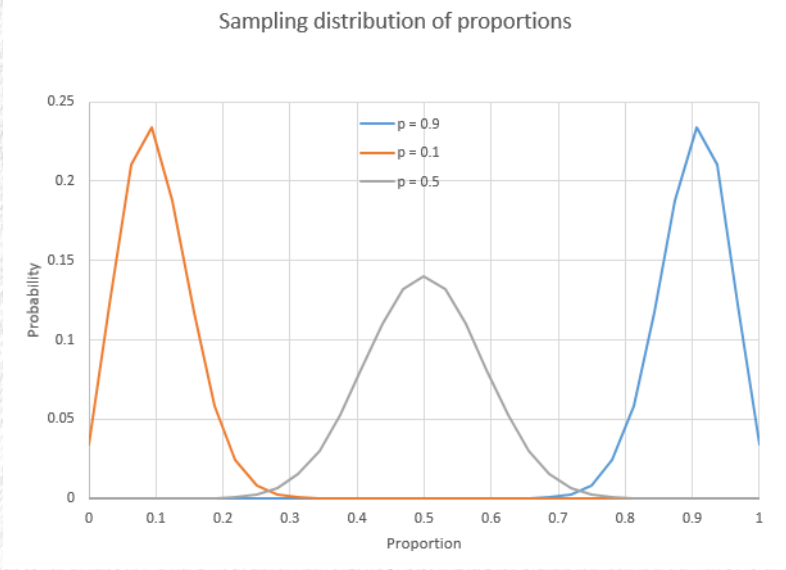
- Resampling statistical methods are based on using only the data in hand to derive p-values
- For comparisons of means use randomization tests
- For confidence intervals use the “bootstrap”

Example: parasitoid wasps

- *Trichogramma brassicae*
- Lay eggs on butterfly eggs
- Ride on legs of butterflies
- Can they tell mated butterflies from virgins?
- Present them with mated and virgins simultaneously, see which they climb on to
- Result: 23 out of 32 chose mated butterflies – proportion is 0.71875
- If they are guessing, expect 50% to be mated – does the 95% confidence interval include 50%?

Confidence intervals for proportions are problematic

- Problem – sampling distribution of proportions only symmetrical at 0.5
- Confidence intervals need to be asymmetrical



- Symmetrical intervals won't represent the possible values of p accurately, and may even go out of bounds (under 0, over 1)
- Various methods to address this analytically, but no theoretically best method

Bootstrapping the confidence interval for a proportion

- Instead of an imperfect analytical solution, we can estimate the interval by resampling
- We'll find this by:
  - Randomly selecting *with replacement* from the observed data – some observations will be included more than once, others not included at all
  - Calculate the proportion of mated butterflies each time
  - Repeat many times (at least 1000)
  - The 2.5%-ile and 97.5%-ile for proportions from the 1000 sets of resampled data are the 95% confidence limits

The data – 23 out of 32 wasps on mated butterflies

	A	B
1	Number	Wasp selections
2	1	Mated
3	2	Mated
4	3	Mated
5	4	Mated
6	5	Mated
7	6	Mated
8	7	Mated
9	8	Mated
10	9	Mated
11	10	Mated
12	11	Mated
13	12	Mated
14	13	Mated
15	14	Mated
16	15	Mated
17	16	Mated
18	17	Mated
19	18	Mated
20	19	Mated
21	20	Mated
22	21	Mated
23	22	Mated
24	23	Mated
25	24	Unmated
26	25	Unmated
27	26	Unmated
28	27	Unmated
29	28	Unmated
30	29	Unmated
31	30	Unmated
32	31	Unmated
33	32	Unmated

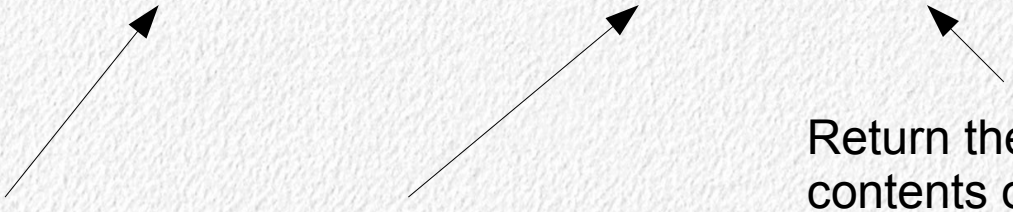
We need to sample with replacement 32 times

Randomly pick numbers from 1 to 32

=randbetween(1,32)

Select the wasp selection that corresponds with this random number

=lookup(randbetween(1,32), A2:A33, B2:B33)



Generated random number

Look up the number in the “Number” column

Return the contents of the “Wasp selections” column

E2 | =LOOKUP(D2,A$2:A$33,B$2:B$33)

