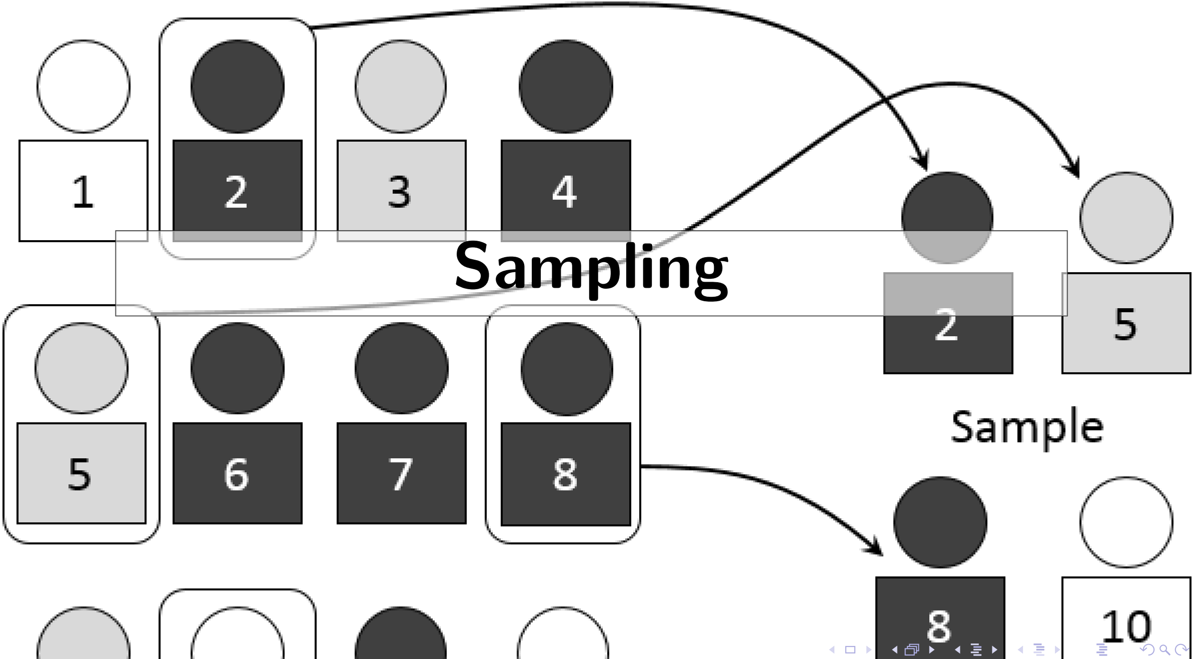
Sampling



Today:

- What is sampling and what do we use it to do?
- What is the effect of sampling variation/sample size on an estimate ?
- What can go wrong with sampling and how can we fix it?

Definitions

- **Population**: a ‘complete’ group of N objects, items, entities, or events of interest – e.g. all adults living in the US;
- **Sample**: a selected subset of n individuals from a population – e.g. 5,000 US adults appearing in a poll;
- **Summary Statistic**: a summary of the information in a set of observations – e.g. mean, median, mode, etc.;

Definitions

- **Population**: a ‘complete’ group of N objects, items, entities, or events of interest – e.g. all adults living in the US;
- **Sample**: a selected subset of n individuals from a population – e.g. 5,000 US adults appearing in a poll;
- **Summary Statistic**: a summary of the information in a set of observations – e.g. mean, median, mode, etc.;
- **Census**: a counting of all elements of the population;

Definitions

- **Population**: a ‘complete’ group of N objects, items, entities, or events of interest – e.g. all adults living in the US;
- **Sample**: a selected subset of n individuals from a population – e.g. 5,000 US adults appearing in a poll;
- **Summary Statistic**: a summary of the information in a set of observations – e.g. mean, median, mode, etc.;
- **Census**: a counting of all elements of the population;
- **Sampling**: the act of collecting a sample of size n from a population of size N;
  - sample only when we can't perform a census;
  - typically the sample size $n << N$;
- **Sample Statistic**: a summary statistic computed from a sample that estimates the unknown population parameter.

A simple example – marbles in a bag

- Consider a bag full of marbles;
  - the number of marbles in the bag is unknown;
  - there are multiple but unknown colors of marbles in the bag;
  - the fraction of any particular color of marbles in the bag is unknown;
- Questions we could ask:
  - How many marbles are in the bag?
  - How many colors of marbles are in the bag?
  - What is the fraction of blue marbles in the bag?
- Assume we can't just dump the bag out or remove marbles from it permanently – can we devise a process to answer any of these questions?

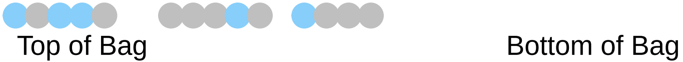
A simple example – marbles in a bag

- Consider a bag full of marbles;
  - the number of marbles in the bag is unknown;
  - there are multiple but unknown colors of marbles in the bag;
  - the fraction of any particular color of marbles in the bag is unknown;
- Questions we could ask:
  - How many marbles are in the bag?
  - How many colors of marbles are in the bag?
  - **What is the fraction of blue marbles in the bag?**
- Assume we can't just dump the bag out or remove marbles from it permanently – can we devise a process to answer any of these questions?

