From Data to Knowledge: A peek under the hood of statistics

Author: Nicholas G Reich

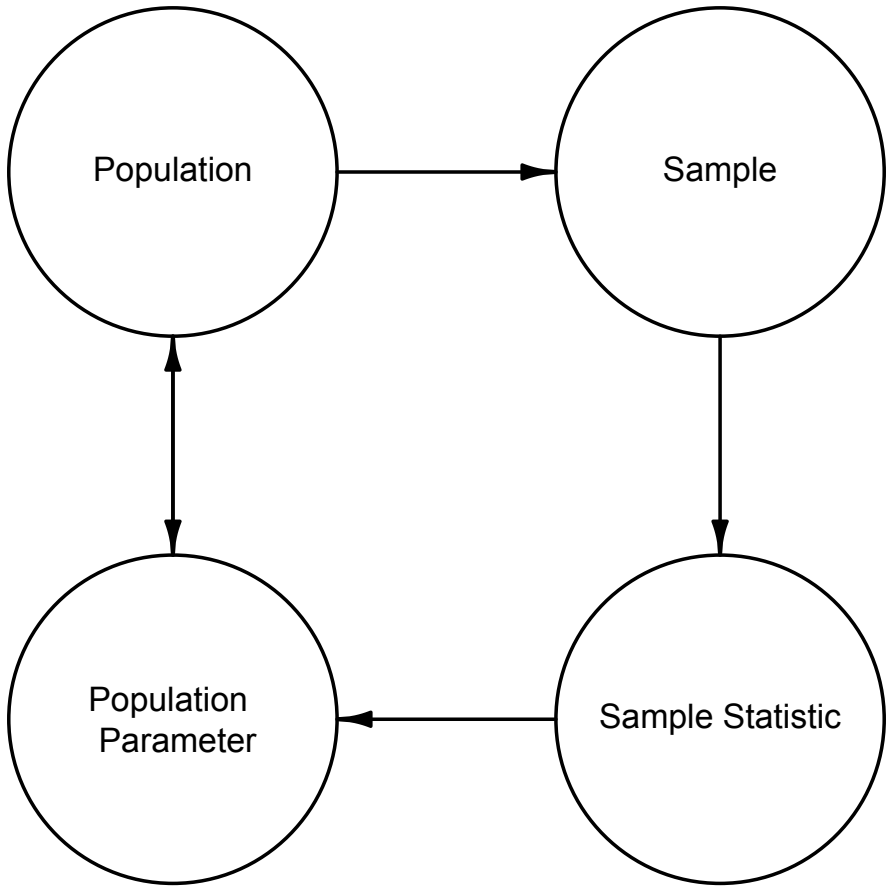
This material is part of the **statsTeachR** *project*

Why do we do statistics

- Statistics is the science of turning data into knowledge.
- Knowing what you do not know is one the most important traits as a scientist/seeker of knowledge through data.

Estimation vs. inference

- Statistical estimation (e.g. the method of least-squares) gives us our best guess at a parameter.
- Inference tells us how certain we should be about these estimates.



There isn't one accepted way of learning from data

Over the next two weeks, we are going to look at a few different methods for measuring uncertainty in relationships that we see in data.

Relationships are characterized by parameters in our models.

These are some tools we use to characterize our uncertainty about these parameters:

- Likelihood
- Posterior distributions
- Null distributions
- Sampling distributions ("classical" approach)

There is no one "best" approach. Some will be right for some circumstances, not right for others. Each make different assumptions.

