Programming II

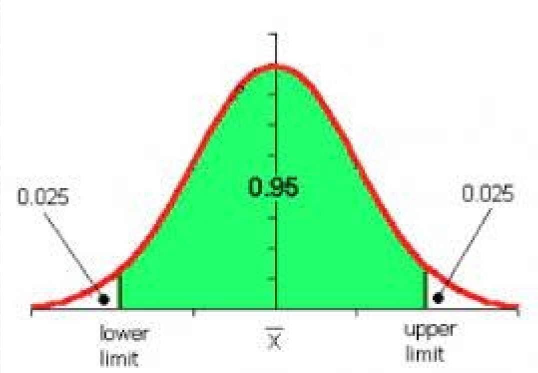
Bootstrapping: using the computer to make hairy statistics easy

Confidence intervals

- Confidence intervals are measures of uncertainty in estimates
- For example, if you want to know how tall people are:
  - Collect a data set and calculate a mean ($\bar{x}$) to estimate the mean for the population (μ)
  - We know that a different set of data will give us a different mean – the mean of our particular sample is probably not exactly equal to the mean for all people
  - The question is, how close to the actual population mean is the sample mean?
- Confidence intervals around the estimate give us information about where the population is expected to be, given the variability in our sampling

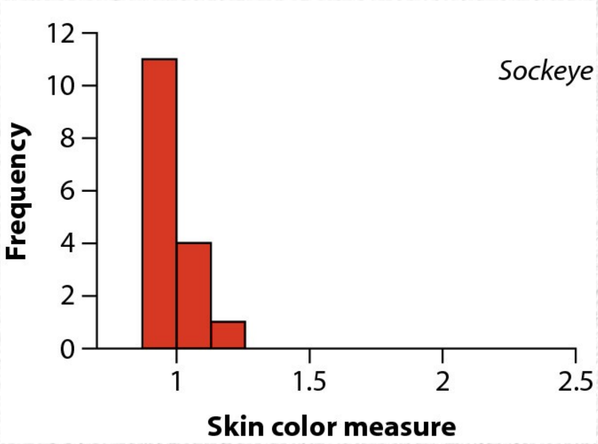
The usual method of calculating confidence intervals

- Neyman-Pearson confidence intervals – work great for a wide range of conditions
- From a sample of data, calculate the mean, standard deviation, and standard error
- Calculate the upper and lower limits of a (usually 95%) confidence interval as $\bar{x} \pm t\, s_{\bar{x}}$
- Uncertainty = $t\, s_{\bar{x}}$
  - Add/subtract uncertainty from the mean → symmetrical interval (2.5% above, 2.5% below the mean excluded from the interval)
  - Standard error is $s/\sqrt{n}$ → need to be able to calculate the standard deviation to calculate the confidence interval
- Sometimes it doesn't work well, sometimes it's not possible to use it at all



Example: sockeye skin color

- Recall that the sockeye salmon skin color data look pretty right-skewed
- Basement at low levels → asymmetrical confidence interval
  - Skin color measurements can be higher than observed by a lot
  - Can’t be lower than observed by much
  - Using symmetrical confidence intervals may not represent the uncertainty in the estimate very well



Bootstrapping the confidence interval for a proportion

- Instead of an imperfect analytical solution, we can estimate the interval numerically by resampling
- We'll find the confidence interval by:
  - Randomly selecting *with replacement* from the observed data
  - Record the mean of the randomly selected data
  - Repeat many times (at least 1000)
  - The 2.5 percentile and 97.5 percentile of proportions from the 1000 resampled data sets are the 95% confidence limits

Setting up the worksheet

- To randomly select the data with replacement…
  - Select a random number from 1 to 16 for each of the 16 rows of the bootstrap sample
  - Use the random numbers to look up the data value
- The mean of each bootstrap sample is recorded

Generated random number

Look up the number in the “Number” column

	A	B	C	D	E	F
1	Sockeye number	skin color		Random salmon numer	Bootstrap sample	
2	1	0.98		=RANDBETWEEN(1,16)	=LOOKUP(D2,A$2:A$17,B$2:B$17)	
3	2	0.88		5	1.02	
4	3	0.97		10	1.03	
5	4	0.99		2	0.88	
6	5	1.02		7	0.99	
7	6	1.03		11	1.08	
8	7	0.99		11	1.08	
9	8	0.97		8	0.97	
10	9	0.98		9	0.98	
11	10	1.03		15	0.94	
12	11	1.08		15	0.94	
13	12	1.15		8	0.97	
14	13	0.9		14	0.95	
15	14	0.95		9	0.98	
16	15	0.94		2	0.88	
17	16	0.99		14	0.95	
18						
19	Mean	0.990625		Bootstrap mean	0.976	