A simple example – marbles in a bag

- Consider a bag full of marbles;
  - the number of marbles in the bag is unknown;
  - there are multiple but unknown colors of marbles in the bag;
  - the fraction of any particular color of marbles in the bag is unknown;
- Questions we could ask:
  - How many marbles are in the bag?
  - How many colors of marbles are in the bag?
  - **What is the fraction of blue marbles in the bag?**
- Assume we can't just dump the bag out or remove marbles from it permanently – can we devise a process to answer any of these questions?
- What about the following:
  1. Stick a hand in the top of the bag and pull out handful of marbles;
  2. Observe them;
  3. Return them to the bag and then mix;
  4. Repeat.

What is the fraction of blue marbles?

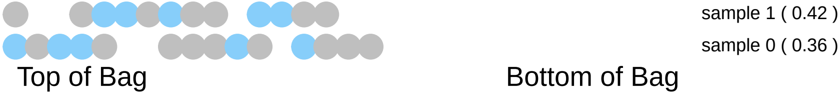


Top of Bag

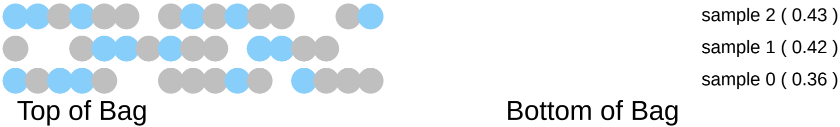
Bottom of Bag

sample 0 (0.36)

What is the fraction of blue marbles?



What is the fraction of blue marbles?



sample 2 (0.43)

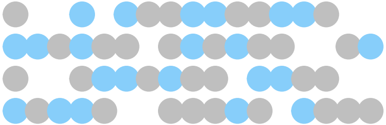
sample 1 (0.42)

sample 0 (0.36)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 3 (0.5)

sample 2 (0.43)

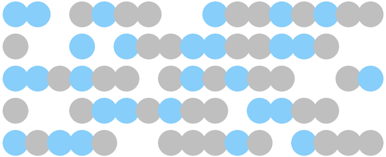
sample 1 (0.42)

sample 0 (0.36)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 4 (0.43)

sample 3 (0.5)

sample 2 (0.43)

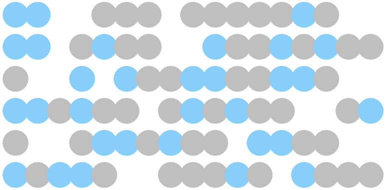
sample 1 (0.42)

sample 0 (0.36)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 5 (0.25)

sample 4 (0.43)

sample 3 (0.5)

sample 2 (0.43)

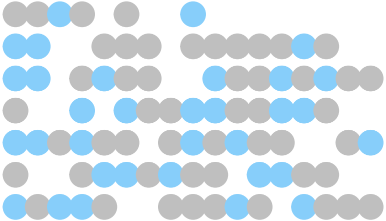
sample 1 (0.42)

sample 0 (0.36)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 6 (0.33)

sample 5 (0.25)

sample 4 (0.43)

sample 3 (0.5)

sample 2 (0.43)

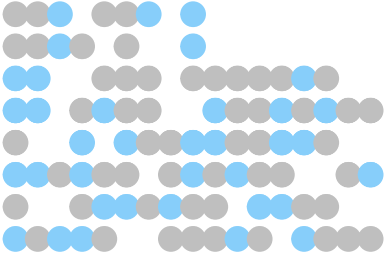
sample 1 (0.42)

sample 0 (0.36)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 7 (0.43)

sample 6 (0.33)

sample 5 (0.25)

sample 4 (0.43)

sample 3 (0.5)

sample 2 (0.43)

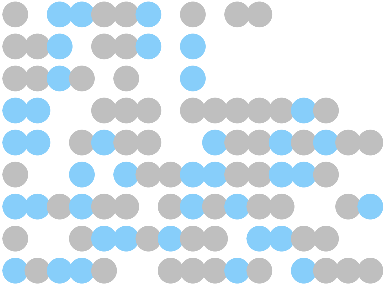
sample 1 (0.42)

sample 0 (0.36)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?

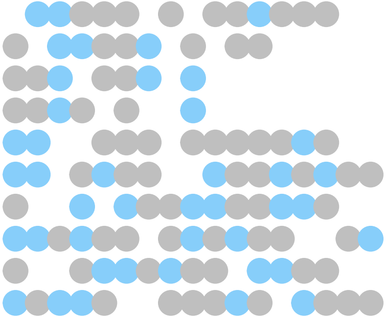


sample 8 (0.33)
sample 7 (0.43)
sample 6 (0.33)
sample 5 (0.25)
sample 4 (0.43)
sample 3 (0.5)
sample 2 (0.43)
sample 1 (0.42)
sample 0 (0.36)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 9 (0.25)

sample 8 (0.33)

sample 7 (0.43)

sample 6 (0.33)

sample 5 (0.25)

sample 4 (0.43)

sample 3 (0.5)

sample 2 (0.43)

