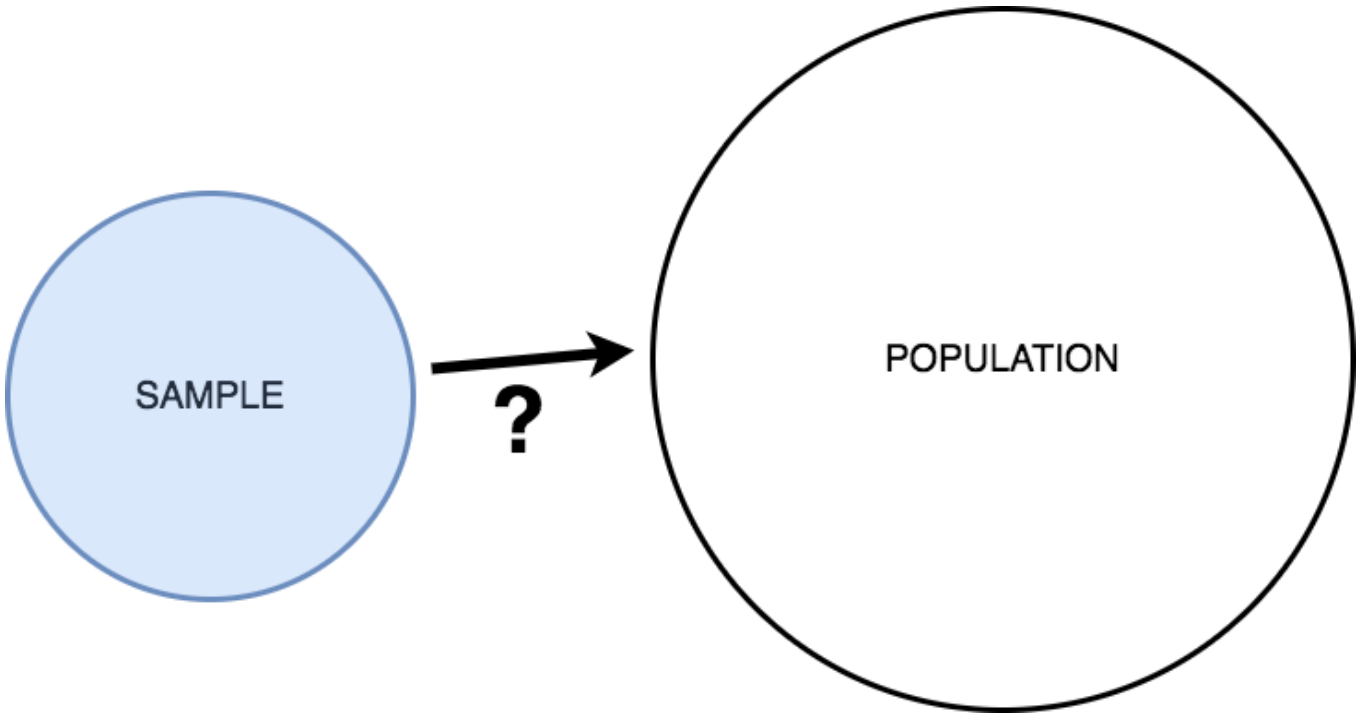
A Statistician's Approach to Model Selection

Erle Holgersen

November 23rd, 2017

Introduction to Statistics



What Does This Mean for Model Selection?

- Can always find a better model in our dataset

What Does This Mean for Model Selection?

- Can always find a better model in our dataset
- Will a model perform better in the population at large?

Basic Toolbox

- Likelihood ratio tests
- Information criteria
- Model evaluation metrics

Likelihood Ratio Tests

- For nested models
- Based on asymptotic theory about likelihood functions
- Limited to models where you have a likelihood function
    - i.e. regression models

```
> anova(m2, m1, test = 'LRT')
Analysis of Variance Table

Model 1: peri ~ shape + perm
Model 2: peri ~ shape
  Res.Df      RSS Df Sum of Sq Pr(>Chi)
1     45 43697562
2     46 78261734 -1 -34564173 2.43e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Information Criteria

- Also used for non-nested models
- Penalize extra terms in the model
- Examples: AIC, BIC

```
> AIC(m1)
[1] 828.8278
> AIC(m2)
[1] 802.855
```

Model Evaluation Metrics

- One number to summarize model performance
  - e.g. area under the curve (AUC), R^2
- Always improves with more parameters in the model
- Need to estimate variability of the metric to be able to say anything about real improvement
  - bootstrapping

Conclusion

- Many data science applications have much larger samples than traditional statistics problems
- Samples still tend to be limited in some way, often in time
- Can use the same model selection techniques