Statistical Inference

Statistics, in a nutshell

- Descriptive statistics: analysis of *observed data only*
  - Computing numerical summaries, using visualization tools, etc.
  - Provides a description of the data, as the name suggests
  - Example: Can be used to compare the average GPAs of students across all the Ivies, by calculating the average of *all* Brown students, *all* Yale students, *all* Columbia students, etc.
- Inferential statistics: analysis of observed data, leading to conclusions about *unobserved data*
  - Used to make reasonable guesses about a population from a data sample
  - Example: Take a random sample of Brown students, calculate their average GPA.
    Use inferential statistics to make estimates and test hypotheses about all Brown students based on this sample.
  - Caveat: sample must be representative of the population (or the inferences will not be valid)
  - There is always sampling error: hence, there is always uncertainty in inferential statistics

Statistical nomenclature

- Statistic: a measure that describes the sample
- Parameter: a measure that describes the population
- Sampling error: the difference between a statistic and a parameter

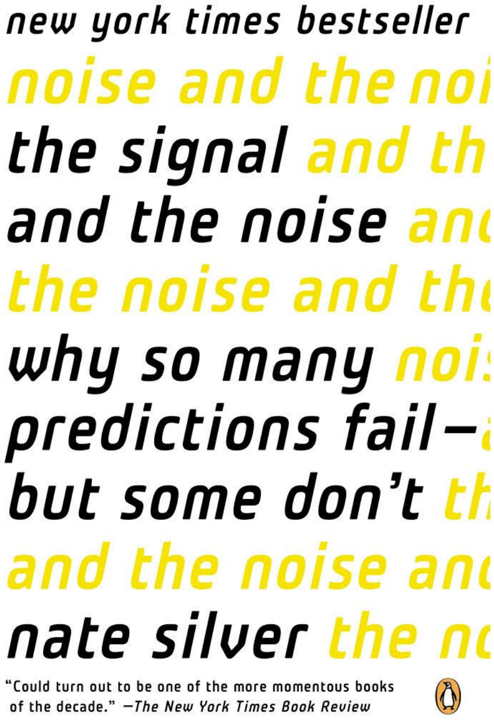
- Point estimate: a single value estimate of a parameter
- Interval estimate: a range of estimated values of a parameter
    - E.g., confidence intervals
- Both forms of estimates are often used in concert with one another

- Hypothesis testing: used to test hypotheses based on samples
    - Comparison tests: assess differences in means, medians, etc. among different groups
    - Correlation tests: test relationships among variables
    - Regression tests: test whether changes in one variable cause changes in another

The Signal vs. the Noise

As Nate Silver will tell you, the difficulty in statistical inference is separating the signal from the noise.

- The signal is meaningful information.
- The noise is random fluctuation.
- If we flip a fair coin, it is possible that the outcome will be a sequence of all heads, just due to random chance.
- If this is the only sample that we see, then how can we separate the signal—the coin is fair—from the noise—the sequence of all heads that we observe.



Paul the Octopus

Paul the Octopus

- An animal oracle who successfully predicted the outcome of all 7 of Germany’s matches in the 2010 World Cup
- He also predicted Spain to win the Cup final

Paul the Octopus

- An animal oracle who successfully predicted the outcome of all 7 of Germany’s matches in the 2010 World Cup
- He also predicted Spain to win the Cup final
- How likely is his success rate?

```
> dbinom(7, 7, 1/2)
[1] 0.0078125
```

Paul the Octopus

- An animal oracle who successfully predicted the outcome of all 7 of Germany's matches in the 2010 World Cup
- He also predicted Spain to win the Cup final
- How likely is his success rate?

```
> dbinom(7, 7, 1/2)
[1] 0.0078125
```

- Other factors: He chose Germany 11/14 times, Spain (twice), and Serbia once. Similar flags. Likely color blind, but perhaps has preference for horizontal shapes.

Image Source

Statistical Modeling

Statistics, in a nutshell

- Descriptive statistics: analysis of *observed data only*
  - Computing numerical summaries, using visualization tools, etc.
- Probability theory: idealized descriptions of unobserved (imagined) data
- Inferential statistics: analysis of observed data, leading to conclusions about *unobserved data*
  - Assume a probabilistic model
  - Estimate the parameters of the model from data
  - Use the ensuing statistical model to draw inferences
    1. E.g., The treatment was effective (or not)
  - The model also allows us to precisely quantify the uncertainty in these inferences
    1. E.g., at the 95% confidence level

Statistical Modeling

- Design a probabilistic model of the data
  - Specify the variables of interest, how they are distributed, and how they relate to one another
- Calculate statistics (i.e., estimation)
  - Because statistics are functions of data, a model also specifies how statistics are distributed
- The distributions over statistics allow us to draw statistical inferences (i.e., inferences with quantifiable uncertainty)

Is it really that easy?

- Model building is an art
  - Exploratory data analysis is a good way to start
  - But it also requires domain knowledge (talking to experts, literature reviews, etc.)
- Given a model, you can (in principle) turn a crank, and draw conclusions: i.e., estimation and inference are each a science
- The stronger the assumptions, the stronger the conclusions
- "There is no virtue to strong conclusions which rest on faulty premises." —Cosma Shalizi

Statistical Modeling

- Part I: Model building
- Part II: Estimation
- Part III: Inference
- Part IV: Model checking
    - Verify that the assumptions of the model hold
    - Modify any that are very wrong, and can easily be modified
    - Qualify any conclusions in light of any false assumptions

“All models are wrong, but some are useful.” -- George Box

Formalization and Examples

Statistical Model: Definition

A statistical model is a set of probabilistic assumptions about how data are generated.

It consists of a sample space S—the set of possible observations—together with a set of probability distributions $\mathcal{P}$ over S.

The probability distributions are parameterized: $\mathcal{P} = \{ P_\theta \mid \theta \in \Theta \}$.

$\mathcal{P}$ need not contain the true distribution:
“All models are wrong, but some are useful.” -- George Box

Statistical Model: Definition & Example

A statistical model is a set of probabilistic assumptions about how data are generated.

It consists of a sample space S—the set of possible observations—together with a set of probability distributions $\mathcal{P}$ over S.

The probability distributions are parameterized: $\mathcal{P} = \{ P_\theta \mid \theta \in \Theta \}$.

For example, consider a sequence of coin flips.

- The sample space is all possible sequences of heads and tails.
- The probability distributions are parameterized by all possible biases of the coin towards heads: $p_H = 0.01$, $p_H = 0.1$, $p_H = 0.15$, etc. So $\Theta = [0, 1]$.

Statistical Model: More Examples

A statistical model is a set of probabilistic assumptions about how data are generated.

- $Y = \mu + \varepsilon$, where μ is an unknown model parameter representing the mean, and ε is a random error term (*a.k.a.* noise) representing everything else.
  - Here, the sample space consists of individual observations, such as heights or weights.
- $Y = \beta_0 + \beta_1 X + \varepsilon$, where β_0 and β_1 are unknown model parameters s.t. $\beta_0 + \beta_1 x$ represents the mean, given $X = x$, and ε is a random error term (*a.k.a.* noise).
  - Here, the sample space consists of observation pairs, such as heights and weights.
- In both models, assumptions about ε, such as it has mean zero and variance σ, define the probability distributions over outcomes.