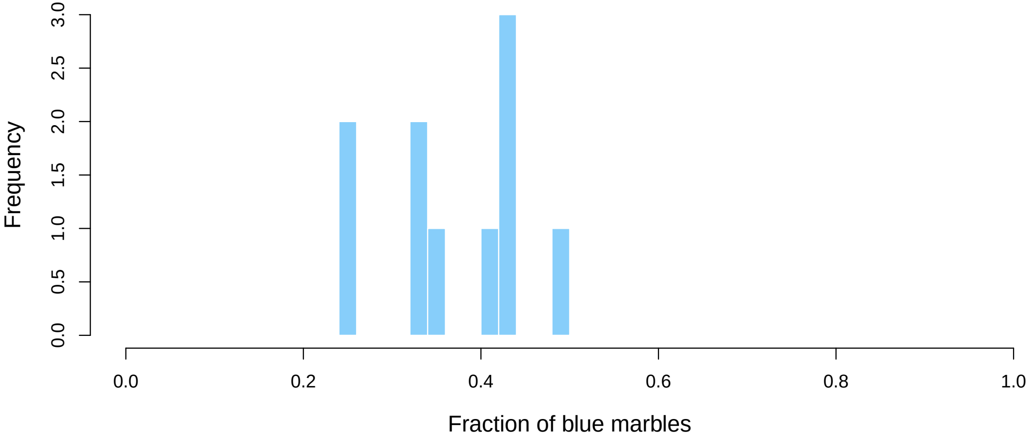
sample 1 (0.42)

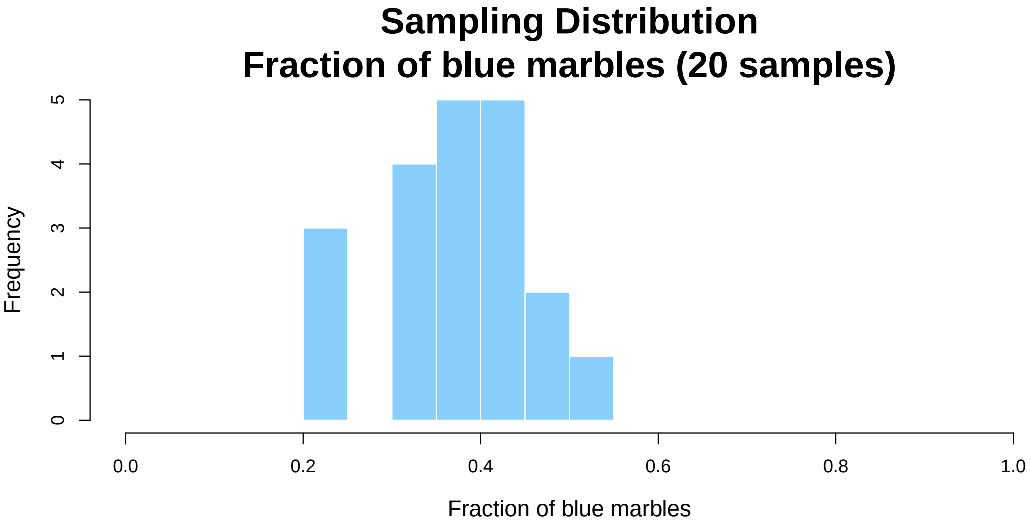
sample 0 (0.36)

Top of Bag

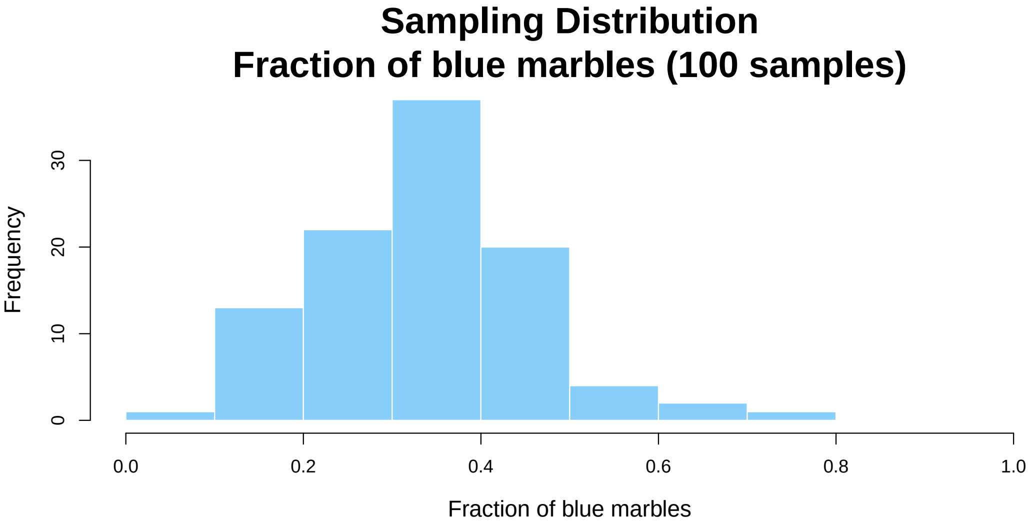
Bottom of Bag

Sampling Distribution
Fraction of blue marbles (10 samples)

