Introduction to Data Science

Data: What, Why & How

Ernst C. Wit

October 6, 2024

Lecture 4:

Simple Random Sampling

Simple Random Sampling (SRS)

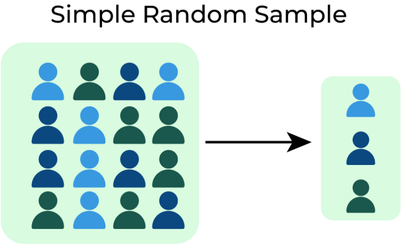
Simple Random Sampling

- Each member of population has equal chance of selection.
- Samples do not “interfere” with each other

Sampling can be done **with** or **without replacement**.

Focus today:

- SRS with replacement and its probabilistic properties.
- Extension to non-equal chance sampling.



Example: Population of Voters

Population: 100 voters

- 40 support candidate T
- 60 support candidate H

Parameter of interest:

Example: Population of Voters

Population: 100 voters

- 40 support candidate T
- 60 support candidate H

Parameter of interest:

$$p = \text{true fraction of T voters} =$$

Example: Population of Voters

Population: 100 voters

- 40 support candidate T
- 60 support candidate H

Parameter of interest:

$$p = \text{true fraction of T voters} = 0.4.$$

Methods:

- We take a sample of 10 voters,
- with replacement (so same voter may appear twice),
- and analyze probability of different T voters.

Probability when sampling one or two Voters

Probabilities when **sampling a single voter**:

- **P(selecting T)** $= P(T) = \frac{40}{100} = 0.4$
- **P(selecting H)** $= P(H) = \frac{60}{100} = 0.6$

Nice: sampling **with replacement**, so probabilities constant.

Probability when sampling one or two Voters

Probabilities when **sampling a single voter**:

- **P(selecting T)** $= P(T) = \frac{40}{100} = 0.4$
- **P(selecting H)** $= P(H) = \frac{60}{100} = 0.6$

Nice: sampling **with replacement**, so probabilities constant.

Probabilities when **sampling two voters**:

- $P(TT) =$

Probability when sampling one or two Voters

Probabilities when **sampling a single voter**:

- **P(selecting T)** $= P(T) = \frac{40}{100} = 0.4$
- **P(selecting H)** $= P(H) = \frac{60}{100} = 0.6$

Nice: sampling **with replacement**, so probabilities constant.

Probabilities when **sampling two voters**:

- $P(TT) = 0.4 \times 0.4 =$

Probability when sampling one or two Voters

Probabilities when **sampling a single voter**:

- **P(selecting T)** $= P(T) = \frac{40}{100} = 0.4$
- **P(selecting H)** $= P(H) = \frac{60}{100} = 0.6$

Nice: sampling **with replacement**, so probabilities constant.

Probabilities when **sampling two voters**:

- $P(TT) = 0.4 \times 0.4 = 0.16$!making use of **independence**!
- $P(TH) =$

Probability when sampling one or two Voters

Probabilities when **sampling a single voter**:

- $\mathbf{P(selecting\ T)} = P(T) = \frac{40}{100} = 0.4$
- $\mathbf{P(selecting\ H)} = P(H) = \frac{60}{100} = 0.6$

Nice: sampling **with replacement**, so probabilities constant.

Probabilities when **sampling two voters**:

- $P(TT) = 0.4 \times 0.4 = 0.16$!making use of **independence**!
- $P(TH) = 0.4 \times 0.6 =$

Probability when sampling one or two Voters

Probabilities when **sampling a single voter**:

- **P(selecting T)** $= P(T) = \frac{40}{100} = 0.4$
- **P(selecting H)** $= P(H) = \frac{60}{100} = 0.6$

Nice: sampling **with replacement**, so probabilities constant.

Probabilities when **sampling two voters**:

- $P(TT) = 0.4 \times 0.4 = 0.16$!making use of **independence**!
- $P(TH) = 0.4 \times 0.6 = 0.24$
- $P(HT) =$

Probability when sampling one or two Voters

Probabilities when **sampling a single voter**:

- **P(selecting T)** $= P(T) = \frac{40}{100} = 0.4$
- **P(selecting H)** $= P(H) = \frac{60}{100} = 0.6$

Nice: sampling **with replacement**, so probabilities constant.