	A	B	C	D	E	F	G
1	Number	Wasp selections		Random number	Bootstrap sample		
2	1	Mated		9	Mated		
3	2	Mated		31	Unmated		
4	3	Mated		6	Mated		
5	4	Mated		23	Mated		
6	5	Mated		14	Mated		
7	6	Mated		31	Unmated		
8	7	Mated		9	Mated		
9	8	Mated		24	Unmated		
10	9	Mated		15	Mated		
11	10	Mated		25	Unmated		
12	11	Mated		27	Unmated		
13	12	Mated		4	Mated		
14	13	Mated		21	Mated		
15	14	Mated		6	Mated		
16	15	Mated		9	Mated		
17	16	Mated		22	Mated		
18	17	Mated		27	Unmated		
19	18	Mated		21	Mated		
20	19	Mated		14	Mated		
21	20	Mated		27	Unmated		
22	21	Mated		10	Mated		
23	22	Mated		6	Mated		
24	23	Mated		6	Mated		
25	24	Unmated		12	Mated		
26	25	Unmated		26	Unmated		
27	26	Unmated		18	Mated		
28	27	Unmated		30	Unmated		
29	28	Unmated		15	Mated		
30	29	Unmated		3	Mated		
31	30	Unmated		20	Mated		
32	31	Unmated		29	Unmated		
33	32	Unmated		23	Mated		
34							
35				Number mated	22		
36							
37							

Randomly sample with replacement, n = 32

=randbetween(1,32)

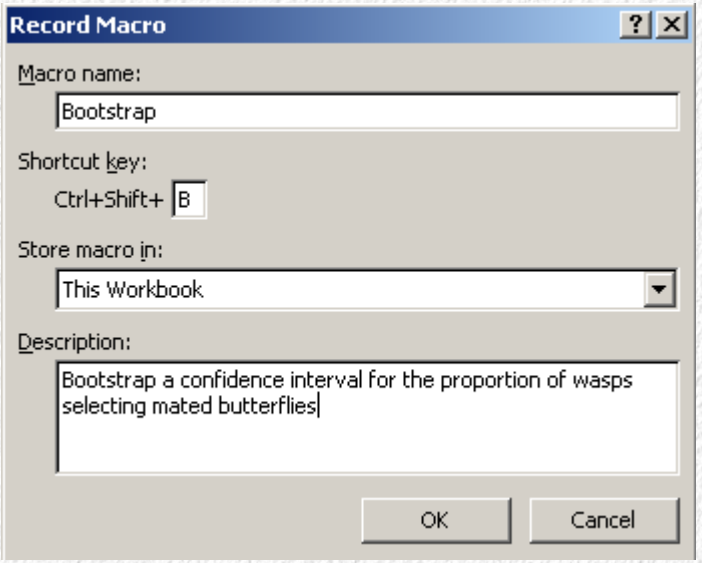
=COUNTIF(D2:D33, "Mated")

Repeat many times using a macro

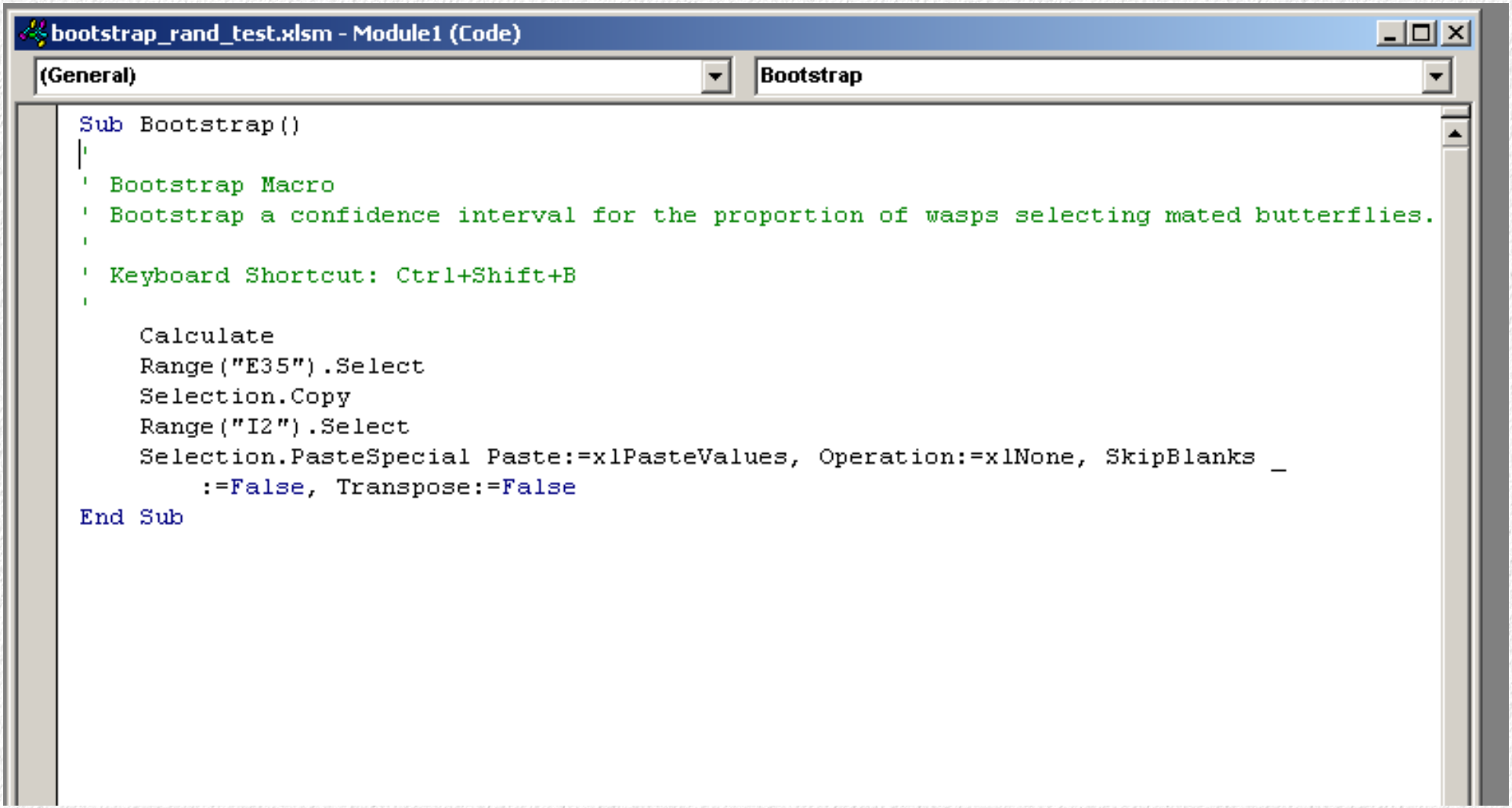
- To repeat this, we will use a “For...next” loop
  - We want to repeat a fixed number of times (1000)
  - Each time through we'll re-calculate the sheet to select a new set of random numbers
  - Each time we select a new sample, we want to record the number mated
- At the end, we will have 1000 numbers of times mated butterflies were selected
- We can calculate the proportions (divide by 32), sort them, and find the endpoints of the interval