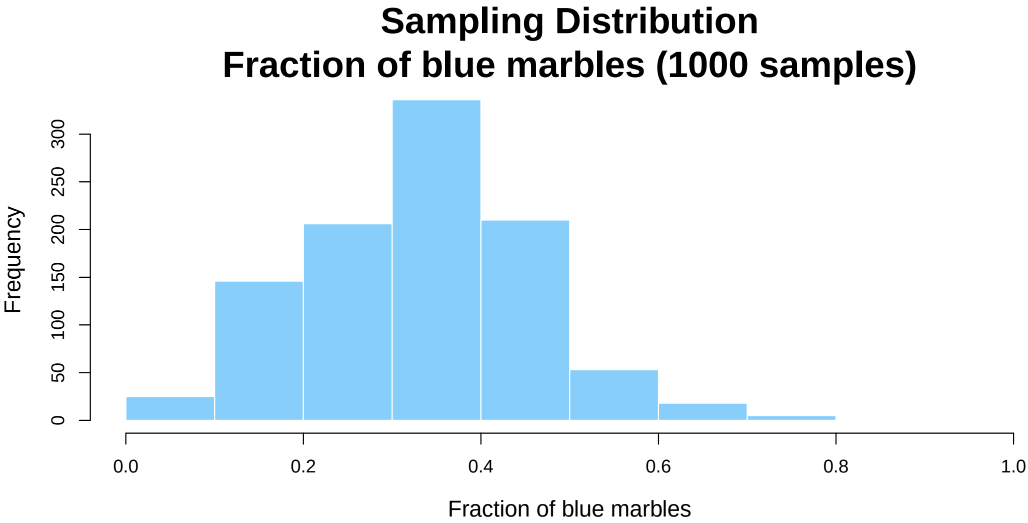




Sampling Distribution
Fraction of blue marbles (20 samples)

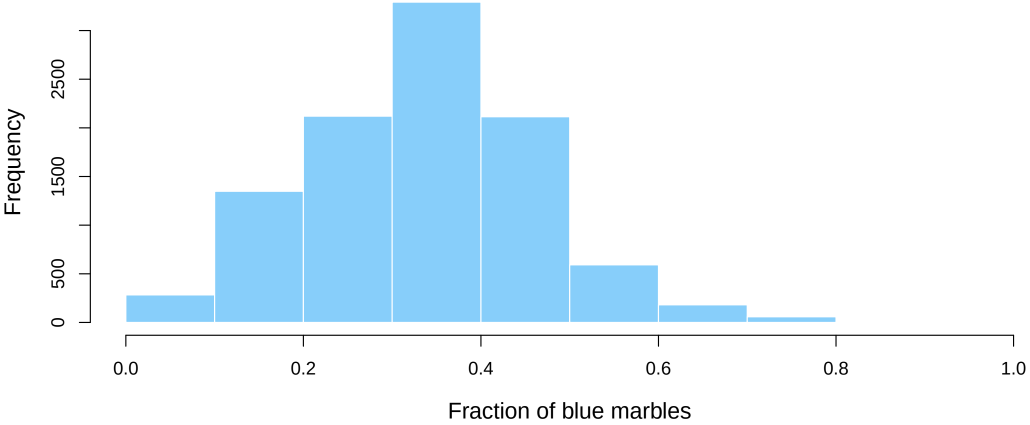


Sampling Distribution
Fraction of blue marbles (100 samples)



Sampling Distribution
Fraction of blue marbles (1000 samples)

Sampling Distribution
Fraction of blue marbles (10000 samples)



Summarizing what we've learned from doing this...

- What should we conclude from our samples about the fraction of blue marbles in the bag? Consider taking the average of the first 10:

$$\frac{0.36 + 0.42 + 0.43 + 0.5 + 0.43 + 0.25 + 0.33 + 0.43 + 0.33 + 0.25}{10} = 0.373;$$

- What is a reasonable measurement of variation around this?

Summarizing what we've learned from doing this...

- What should we conclude from our samples about the fraction of blue marbles in the bag? Consider taking the average of the first 10:

$$\frac{0.36+0.42+0.43+0.5+0.43+0.25+0.33+0.43+0.33+0.25}{10} = 0.373;$$

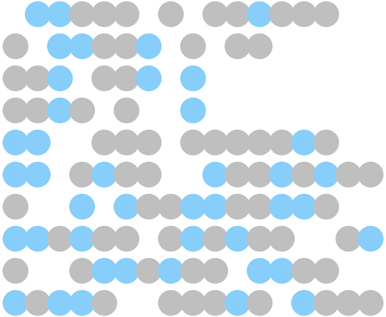
- What is a reasonable measurement of variation around this?
    - Option 1 – look at it empirically:

$$0.25, \underbrace{\mathbf{0.25}, 0.33, 0.33, 0.36, 0.42, 0.43, 0.43, \mathbf{0.43}}_{\text{middle 80\% of data}}, 0.5$$

    - Option 2 – compute the sample standard deviation:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{(0.25-0.373)^2 + \ldots + (0.5-0.373)^2}{10-1}} \approx 0.083.$$

What is the fraction of blue marbles?



sample 9 (0.25)

sample 8 (0.33)

sample 7 (0.43)

sample 6 (0.33)

sample 5 (0.25)

sample 4 (0.43)

sample 3 (0.5)

sample 2 (0.43)