Return the contents of the “Wasp selections” column

Repeat 1000 times

- Each time a mean is copied/pasted the worksheet recalculates → new bootstrap sample
- We just need to copy/paste 1000 times

The macro

```
Sub BootstrapCI()
'
' BootstrapCI Macro
'
' Keyboard Shortcut: Ctrl+Shift+B
'

Application.ScreenUpdating = False

For i = 1 To 1000
    Range("G" & i + 1) = Range("E19").Value
Next i

Columns("G").Sort key1:=Range("G2"), order1:=xlAscending, Header:=xlYes

Range("B20") = Range("G26").Value
Range("B21") = Range("G976").Value

Application.ScreenUpdating = True

End Sub
```

Enter the bootstrap mean into column G

After the loop is done, sort the bootstrap means in G

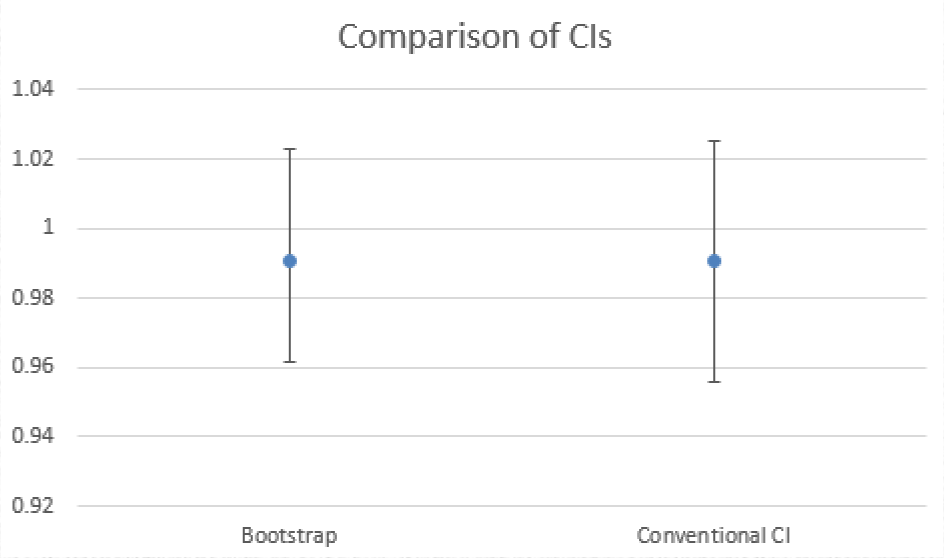
Record the lower and upper limits in B20 and B21

	A	B	C	D	E	F	G	H
1	Sockeye number	skin color		Random salmon numer	Bootstrap sample		Bootstrap means	
2	1	0.98		6	1.03		0.943125	
3	2	0.88		9	0.98		0.94375	
4	3	0.97		9	0.98		0.95	
5	4	0.99		13	0.9		0.95125	
6	5	1.02		16	0.99		0.95125	
7	6	1.03		9	0.98		0.955	
8	7	0.99		3	0.97		0.955	
9	8	0.97		8	0.97		0.955625	
10	9	0.98		2	0.88		0.955625	
11	10	1.03		6	1.03		0.955625	
12	11	1.08		12	1.15		0.955625	
13	12	1.15		7	0.99		0.95625	
14	13	0.9		9	0.98		0.956875	
15	14	0.95		11	1.08		0.956875	
16	15	0.94		15	0.94		0.958125	
17	16	0.99		4	0.99		0.958125	
18							0.959375	
19	Mean	0.990625		Bootstrap mean	0.99		0.96	
20	Lower limit	0.961875					0.96	
21	Upper limit	1.023125					0.96	
22							0.96	
23							0.96	
24							0.960625	
25							0.96125	
26							0.961875	
27							0.961875	
975							1.023125	
976							1.023125	

25th mean is the 2.5th percentile = lower limit

975th mean is the 97.5th percentile = upper limit

How does it work?



Little smaller, slightly asymmetrical
(0.029 below mean, 0.033 above)

Little larger, symmetrical
(0.035 above and below mean)

Now it’s your turn

- You will bootstrap the confidence interval for these data today