

# Data Science Inference and Modeling

The textbook for the Data Science course series is [freely available online](#).

This course corresponds to the textbook chapters [Statistical Inference](#) and [Statistical Models](#).

## Learning Objectives

- The concepts necessary to define estimates and margins of errors of populations, parameters, estimates, and standard errors in order to make predictions about data
- How to use models to aggregate data from different sources
- The very basics of Bayesian statistics and predictive modeling

## Course Overview

### Section 1: Parameters and Estimates

You will learn how to estimate population parameters.

### Section 2: The Central Limit Theorem in Practice

You will apply the central limit theorem to assess how close a sample estimate is to the population parameter of interest.

### Section 3: Confidence Intervals and p-Values

You will learn how to calculate confidence intervals and learn about the relationship between confidence intervals and p-values.

### Section 4: Statistical Models

You will learn about statistical models in the context of election forecasting.

### Section 5: Bayesian Statistics

You will learn about Bayesian statistics through looking at examples from rare disease diagnosis and baseball.

### Section 6: Election Forecasting

You will learn about election forecasting, building on what you've learned in the previous sections about statistical modeling and Bayesian statistics.

## Section 7: Association Tests

You will learn how to use association and chi-squared tests to perform inference for binary, categorical, and ordinal data through an example looking at research funding rates.

## Introduction to Inference

The textbook for this section is available [here](#)

In this course, we will learn:

- *statistical inference*, the process of deducing characteristics of a population using data from a random sample
- the statistical concepts necessary to define *estimates* and *margins of errors*
- how to *forecast future results* and estimate the precision of our forecast
- how to calculate and interpret *confidence intervals* and *p-values*

### Key points

- Information gathered from a small random sample can be used to infer characteristics of the entire population.
- Opinion polls are useful when asking everyone in the population is impossible.
- A common use for opinion polls is determining voter preferences in political elections for the purposes of forecasting election results.
- The *spread* of a poll is the estimated difference between support two candidates or options.

## Section 1 Overview

Section 1 introduces you to parameters and estimates.

After completing Section 1, you will be able to:

- Understand how to use a sampling model to perform a poll.
- Explain the terms **population**, **parameter**, and **sample** as they relate to statistical inference.
- Use a sample to estimate the population proportion from the sample average.
- Calculate the expected value and standard error of the sample average.

## Sampling Model Parameters and Estimates

The textbook for this section is available [here](#) and [here; first part](#)

### Key points

- The task of statistical inference is to estimate an unknown population parameter using observed data from a sample.
- In a sampling model, the collection of elements in the urn is called the *population*.
- A *parameter* is a number that summarizes data for an entire population.
- A *sample* is observed data from a subset of the population.
- An *estimate* is a summary of the observed data about a parameter that we believe is informative. It is a data-driven guess of the population parameter.
- We want to predict the proportion of the blue beads in the urn, the parameter  $p$ . The proportion of red beads in the urn is  $1 - p$  and the *spread* is  $2p - 1$ .

- The sample proportion is a random variable. Sampling gives random results drawn from the population distribution.

*Code: Function for taking a random draw from a specific urn*

The **dslabs** package includes a function for taking a random draw of size  $n$  from the urn:

```
if(!require(tidyverse)) install.packages("tidyverse")
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages -----
```

```
## v ggplot2 3.3.2      v purrr   0.3.4
## v tibble  3.0.3      v dplyr   1.0.1
## v tidyr   1.1.1      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.5.0
```

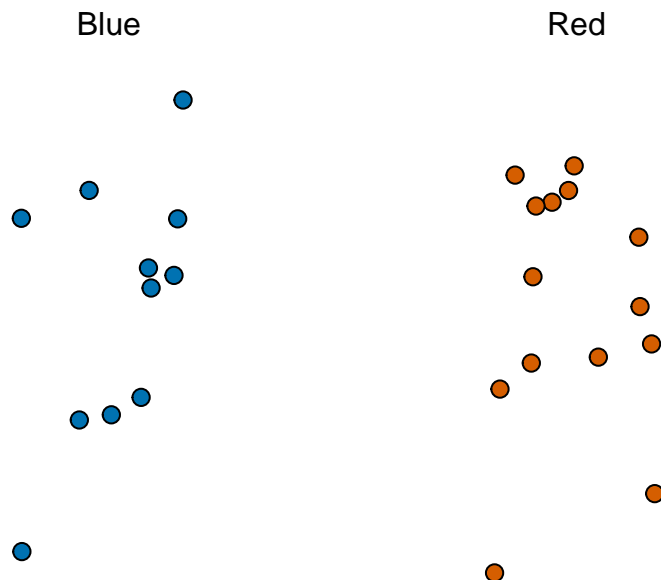
```
## -- Conflicts -----
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
if(!require(dslabs)) install.packages("dslabs")
```

```
## Loading required package: dslabs
```

```
library(tidyverse)
library(dslabs)
take_poll(25)      # draw 25 beads
```



## The Sample Average

The textbook for this section is available [here](#) and [here](#)

### Key points

- Many common data science tasks can be framed as estimating a parameter from a sample.
- We illustrate statistical inference by walking through the process to estimate  $p$ . From the estimate of  $p$ , we can easily calculate an estimate of the spread,  $2p - 1$ .
- Consider the random variable  $X$  that is 1 if a blue bead is chosen and 0 if a red bead is chosen. The proportion of blue beads in  $N$  draws is the average of the draws  $X_1, \dots, X_N$ .
- $\bar{X}$  is the *sample average*. In statistics, a bar on top of a symbol denotes the average.  $\bar{X}$  is a random variable because it is the average of random draws - each time we take a sample,  $\bar{X}$  is different.

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N}$$

- The number of blue beads drawn in  $N$  draws,  $N\bar{X}$ , is  $N$  times the proportion of values in the urn. However, we do not know the true proportion: we are trying to estimate this parameter  $p$ .

## Polling versus Forecasting

The textbook for this section is available [here](#)

### Key points

- A poll taken in advance of an election estimates  $p$  for that moment, not for election day.
- In order to predict election results, forecasters try to use early estimates of  $p$  to predict  $p$  on election day. We discuss some approaches in later sections.

## Assessment 1.1: Parameters and Estimates

### 1. Polling - expected value of $S$

Suppose you poll a population in which a proportion  $p$  of voters are Democrats and  $1-p$  are Republicans. Your sample size is  $N=25$ . Consider the random variable  $S$ , which is the total number of Democrats in your sample.

What is the expected value of this random variable  $S$ ?

Possible Answers - [ ] A.  $E(S)=25(1-p)$  - [X] B.  $E(S)=25p$  - [ ] C.  $E(S)=\sqrt{(25p(1-p))}$  - [ ] D.  $E(S)=p$

### 2. Polling - standard error of $S$

Again, consider the random variable  $S$ , which is the total number of Democrats in your sample of 25 voters. The variable  $p$  describes the proportion of Democrats in the sample, whereas  $1-p$  describes the proportion of Republicans.

What is the standard error of  $S$ ?

Possible Answers - [ ] A.  $SE(S)=25p(1-p)$  - [ ] B.  $SE(S)=\sqrt{25p}$  - [ ] C.  $SE(S)=25(1-p)$  - [X] D.  $SE(S)=\sqrt{(25p(1-p))}$

### 3. Polling - expected value of $\bar{X}$