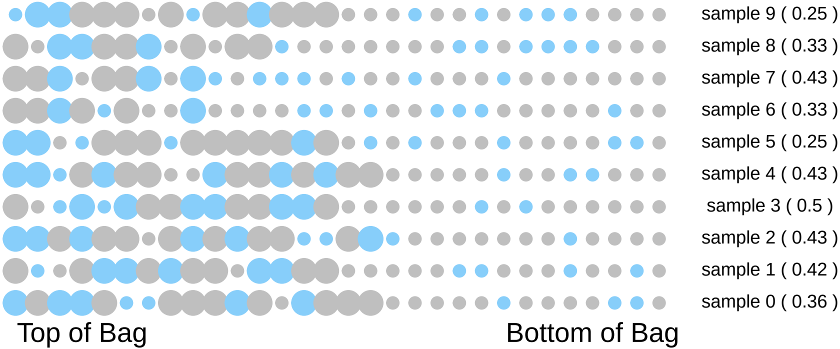
sample 1 (0.42)

sample 0 (0.36)

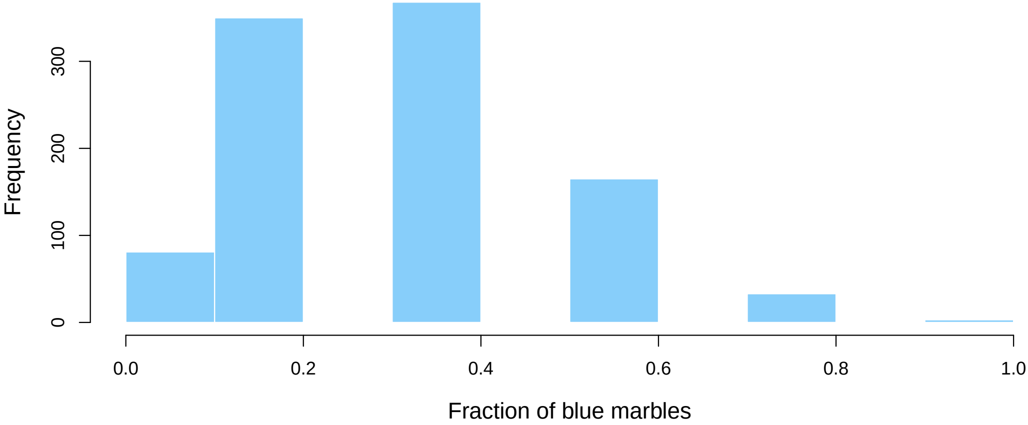
Top of Bag

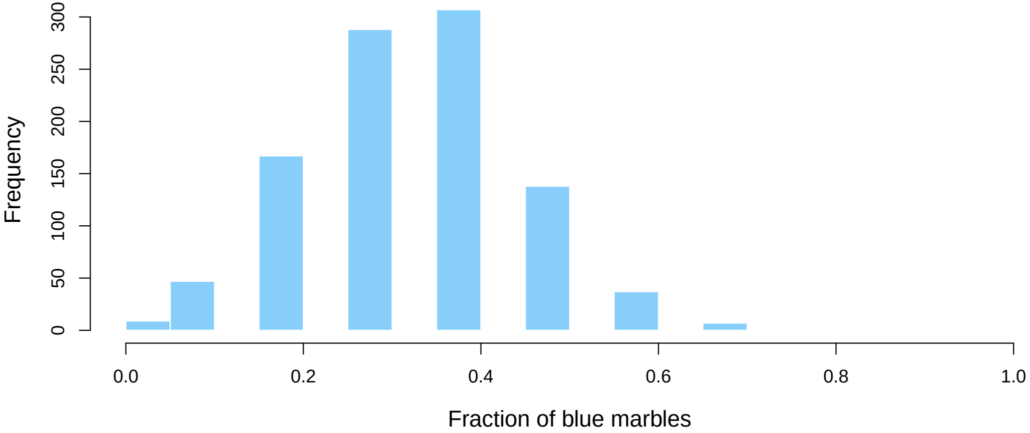
Bottom of Bag

What is the fraction of blue marbles?



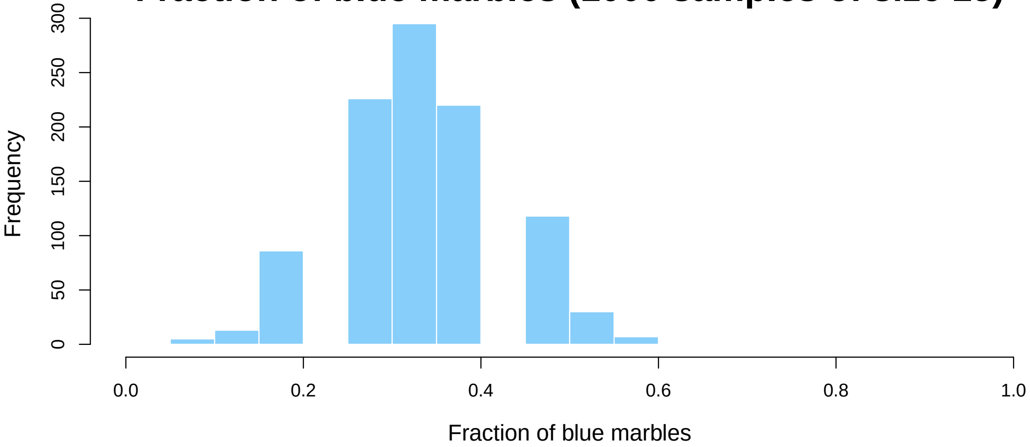
Sampling Distribution
Fraction of blue marbles (1000 samples of size 5)