Probabilities when **sampling two voters**:

- $P(TT) = 0.4 \times 0.4 = 0.16$!making use of **independence**!
- $P(TH) = 0.4 \times 0.6 = 0.24$
- $P(HT) = 0.6 \times 0.4 =$

Probability when sampling one or two Voters

Probabilities when **sampling a single voter**:

- **P(selecting T)** $= P(T) = \frac{40}{100} = 0.4$
- **P(selecting H)** $= P(H) = \frac{60}{100} = 0.6$

Nice: sampling **with replacement**, so probabilities constant.

Probabilities when **sampling two voters**:

- $P(TT) = 0.4 \times 0.4 = 0.16$!making use of **independence**!
- $P(TH) = 0.4 \times 0.6 = 0.24$
- $P(HT) = 0.6 \times 0.4 = 0.24$
- $P(HH) =$

Probability when sampling one or two Voters

Probabilities when **sampling a single voter**:

- **P(selecting T)** $= P(T) = \frac{40}{100} = 0.4$
- **P(selecting H)** $= P(H) = \frac{60}{100} = 0.6$

Nice: sampling **with replacement**, so probabilities constant.

Probabilities when **sampling two voters**:

- $P(TT) = 0.4 \times 0.4 = 0.16$!making use of **independence**!
- $P(TH) = 0.4 \times 0.6 = 0.24$
- $P(HT) = 0.6 \times 0.4 = 0.24$
- $P(HH) = 0.6 \times 0.6 =$

Probability when sampling one or two Voters

Probabilities when **sampling a single voter**:

- **P(selecting T)** $= P(T) = \frac{40}{100} = 0.4$
- **P(selecting H)** $= P(H) = \frac{60}{100} = 0.6$

Nice: sampling **with replacement**, so probabilities constant.

Probabilities when **sampling two voters**:

- $P(TT) = 0.4 \times 0.4 = 0.16$!making use of **independence**!
- $P(TH) = 0.4 \times 0.6 = 0.24$
- $P(HT) = 0.6 \times 0.4 = 0.24$
- $P(HH) = 0.6 \times 0.6 = 0.36$

Independence

If two random events E_1, E_2 do not interfere, then

$$P(E_1 \text{ and } E_2) = P(E_1) \times P(E_2).$$

Random variables

Notation "HTHHT..." becomes awkward when sample large, so:

Random Variable

Let S be a random sample, then any function

$$X : S \longrightarrow \mathbb{R}$$

is called a random variable.

Random variables

Notation "HTHHT..." becomes awkward when sample large, so:

Random Variable

Let S be a random sample, then any function

$$X : S \longrightarrow \mathbb{R}$$

is called a random variable.

Example: Random sample of 10 voters *with replacement*, then

$$X = \text{number of T voters among random sample of 10 voters}$$

Clearly, X is a random variable.

Random variables

Notation “HTHHT...” becomes awkward when sample large, so:

Random Variable

Let S be a random sample, then any function

$$X : S \longrightarrow \mathbb{R}$$

is called a random variable.

Example: Random sample of 10 voters *with replacement*, then

X = number of T voters among random sample of 10 voters

Clearly, X is a random variable.

Y = number of Swiss nationals among sample of 10 voters

Random variables

Notation "HTHHT..." becomes awkward when sample large, so:

Random Variable

Let S be a random sample, then any function

$$X : S \longrightarrow \mathbb{R}$$

is called a random variable.

Example: Random sample of 10 voters *with replacement*, then

$X =$ number of T voters among random sample of 10 voters

Clearly, X is a random variable.

$Y =$ number of Swiss nationals among sample of 10 voters

is also a random variable.

Random variables

Notation "HTHHT..." becomes awkward when sample large, so:

Random Variable

Let S be a random sample, then any function

$$X : S \longrightarrow \mathbb{R}$$

is called a random variable.