Record macro to start

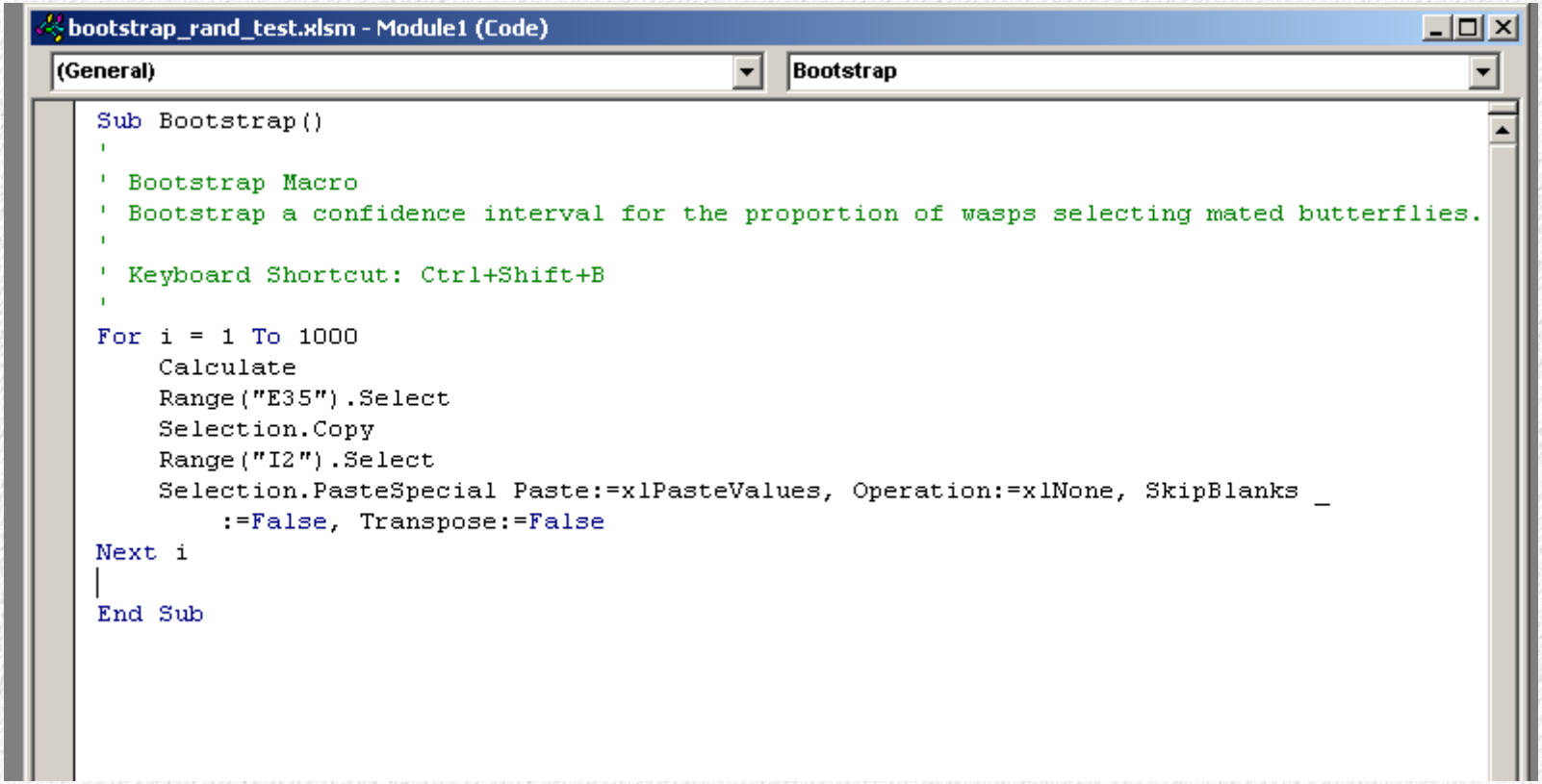
- Start the macro recorder, give it a name and a shortcut key (CTRL+SHIFT+B)
- Hit the “recalculate” key (F9) to select a new sample
- Copy the number mated, paste-special to the first row of a results column
- Stop the recorder
- Open the macro for editing