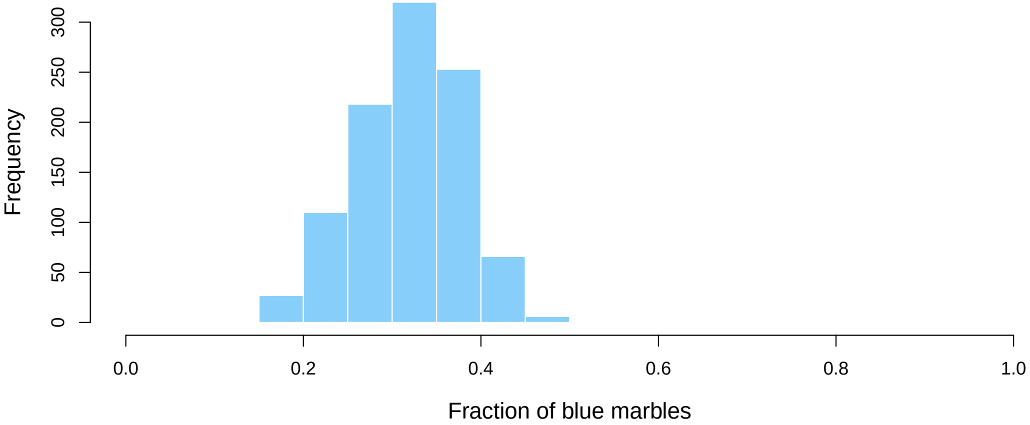


Sampling Distribution
Fraction of blue marbles (1000 samples of size 10)

Sampling Distribution
Fraction of blue marbles (1000 samples of size 15)



Sampling Distribution
Fraction of blue marbles (1000 samples of size 20)



A simple example – marbles in a bag (take 2)

- Consider a bag full of marbles;
  - the number of marbles in the bag is unknown;
  - there are multiple but unknown colors of marbles in the bag;
  - the fraction of any particular color of marbles in the bag is unknown;
- Questions we could ask:
  - How many marbles are in the bag?
  - How many colors of marbles are in the bag?
  - **What is the fraction of blue marbles in the bag?**
- Assume we can't just dump the bag out or remove marbles from it permanently – can we devise a process to answer any of these questions?
- What about the following:
  1. Stick a hand in the top of the bag and pull out handful of marbles;
  2. Observe them;
  3. Return them to the bag and then mix;
  4. Repeat.

A simple example – marbles in a bag (take 2)

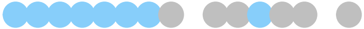
- Consider a bag full of marbles;
  - the number of marbles in the bag is unknown;
  - there are multiple but unknown colors of marbles in the bag;
  - the fraction of any particular color of marbles in the bag is unknown;

- Questions we could ask:
  - How many marbles are in the bag?
  - How many colors of marbles are in the bag?
  - **What is the fraction of blue marbles in the bag?**

- Assume we can't just dump the bag out or remove marbles from it permanently – can we devise a process to answer any of these questions?

- What about the following:
  1. Stick a hand in the top of the bag and pull out handful of marbles;
  2. Observe them;
  3. Return them to the bag ~~and then mix~~;
  4. Repeat.

What is the fraction of blue marbles?

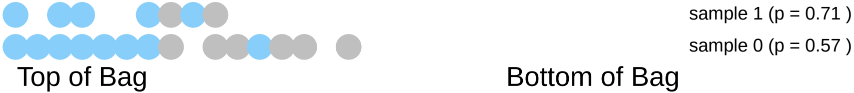


sample 0 (p = 0.57)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



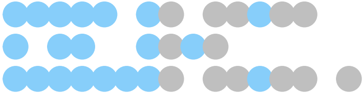
sample 1 (p = 0.71)

sample 0 (p = 0.57)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 2 (p = 0.58)

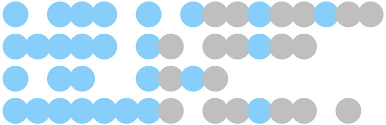
sample 1 (p = 0.71)

sample 0 (p = 0.57)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 3 (p = 0.57)

sample 2 (p = 0.58)

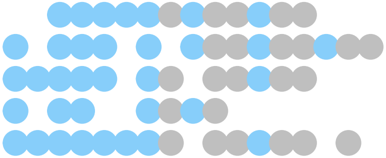
sample 1 (p = 0.71)

sample 0 (p = 0.57)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 4 (p = 0.58)

sample 3 (p = 0.57)

sample 2 (p = 0.58)