Example: Random sample of 10 voters *with replacement*, then

$$X = \text{number of T voters among random sample of 10 voters}$$

Clearly, X is a random variable.

$$Y = \text{number of Swiss nationals among sample of 10 voters}$$

is also a random variable.

$$Z = \text{number of blond haired people among 100 voters}$$

Random variables

Notation "HTHHT..." becomes awkward when sample large, so:

Random Variable

Let S be a random sample, then any function

$$X : S \longrightarrow \mathbb{R}$$

is called a random variable.

Example: Random sample of 10 voters *with replacement*, then

$X =$ number of T voters among random sample of 10 voters

Clearly, X is a random variable.

$Y =$ number of Swiss nationals among sample of 10 voters

is also a random variable.

$Z =$ number of blond haired people among 100 voters

... is **not** a random variable (at least not on S)!

Features of a random variable

Features of a (discrete) random variable

- **probability mass function:**

$$p_X(k) = P(X = k)$$

- **support/domain of X:** all possible values of X

$$D_X = \{k \mid p_X(k) > 0\}$$

- **total probability:**

$$\sum_{k \in D_X} p_X(k) = 1$$

- **mean/expectation:**

$$EX = \sum_{k \in D_X} k \times p_X(k)$$

Sampling 10 Voters (With Replacement)

Let

$$X = \text{number of T voters among random sample of 10 voters}$$

What is its *support*?

Sampling 10 Voters (With Replacement)

Let

$$X = \text{number of T voters among random sample of 10 voters}$$

What is its *support*?

$$D_X = \{0, 1, 2, 3, \ldots, 9, 10\}$$

Sampling 10 Voters (With Replacement)

Let

$$X = \text{number of T voters among random sample of 10 voters}$$

What is its *support*?

$$D_X = \{0, 1, 2, 3, \ldots, 9, 10\}$$

What is its *probability mass function*?

$$p_X(0) \quad =$$

Sampling 10 Voters (With Replacement)

Let

$$X = \text{number of T voters among random sample of 10 voters}$$

What is its *support*?

$$D_X = \{0, 1, 2, 3, \ldots, 9, 10\}$$

What is its *probability mass function*?

$$p_X(0) \quad = \quad P(HH \ldots H) =$$

Sampling 10 Voters (With Replacement)

Let

$$X = \text{number of T voters among random sample of 10 voters}$$

What is its *support*?

$$D_X = \{0, 1, 2, 3, \dots, 9, 10\}$$

What is its *probability mass function*?

$$\begin{aligned} p_X(0) &= P(HH\dots H) = 0.6 \times 0.6 \times \dots \times 0.6 \\ &= \end{aligned}$$

Sampling 10 Voters (With Replacement)

Let

$$X = \text{number of T voters among random sample of 10 voters}$$

What is its *support*?

$$D_X = \{0, 1, 2, 3, \ldots, 9, 10\}$$

What is its *probability mass function*?

$$\begin{aligned} p_X(0) &= P(HH\ldots H) = 0.6 \times 0.6 \times \ldots \times 0.6 \\ &= 0.6^{10} \\ p_X(1) &= \end{aligned}$$

Sampling 10 Voters (With Replacement)

Let

$$X = \text{number of T voters among random sample of 10 voters}$$

What is its *support*?

$$D_X = \{0, 1, 2, 3, \ldots, 9, 10\}$$

What is its *probability mass function*?

$$\begin{aligned} p_X(0) &= P(HH\ldots H) = 0.6 \times 0.6 \times \ldots \times 0.6 \\ &= 0.6^{10} \\ p_X(1) &= P(TH\ldots H) + \end{aligned}$$

Sampling 10 Voters (With Replacement)

Let

$$X = \text{number of T voters among random sample of 10 voters}$$

What is its *support*?

$$D_X = \{0, 1, 2, 3, \ldots, 9, 10\}$$

What is its *probability mass function*?

$$\begin{aligned} p_X(0) &= P(HH\ldots H) = 0.6 \times 0.6 \times \ldots \times 0.6 \\ &= 0.6^{10} \\ p_X(1) &= P(TH\ldots H) + P(HT\ldots H) + \end{aligned}$$

Sampling 10 Voters (With Replacement)

Let

$$X = \text{number of T voters among random sample of 10 voters}$$

What is its *support*?

$$D_X = \{0, 1, 2, 3, \ldots, 9, 10\}$$

What is its *probability mass function*?