As recorded...


bootstrap_rand_test.xlsm - Module1 (Code)
(General)
Bootstrap
Sub Bootstrap()
'
' Bootstrap Macro
' Bootstrap a confidence interval for the proportion of wasps selecting mated butterflies.
'
' Keyboard Shortcut: Ctrl+Shift+B
'
Calculate
Range("E35").Select
Selection.Copy
Range("I2").Select
Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
:=False, Transpose:=False
End Sub

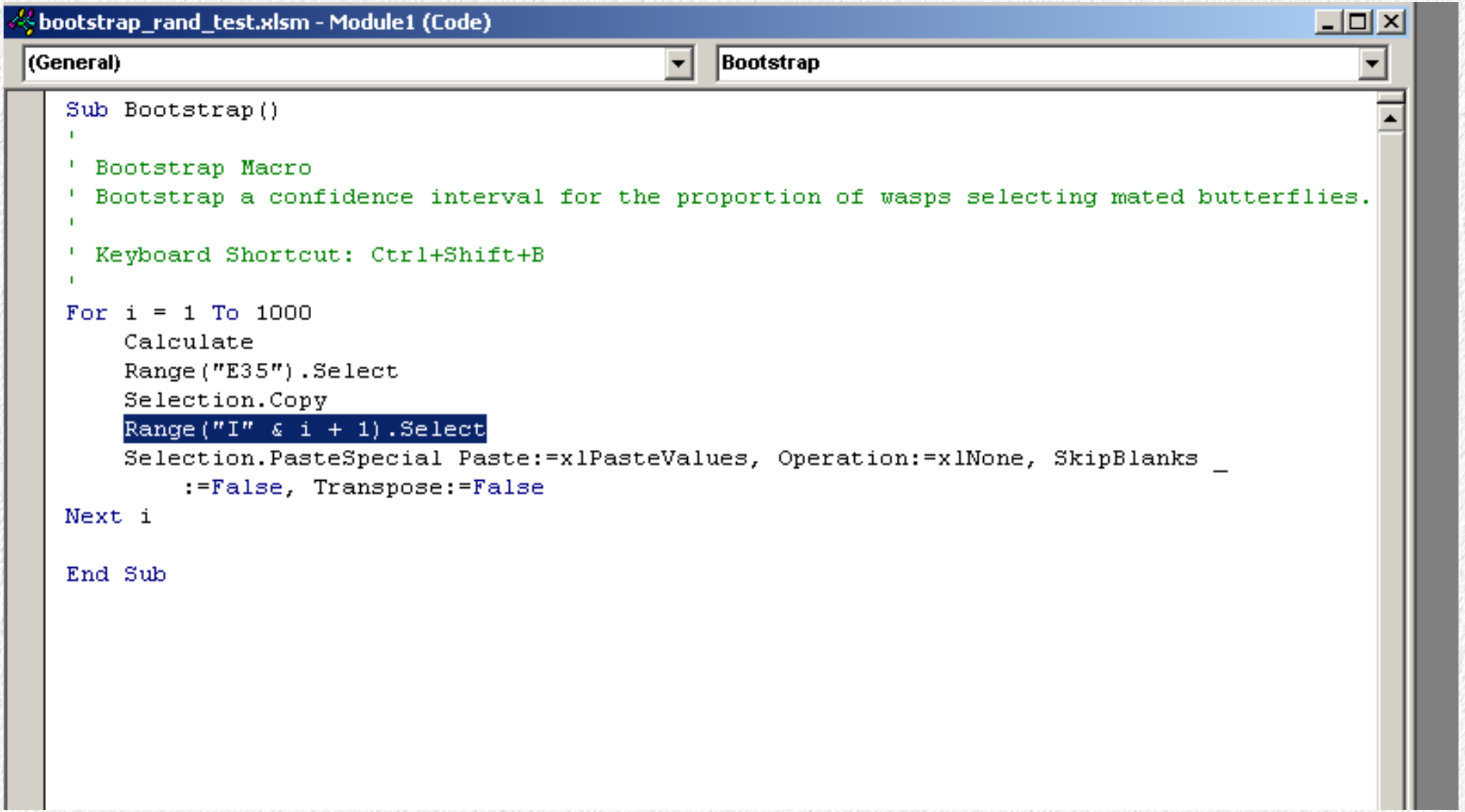
Add a loop



Confirm it worked by hitting CTRL+SHIFT+B

1,000 samples are drawn, and the number mated is pasted each time in I2
Need to move down a row each time