Using models to learn about the world

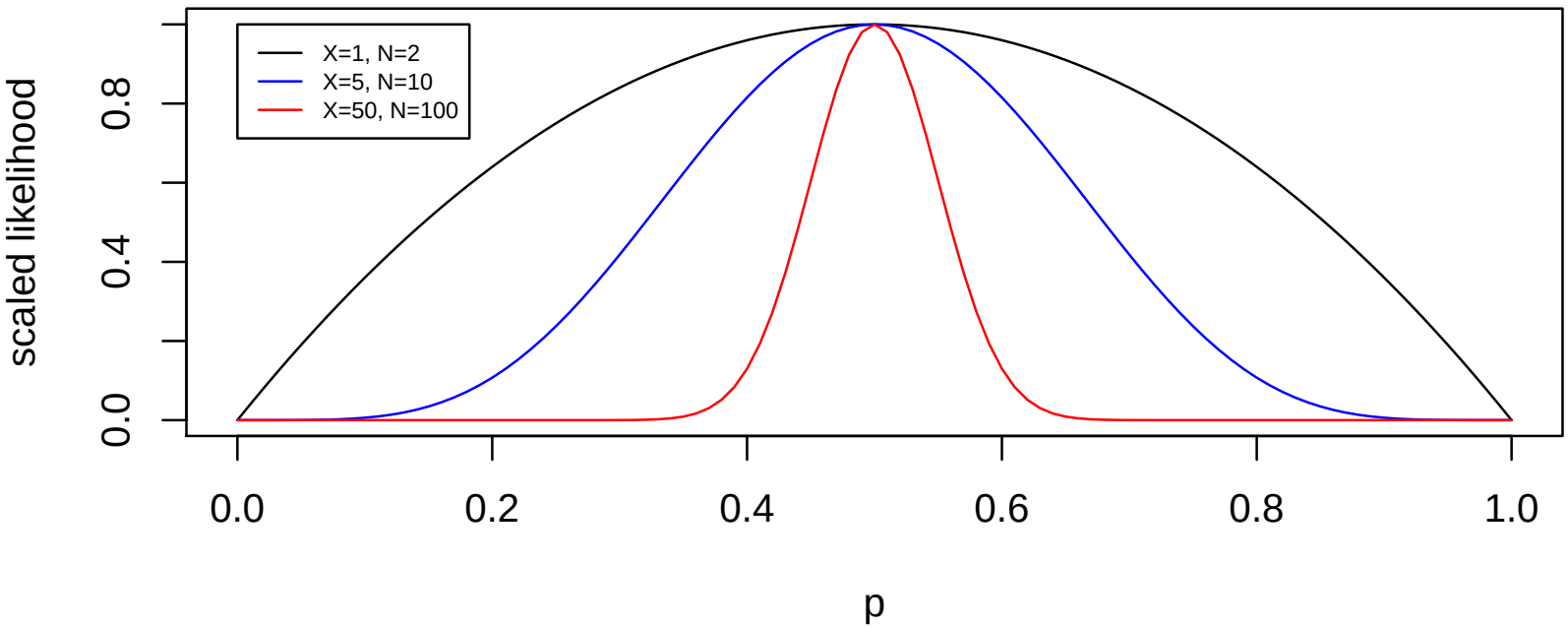
- One model might describe the relationship between smoking status and forced expiratory volume.
- Another model could describe the probability that this coin will land heads.

Underlying most models is a likelihood

- Likelihood is a mathematical function that gives you the likelihood of a particular parameter given the data you have seen.
- In regression, when you find the "least squares" parameters that minimize the residual sum of squares, these also maximize the likelihood function. They are mathematically equivalent.
- Likelihood is driven by assumptions that you make about the distribution and structure of your data. e.g. residuals follow a Normal distribution, coin flips follow a binomial distribution.

Likelihood for coin-flipping

$$L(p|X = 5, n = 10) = \binom{10}{5} \cdot p^5 \cdot (1 - p^{10-5})$$



Maximum for all curves occurs when $p = .5$.
The more pointy the likelihood, the more knowledge you have about the parameter.
Binomial App: http://shiny.stat.calpoly.edu/MLE_Binomial/

Schools of inference

Most (not all) statisticians use likelihood to translate data into knowledge.

- Frequentists use the likelihood to approximate a sampling distribution.
- Bayesians modify the likelihood based on prior belief to create a posterior distribution.

Bayesian thinking

A Bayesian incorporates prior belief into the likelihood. What is your prior belief about what this parameter is?

- Based on prior scientific studies or observations. e.g. "The laws of physics dictate that this coin is more or less fair, so I think the probability of getting a head should be about 0.5."
- Little or no knowledge can be described as having a uniform prior. In this case, Bayesian inference is equivalent to just using the likelihood.

Let's establish the classes prior beliefs

Go here to submit your guess: https://goo.gl/RKW8dJ
We are going to use these guesses to create a prior distribution for the collective belief in the class about this coin.

```
probs <- read.csv(file="https://goo.gl/jNDbrv")
(probs$prob)

library(MASS)
fitdistr(probs$prob, "beta", list(shape1=1,shape2=1))
```

Now, update, using Bayesian reasoning

Every coin flip we observe will update the likelihood and therefore the posterior distribution as well.
Link to app

We just used Bayesian reasoning to learn about the probability that this coin lands heads.
Now we are going to use a different kind of statistical reasoning to evaluate a similar question.
Go to the class activity for today.