$$\begin{aligned} p_X(0) &= P(HH\ldots H) = 0.6 \times 0.6 \times \ldots \times 0.6 \\ &= 0.6^{10} \\ p_X(1) &= P(TH\ldots H) + P(HT\ldots H) + \ldots P(HH\ldots T) \\ &= \end{aligned}$$

Sampling 10 Voters (With Replacement)

Let

$$X = \text{number of T voters among random sample of 10 voters}$$

What is its *support*?

$$D_X = \{0, 1, 2, 3, \ldots, 9, 10\}$$

What is its *probability mass function*?

$$\begin{aligned}
p_X(0) &= P(HH\ldots H) = 0.6 \times 0.6 \times \ldots \times 0.6 \\
&= 0.6^{10} \\
p_X(1) &= P(TH\ldots H) + P(HT\ldots H) + \ldots P(HH\ldots T) \\
&= 0.4 \times 0.6 \times \ldots \times 0.6 \times 10 \\
&=
\end{aligned}$$

Sampling 10 Voters (With Replacement)

Let

$$X = \text{number of T voters among random sample of 10 voters}$$

What is its *support*?

$$D_X = \{0, 1, 2, 3, \ldots, 9, 10\}$$

What is its *probability mass function*?

$$\begin{aligned}
p_X(0) &= P(HH \ldots H) = 0.6 \times 0.6 \times \ldots \times 0.6 \\
&= 0.6^{10} \\
p_X(1) &= P(TH \ldots H) + P(HT \ldots H) + \ldots P(HH \ldots T) \\
&= 0.4 \times 0.6 \times \ldots \times 0.6 \times 10 \\
&= 10 \times 0.4 \times 0.6^9
\end{aligned}$$

Note:

- Probabilities are easy to generalize (we use independence)
- **Main difficulty:** multiplication factor (i.e. 10 for $p_X(1)$)

Intermezzo: Permutations

Permutations

Given n distinct items, a **permutation** is an arrangement of all items in a specific order.

Example. Arrange 3 people (A, B, C) in different ways:

$$ABC, ACB, BAC, BCA, CAB, CBA.$$

The number of permutations is:

Intermezzo: Permutations

Permutations

Given n distinct items, a **permutation** is an arrangement of all items in a specific order.

Example. Arrange 3 people (A, B, C) in different ways:

$$ABC, ACB, BAC, BCA, CAB, CBA.$$

The number of permutations is:

$$3 \cdot 2 \cdot 1 =$$

Intermezzo: Permutations

Permutations

Given n distinct items, a **permutation** is an arrangement of all items in a specific order.

Example. Arrange 3 people (A, B, C) in different ways:

$$ABC, ACB, BAC, BCA, CAB, CBA.$$

The number of permutations is:

$$3 \cdot 2 \cdot 1 = 3!$$

In general:

- Total number of permutations of n items is given by:

$$n! = n \cdot (n-1) \cdot \ldots \cdot 1$$

Intermezzo: Combinations

Combinations:

Given n items of two types:

- r indistinguishable items of type 1
- $n - r$ indistinguishable items of type 2

A **combination** is an ordering of those n items.

Example. Combinations for four items (A, A, B, B),

Intermezzo: Combinations

Combinations:

Given n items of two types:

- r indistinguishable items of type 1
- $n - r$ indistinguishable items of type 2

A **combination** is an ordering of those n items.

Example. Combinations for four items (A, A, B, B),

$$AABB, ABAB, ABBA, BAAB, BABA, BBAA$$

Number of combinations are

$$4!$$

Intermezzo: Combinations

Combinations:

Given n items of two types:

- r indistinguishable items of type 1
- $n - r$ indistinguishable items of type 2

A **combination** is an ordering of those n items.

Example. Combinations for four items (A, A, B, B),

$$AABB, ABAB, ABBA, BAAB, BABA, BBAA$$

Number of combinations are

$$\frac{4!}{2 \times 2} =$$

Intermezzo: Combinations

Combinations:

Given n items of two types:

- r indistinguishable items of type 1
- $n - r$ indistinguishable items of type 2

A **combination** is an ordering of those n items.