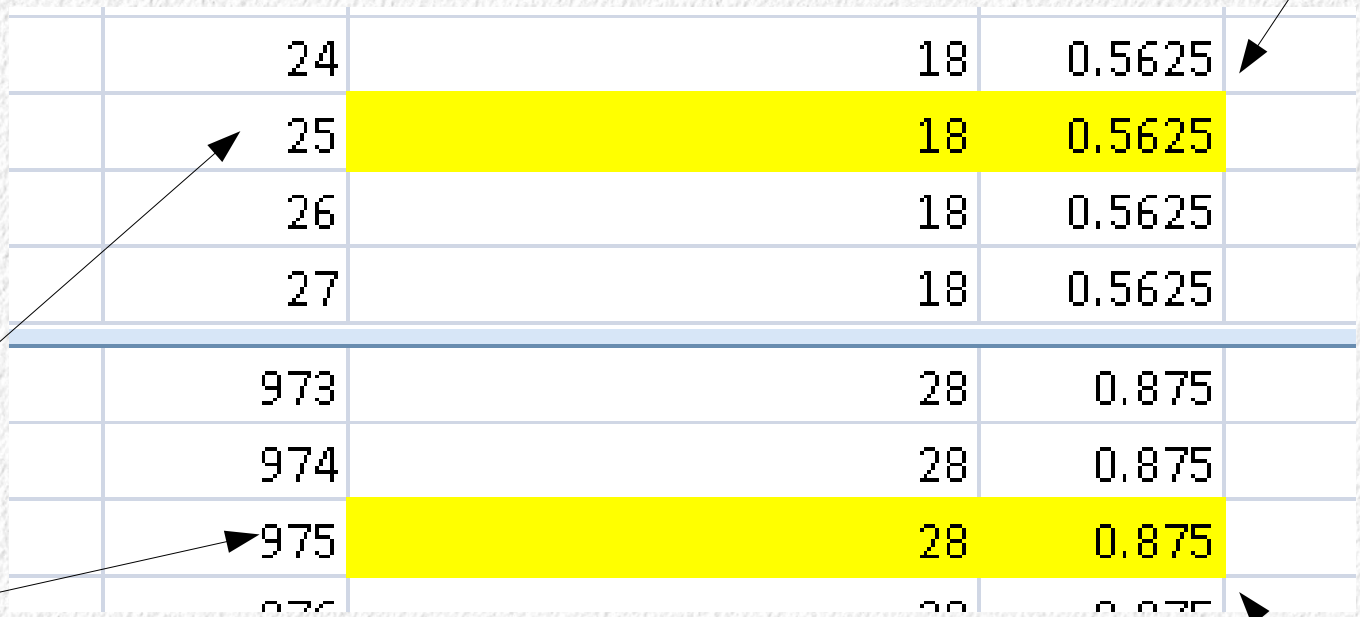
Selecting a new row each round



	A	B	C	D	E	F	G	H	I	J	K
1	Number	Wasp selections		Random number	Bootstrap sample				Bootstrap number mated		
2	1	Mated		7	Mated				20		
3	2	Mated		19	Mated				23		
4	3	Mated		31	Unmated				23		
5	4	Mated		30	Unmated				20		
6	5	Mated		18	Mated				20		
7	6	Mated		19	Mated				19		
8	7	Mated		12	Mated				22		
9	8	Mated		32	Unmated				19		
10	9	Mated		2	Mated				24		
11	10	Mated		20	Mated				29		
12	11	Mated		25	Unmated				26		
13	12	Mated		15	Mated				24		
14	13	Mated		21	Mated				25		
15	14	Mated		10	Mated				28		
16	15	Mated		14	Mated				16		
17	16	Mated		5	Mated				21		
18	17	Mated		5	Mated				24		
19	18	Mated		16	Mated				25		
20	19	Mated		7	Mated				21		
21	20	Mated		24	Unmated				23		
22	21	Mated		24	Unmated				21		
23	22	Mated		11	Mated				26		
24	23	Mated		12	Mated				20		
25	24	Unmated		27	Unmated				20		
26	25	Unmated		26	Unmated				25		
27	26	Unmated		23	Mated				20		
28	27	Unmated		24	Unmated				18		
29	28	Unmated		13	Mated				23		
30	29	Unmated		24	Unmated				23		
31	30	Unmated		27	Unmated				27		
32	31	Unmated		3	Mated				23		
33	32	Unmated		3	Mated				22		
34									21		
998									23		
999									22		
1000									21		
1001									24		
1002											

Number mated from 1,000 bootstrap samples

The upper and lower limits



Estimate: 0.719

Lower limit: 0.562
Upper limit: 0.875

Lower limit is above 0.5, so the wasps are not guessing!