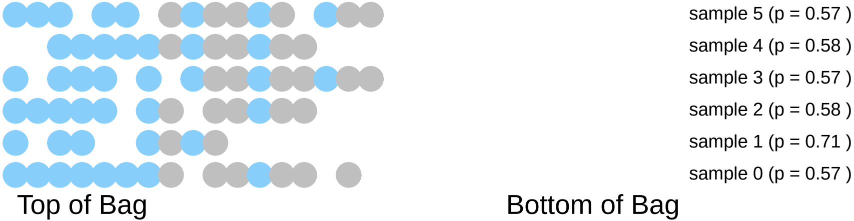
sample 1 (p = 0.71)

sample 0 (p = 0.57)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 5 (p = 0.57)

sample 4 (p = 0.58)

sample 3 (p = 0.57)

sample 2 (p = 0.58)

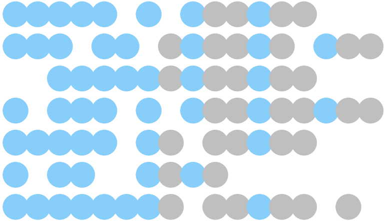
sample 1 (p = 0.71)

sample 0 (p = 0.57)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 6 (p = 0.67)

sample 5 (p = 0.57)

sample 4 (p = 0.58)

sample 3 (p = 0.57)

sample 2 (p = 0.58)

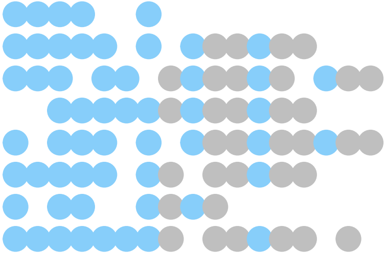
sample 1 (p = 0.71)

sample 0 (p = 0.57)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 7 (p = 1)

sample 6 (p = 0.67)

sample 5 (p = 0.57)

sample 4 (p = 0.58)

sample 3 (p = 0.57)

sample 2 (p = 0.58)

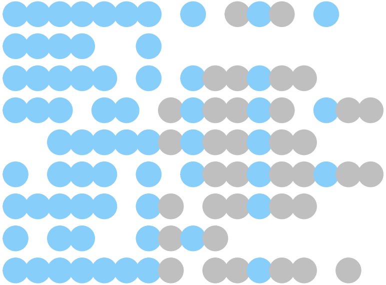
sample 1 (p = 0.71)

sample 0 (p = 0.57)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 8 (p = 0.83)

sample 7 (p = 1)

sample 6 (p = 0.67)

sample 5 (p = 0.57)

sample 4 (p = 0.58)

sample 3 (p = 0.57)

sample 2 (p = 0.58)

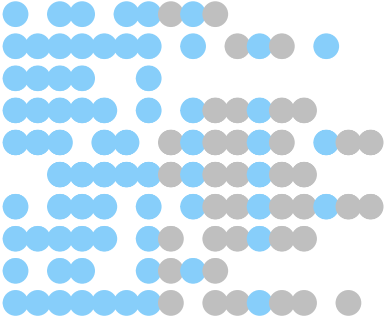
sample 1 (p = 0.71)

sample 0 (p = 0.57)

Top of Bag

Bottom of Bag

What is the fraction of blue marbles?



sample 9 (p = 0.75)

sample 8 (p = 0.83)

sample 7 (p = 1)

sample 6 (p = 0.67)

sample 5 (p = 0.57)

sample 4 (p = 0.58)

sample 3 (p = 0.57)

sample 2 (p = 0.58)

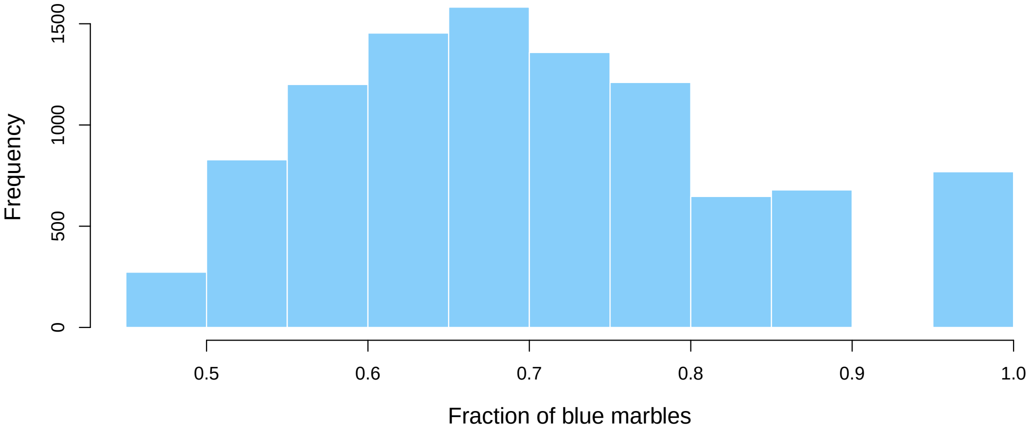
sample 1 (p = 0.71)

sample 0 (p = 0.57)

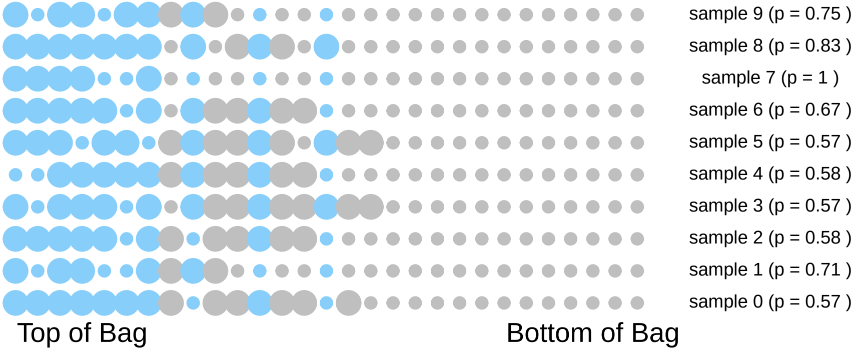
Top of Bag

Bottom of Bag

Histogram: fraction of blue marbles (10000 samples)



What is the fraction of blue marbles?



