Sampling and Standard Error

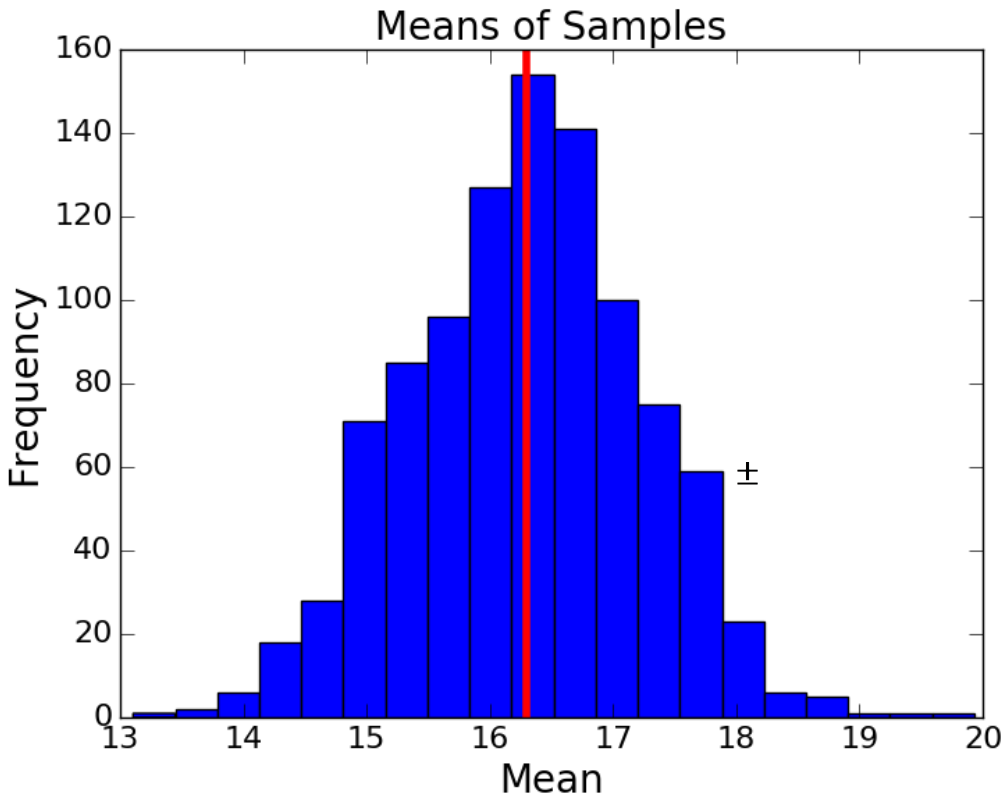
Means and Standard Deviations

- Population mean = 16.3
- Sample mean = 17.1
- Standard deviation of population = 9.44
- Standard deviation of sample = 10.4
- A happy accident, or something we should expect?
- Let's try it 1000 times and plot the results

Notice in Code

- pylab.axvline(x = popMean, color = 'r') draws a red vertical line at popMean on the x-axis
- There's also a pylab.axhline function

Try It 1000 Times



What's the 95% confidence interval?

16.28 +- 1.96*0.94
14.5 - 18.1
Includes population mean

Suppose we want a tighter bound?

Mean of sample Means = 16.3
Standard deviation of sample means = 0.94
Maximum difference in means = 3.63
Maximum difference in standard deviations = 2.46