Another example: species diversity

- Diversity of species at a site can be thought of as a combination of:
  - Species richness = the number of species there (the more species the more diverse)
  - Evenness = the relative number of each species (the more even the more diverse)
- Shannon-Wiener index combines these two characteristics into one value

$$H' = -\sum p_i \log p_i$$

K1

	A	B
1	Number	Species
2	1	Chamise
3	2	Chamise
4	3	Chamise
5	4	Chamise
6	5	Chamise
7	6	Chamise
8	7	Chamise
9	8	Chamise
10	9	Chamise
11	10	Chamise
12	11	Chamise
13	12	Chamise
14	13	Chamise
15	14	Chamise
16	15	Chamise
17	16	Chamise
18	17	Chamise
19	18	Chamise
20	19	Chamise
21	20	Chamise
22	21	Chamise
23	22	Chamise
24	23	Chamise
25	24	Chamise
26	25	Chamise
27	26	Chamise
28	27	Chamise
29	28	Chamise
30	29	Chamise
31	30	Chamise
32	31	Chamise
33	32	Chamise
34	33	White sage
35	34	White sage
36	35	White sage

Calculating the index

B8 f_x {=-SUM(B2:B6*LN(B2:B6))}

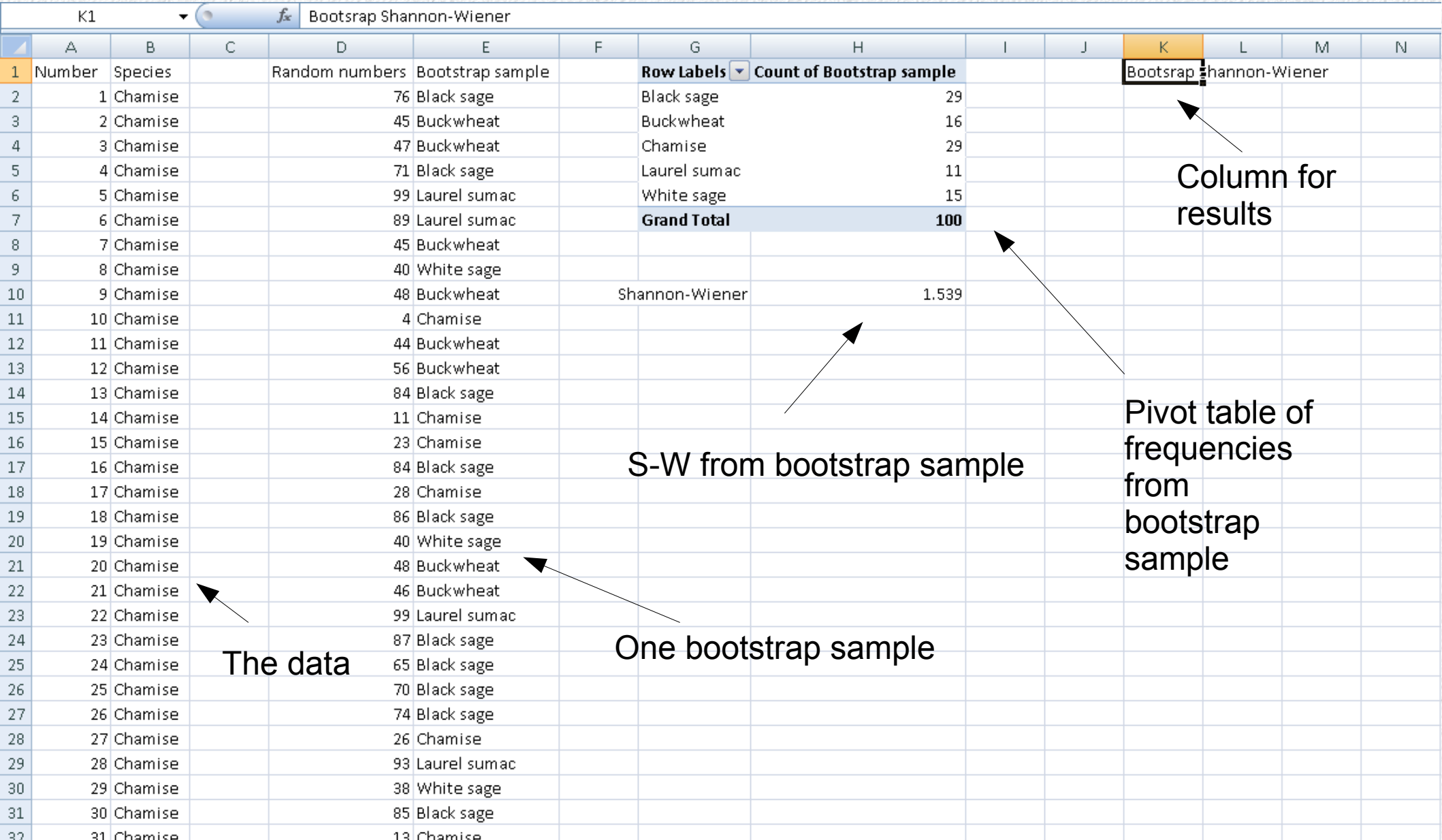
	A	B	C	D	E
1	Species	Relative frequency			
2	Chamise	0.32			
3	White sage	0.10			
4	Buckwheat	0.15			
5	Black sage	0.30			
6	Laurel sumac	0.13			
7					
8	Shannon-Wiener	1.506			
9					