Sampling methodology

- **Representative sample**: a sample is representative if its characteristics "look like" the population;
- **Generalizable**: a sample is generalizable if we can make "good" guesses about the population using the characteristics of the sample;
- **Bias**: a sample is biased if certain individuals in a population have a higher chance of being included in a sample than others;

Sampling methodology

- **Representative sample**: a sample is representative if its characteristics "look like" the population;

- **Generalizable**: a sample is generalizable if we can make "good" guesses about the population using the characteristics of the sample;

- **Bias**: a sample is biased if certain individuals in a population have a higher chance of being included in a sample than others;

- In general if we create a sample of size *n* randomly then...
    - ...the sample will be unbiased and representative of the population of size *N* so...
    - ...any result based on the sample with generalize to the population and therefore...
    - ...the sample statistic is a good guess for the population parameter which means...
    - ...that we can **INFER** about the population using the sample.

- Our second experiment is an example of a **selection effect** – selection of a biased sample that was not representative and so not generalizable.

An IRL example: polling in 1936...

- *The Literary Digest*:
    - A weekly magazine that started in 1890 w/ circulation $> 1,000,000$;
    - Correctly predicted US presidential elections from 1916 – 1932;
- 1936 Election: Langdon v Roosevelt;
    - *The Literary Digest* polled 10 million and got 2.3 million responses;
    - Langdon predicted to be the decisive winner – but Roosevelt crushed him!
- The magazine folded within 18 months – what happened?!

October 6, 1923
The Literary Digest
A PROBLEM
New York FUNK & WAGNALLS COMPANY London
PUBLIC OPINION New York combined with The LITERARY DIGEST
Vol. 79, No. 1. Whole No. 1746
October 6, 1923
Price 10 Cents

An IRL example: polling in 1936...

- *The Literary Digest*:
  - A weekly magazine that started in 1890 w/ circulation $> 1,000,000$;
  - Correctly predicted US presidential elections from 1916 – 1932;
- 1936 Election: Langdon v Roosevelt;
  - *The Literary Digest* polled 10 million and got 2.3 million responses;
  - Langdon predicted to be the decisive winner – but Roosevelt crushed him!
- The magazine folded within 18 months – what happened?! Sampled:
  - Auto registrations;
  - Phone number lists;
  - Country club memberships;
  - Its own subscriber list.

October 6, 1923
The Literary Digest
A PROBLEM
New York FUNK & WAGNALLS COMPANY London
PUBLIC OPINION New York combined with The LITERARY DIGEST

What happens as we take more samples?

- So why is it that the most frequent value in the histogram appeared to get closer and closer to the actual value of $\frac{1}{3}$ as we added samples?

- **Theorem** (Weak Law of Large Numbers, informal): the sample average “moves towards” the “true” value as the number of samples grows.

What happens as sample size grows?

- Why is it that the sampling distribution got more peaked as we increased the sample size?

- **Theorem** (Central Limit Theorem): As sample size gets bigger the sampling distribution of a sample statistic increasingly follows a normal distribution and the standard deviation of that normal distribution gets smaller.

Data Generating Process

- A useful (and ubiquitous) construct: **the data generating process** (DGP) – the set of all operations that lead to:
  1. the particular observations that appear in the dataset...
  2. ...and their structure;

- Occurs both IRL and at the researcher's desk – usually we know at most only part of the DGP;

- Selected examples:
  1. Generation of events (could become data) – e.g. some countries fight wars;
  2. Selection of population units into the data;
  3. Categorization/Binarization – e.g. a Likert scale representation of preference;
  4. Analyst decisions to aggregate, group, or drop data;
  5. Assignment of independent variables to observations (e.g. selection of treatment and control groups);

Data Generating Process: Sampling

- A useful (and ubiquitous) construct: **the data generating process** (DGP) – the set of all operations that lead to:
  1. the particular observations that appear in the dataset...
     ...and their structure;

  Occurs both IRL and at the researcher's desk – usually we know at most only part of the DGP;

- Selected examples:
  1. **Generation of events (could become data) – e.g. some countries fight wars**;
  2. **Selection of selection of population units into the data**;
     Categorization/Binarization – e.g. a Likert scale representation of preference;
     Analyst decisions to aggregate, group, or drop data;
     Assignment of independent variables to observations (e.g. selection of treatment and control groups);

Why should we care?

Concepts of sampling will be the basis of things like probability and hypothesis testing which we will cover next! Understanding sampling methodology will help you decide whether to trust polling results etc.