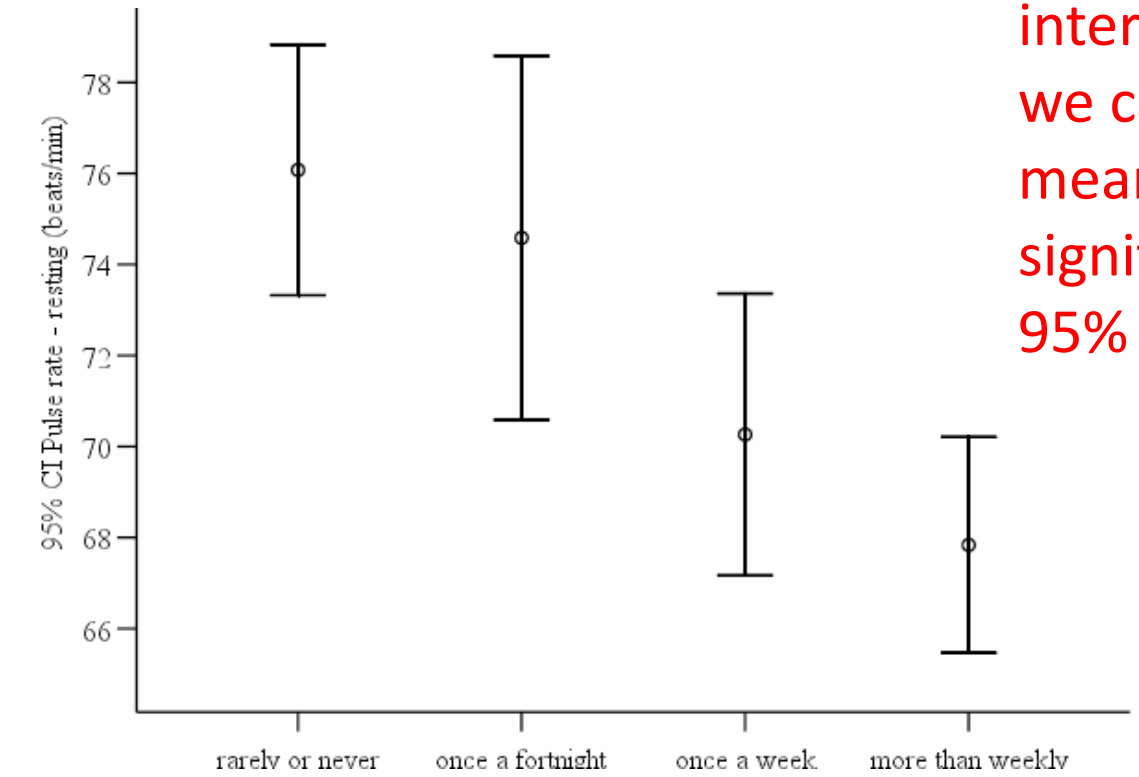
Getting a Tighter Bound

- Will drawing more samples help?
  - Let's try increasing from 1000 to 2000
- How about larger samples?
  - Let's try increasing sample size from 100 to 200
  - Standard deviation of sample means drops from 0.94 to 0.66

Error Bars, a Digression

- Graphical representation of the variability of data
- Way to visualize uncertainty

When confidence intervals don't overlap, we can conclude that means are statistically significantly different at 95% level.

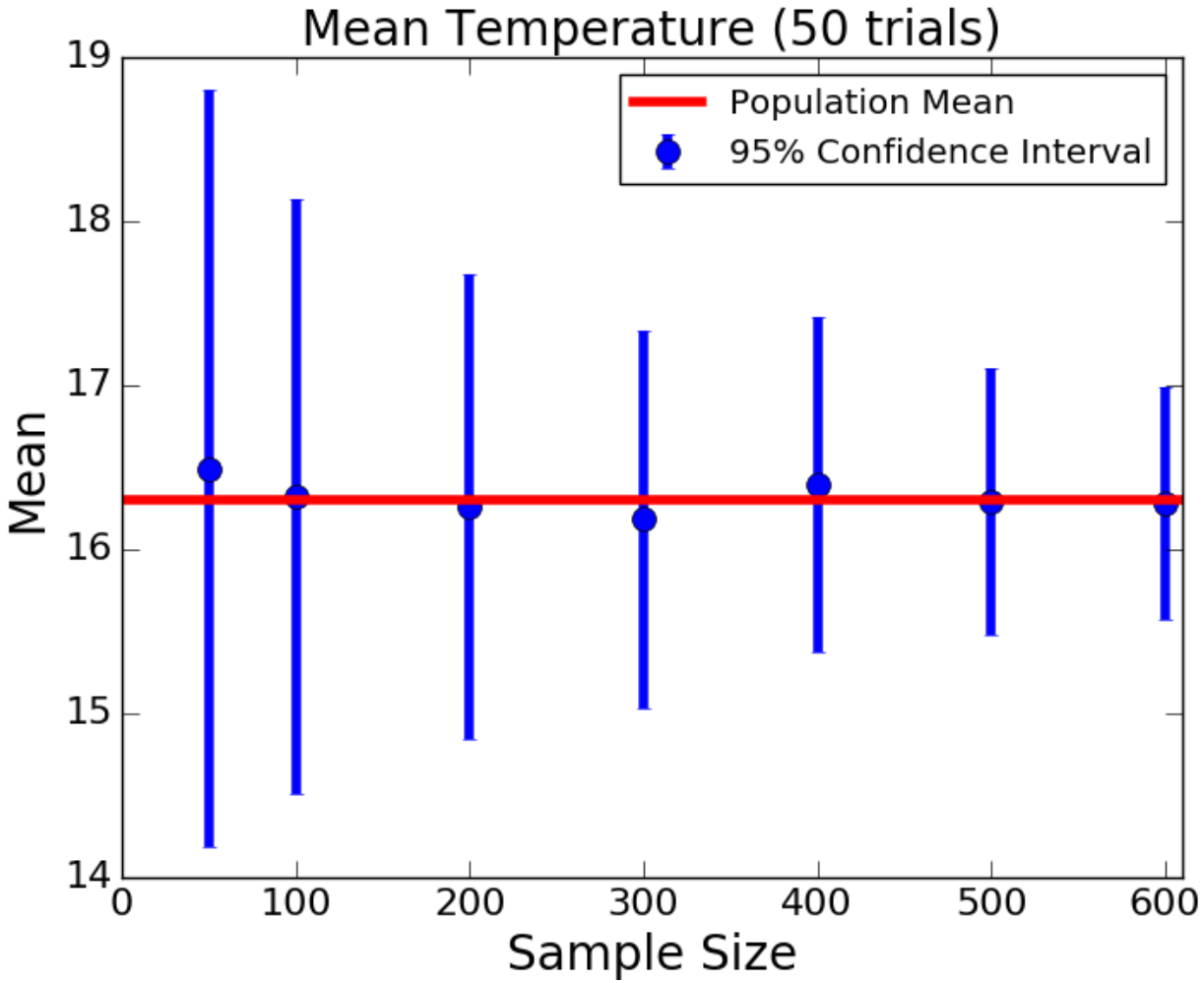


https://upload.wikimedia.org/wikipedia/commons/1/1d/Pulse_Rate_Error_Bar_By_Exercise_Level.png

Key Line of Code

```
pylab.errorbar(xVals, sizeMeans,
               yerr = 1.96*pylab.array(sizeSDs),
               fmt = 'o',
               label = '95% Confidence Interval')
```

Sample Size and Standard Deviation



Bigger Seems to Be Better

- Going from a sample size of 100 to 400 reduced the confidence interval from 1.8C to about 1C.
- But we are now looking at 400,000 examples
  - What has sampling bought us?
  - Absolutely Nothing!

What Can We Conclude from 1 Sample?

- More than you might think