- List of species, and the relative frequency of occurrence
- There were 100 total shrubs at the site

Calculating a confidence interval

- There is no formula for calculating the standard error for a Shannon index
- But, we can resample the data, and calculate the index each time
- Variation in the index can then be used to give us confidence intervals

Set up the sheet



K1 | Bootsrap Shannon-Wiener

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Number	Species		Random numbers	Bootstrap sample		Row Labels	Count of Bootstrap sample			Bootsrap	hannon-Wiener		
2	1	Chamise		76	Black sage		Black sage	29						
3	2	Chamise		45	Buckwheat		Buckwheat	16						
4	3	Chamise		47	Buckwheat		Chamise	29						
5	4	Chamise		71	Black sage		Laurel sumac	11						
6	5	Chamise		99	Laurel sumac		White sage	15						
7	6	Chamise		89	Laurel sumac		Grand Total	100						
8	7	Chamise		45	Buckwheat									
9	8	Chamise		40	White sage									
10	9	Chamise		48	Buckwheat		Shannon-Wiener	1.539						
11	10	Chamise		4	Chamise									
12	11	Chamise		44	Buckwheat									
13	12	Chamise		56	Buckwheat									
14	13	Chamise		84	Black sage									
15	14	Chamise		11	Chamise									
16	15	Chamise		23	Chamise									
17	16	Chamise		84	Black sage									
18	17	Chamise		28	Chamise									
19	18	Chamise		86	Black sage									
20	19	Chamise		40	White sage									
21	20	Chamise		48	Buckwheat									
22	21	Chamise		46	Buckwheat									
23	22	Chamise		99	Laurel sumac									
24	23	Chamise		87	Black sage									
25	24	Chamise		65	Black sage									
26	25	Chamise		70	Black sage									
27	26	Chamise		74	Black sage									
28	27	Chamise		26	Chamise									
29	28	Chamise		93	Laurel sumac									
30	29	Chamise		38	White sage									
31	30	Chamise		85	Black sage									
32	31	Chamise		13	Chamise									

Column for results

Pivot table of frequencies from bootstrap sample

S-W from bootstrap sample

One bootstrap sample

The data

Record the macro

- Each time, will need to:
  - Refresh the pivot table to reflect the new bootstrap sample – a new sample will automatically be generated when this is done for the next run
  - Copy the S-W value
  - Copy-Paste-special the value to column K

As recorded

```
Sub BootstrapSW()
'
' BootstrapSW Macro
' Bootstrap a confidence interval for the Shannon Wiener index.
'
' Keyboard Shortcut: Ctrl+Shift+S
'
    Range("H5").Select
    ActiveSheet.PivotTables("PivotTable1").PivotCache.Refresh
    Range("H10").Select
    Selection.Copy
    Range("K2").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
End Sub
```

Add a loop, record the results

```
Sub BootstrapSW()
'
' BootstrapSW Macro
' Bootstrap a confidence interval for the Shannon Wiener index.
'
' Keyboard Shortcut: Ctrl+Shift+S
'
For i = 1 To 1000
    ActiveSheet.PivotTables("PivotTable1").PivotCache.Refresh
    Range("H10").Select
    Selection.Copy
    Range("K" & i + 1).Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
Next i

End Sub
```

Run the macro, sort the output values, pick the endpoints

	J	K	L
	Order	Bootsrap Shannon-V	
	1	1.265197	
	2	1.278225	
	3	1.318246	
	4	1.319591	
	5	1.333683	
	6	1.336067	
	7	1.336185	
	8	1.344502	
	9	1.346924	
	10	1.355853	
	11	1.35641	
	12	1.35668	
	13	1.356981	
	14	1.358606	
	15	1.360736	
	16	1.361989	
	17	1.362091	
	18	1.362475	
	19	1.362633	
	20	1.364016	
	21	1.365962	
	22	1.366737	
	23	1.3715	
	24	1.371836	
	25	1.372095	
	26	1.372224	
	27	1.373081	
	28	1.373165	
	29	1.374082	
	972	1.562927	
	973	1.563441	
	974	1.563572	
	975	1.564675	
	976	1.564732	
	977	1.56513	
	978	1.56513	

Lower

Upper

Estimate: 1.539

Lower limit: 1.372
Upper limit: 1.565

Testing hypotheses with bootstrapping

- We used randomization testing last week to test hypotheses
- Can also bootstrap the difference between groups
- If the 95% CI of the differences doesn't include 0, the groups are different

Do these two locations have significantly different diversity?

	A	B	C	D	E
1		Site 1		Site 2	
2					
3	Species	Relative frequency		Relative frequency	
4	Chamise	0.32		0.40	
5	White sage	0.10		0.20	
6	Buckwheat	0.15		0.20	
7	Black sage	0.30		0.20	
8	Laurel sumac	0.13			
9					
10	Shannon-Wie	1.506		1.332	
11					
12					

Bootstrap the difference between them, see if the interval contains 0

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Site 1		Site 2		Site 1			Site 2			Site 1			Order	Difference
2	Number	Species	Species		Random numbers	Bootstrap sample		Random numbers	Bootstrap sample		**Row Labels**	**Count of Bootstrap sample**			
3	1	Chamise	Chamise		96	Laurel sumac		54	White sage		Black sage	39			
4	2	Chamise	Chamise		30	Chamise		70	Buckwheat		Buckwheat	12			
5	3	Chamise	Chamise		12	Chamise		49	White sage		Chamise	23			
6	4	Chamise	Chamise		74	Black sage		41	White sage		Laurel sumac	14			
7	5	Chamise	Chamise		76	Black sage		15	Chamise		White sage	12			
8	6	Chamise	Chamise		41	White sage		30	Chamise		**Grand Total**	**100**			
9	7	Chamise	Chamise		73	Black sage		12	Chamise						
10	8	Chamise	Chamise		26	Chamise		56	White sage						
11	9	Chamise	Chamise		71	Black sage		83	Black sage		Shannon-Wiener	1.489			
12	10	Chamise	Chamise		27	Chamise		45	White sage						
13	11	Chamise	Chamise		69	Black sage		16	Chamise						
14	12	Chamise	Chamise		77	Black sage		36	Chamise		Site 2				
15	13	Chamise	Chamise		22	Chamise		51	White sage		**Row Labels**	**Count of Bootstrap sample**			
16	14	Chamise	Chamise		61	Black sage		58	White sage		Black sage	16			
17	15	Chamise	Chamise		40	White sage		74	Buckwheat		Buckwheat	29			
18	16	Chamise	Chamise		26	Chamise		21	Chamise		Chamise	45			
19	17	Chamise	Chamise		42	White sage		22	Chamise		White sage	10			
20	18	Chamise	Chamise		76	Black sage		57	White sage		**Grand Total**	**100**			
21	19	Chamise	Chamise		50	Buckwheat		83	Black sage						
22	20	Chamise	Chamise		26	Chamise		25	Chamise						
23	21	Chamise	Chamise		48	Buckwheat		21	Chamise		Shannon-Wiener	1.242			
24	22	Chamise	Chamise		87	Black sage		79	Buckwheat						
25	23	Chamise	Chamise		5	Chamise		73	Buckwheat						
26	24	Chamise	Chamise		58	Black sage		33	Chamise		Difference in diversity	0.248			
27	25	Chamise	Chamise		39	White sage		44	White sage						
28	26	Chamise	Chamise		61	Black sage		70	Buckwheat						
29	27	Chamise	Chamise		79	Black sage		33	Chamise						
30	28	Chamise	Chamise		21	Chamise		99	Black sage						
31	29	Chamise	Chamise		4	Chamise		33	Chamise						
32	30	Chamise	Chamise		78	Black sage		95	Black sage						
33	31	Chamise	Chamise		92	Laurel sumac		28	Chamise						
34	32	Chamise	Chamise		72	Black sage		76	Buckwheat						
35	33	White sag	Chamise		50	Buckwheat		19	Chamise						
36	34	White sag	Chamise		70	Black sage		20	Chamise						
37	35	White sag	Chamise		82	Black sage		66	Buckwheat						
38	36	White sag	Chamise		90	Laurel sumac		37	Chamise						

Macro recorder

- Like before, except...
  - Refresh both pivot tables
  - Save a copy of the difference each time

As recorded

```
Sub DiversityDifference()
'
' DiversityDifference Macro
' Difference in diversity between two sites
'
' Keyboard Shortcut: Ctrl+Shift+D
'
    Range("L4").Select
    ActiveSheet.PivotTables("PivotTable1").PivotCache.Refresh
    Range("L17").Select
    ActiveSheet.PivotTables("PivotTable2").PivotCache.Refresh
    Range("L26").Select
    Selection.Copy
    Range("O2").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
End Sub
```

Add a loop, record results each time

```
Sub DiversityDifference()
'
' DiversityDifference Macro
' Difference in diversity between two sites
'
' Keyboard Shortcut: Ctrl+Shift+D
'
For i = 1 To 1000
    ActiveSheet.PivotTables("PivotTable1").PivotCache.Refresh
    ActiveSheet.PivotTables("PivotTable2").PivotCache.Refresh
    Range("L26").Select
    Selection.Copy
    Range("O" & i + 1).Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
Next i

End Sub
```

Sort results, find the endpoints

M	N	O	P
	Order	Difference	
	1	-0.01889	
	2	-0.00709	
	3	0.010641	
	4	0.012019	
	5	0.018965	
	6	0.02457	
	7	0.025945	
	8	0.028224	
	9	0.028975	
	10	0.034436	
	11	0.035053	
	12	0.03512	
	13	0.035837	
	14	0.038098	
	15	0.038373	
	16	0.040118	
	17	0.041657	
	18	0.041712	
	19	0.04396	
	20	0.044421	
	21	0.044893	
	22	0.047778	
	23	0.047795	
	24	0.047975	
	25	0.051934	
	26	0.052551	
	27	0.053051	
	973	0.275159	
	974	0.27543	
	975	0.275549	
	976	0.275576	
	977	0.275872	

Zero difference is not within the confidence interval, so site 1 is significantly more diverse than site 2