Example. Combinations for four items (A, A, B, B),

$$AABB, ABAB, ABBA, BAAB, BABA, BBAA$$

Number of combinations are

$$\frac{4!}{2 \times 2} = \binom{4}{2} = 6$$

In general:

- Total number of combinations is given by: $\binom{n}{r} = \frac{n!}{r! \cdot (n-r)!}$

Back to our 10 voters: probability mass function

We had found $p_X(k)$ for $k = 0, 1$. Let's continue:

$$p_X(2) =$$

Back to our 10 voters: probability mass function

We had found $p_X(k)$ for $k = 0, 1$. Let's continue:

$$\begin{aligned} p_X(2) &= P(\text{combinations of 2 Ts and 8 Hs}) \\ &= \end{aligned}$$

Back to our 10 voters: probability mass function

We had found $p_X(k)$ for $k = 0, 1$. Let's continue:

$$
\begin{aligned}
p_X(2) &= P(\text{combinations of 2 Ts and 8 Hs}) \\
&= 0.4 \times 0.4 \times 0.6 \times \ldots \times 0.6 \times \binom{10}{2} \\
&=
\end{aligned}
$$

Back to our 10 voters: probability mass function

We had found $p_X(k)$ for $k = 0, 1$. Let's continue:

$$\begin{aligned} p_X(2) &= P(\text{combinations of 2 Ts and 8 Hs}) \\ &= 0.4 \times 0.4 \times 0.6 \times \ldots \times 0.6 \times \binom{10}{2} \\ &= \binom{10}{2} 0.4^2 0.6^8 \\ p_X(3) &= \end{aligned}$$

Back to our 10 voters: probability mass function

We had found $p_X(k)$ for $k = 0, 1$. Let's continue:

$$
\begin{aligned}
p_X(2) &= P(\text{combinations of 2 Ts and 8 Hs}) \\
&= 0.4 \times 0.4 \times 0.6 \times \ldots \times 0.6 \times \binom{10}{2} \\
&= \binom{10}{2} 0.4^2 0.6^8 \\
p_X(3) &= P(\text{combinations of 3 Ts and 7 Hs}) \\
&=
\end{aligned}
$$

Back to our 10 voters: probability mass function

We had found $p_X(k)$ for $k = 0, 1$. Let's continue:

$$
\begin{aligned}
p_X(2) &= P(\text{combinations of 2 Ts and 8 Hs}) \\
&= 0.4 \times 0.4 \times 0.6 \times \ldots \times 0.6 \times \binom{10}{2} \\
&= \binom{10}{2} 0.4^2 0.6^8 \\
p_X(3) &= P(\text{combinations of 3 Ts and 7 Hs}) \\
&= 0.4 \times 0.4 \times 0.4 \times 0.6 \times \ldots \times 0.6 \times \binom{10}{3} \\
&=
\end{aligned}
$$

Back to our 10 voters: probability mass function

We had found $p_X(k)$ for $k = 0, 1$. Let's continue:

$$\begin{aligned}
p_X(2) &= P(\text{combinations of 2 Ts and 8 Hs}) \\
&= 0.4 \times 0.4 \times 0.6 \times \ldots \times 0.6 \times \binom{10}{2} \\
&= \binom{10}{2} 0.4^2 0.6^8 \\
p_X(3) &= P(\text{combinations of 3 Ts and 7 Hs}) \\
&= 0.4 \times 0.4 \times 0.4 \times 0.6 \times \ldots \times 0.6 \times \binom{10}{3} \\
&= \binom{10}{3} 0.4^3 0.6^7
\end{aligned}$$

The Binomial Distribution

$$X = \text{number of T voters among sample of 10}$$

This is a **binomial experiment**, where:

- Number of trials $n = 10$
- Success probability $p = 0.4$ (sampling T voter)

Binomial Distribution Bin(n, p)

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Where:

- $n = 10$ (sample size)
- k is the number of successes (number of "T"s)
- $p = 0.4$ (probability of sampling T voter)

Example: 5 "T"s in sample of 10 voters

Probability of finding exactly 5 T voters:

$$P(X = 5) = \binom{10}{5}(0.4)^5(0.6)^5$$

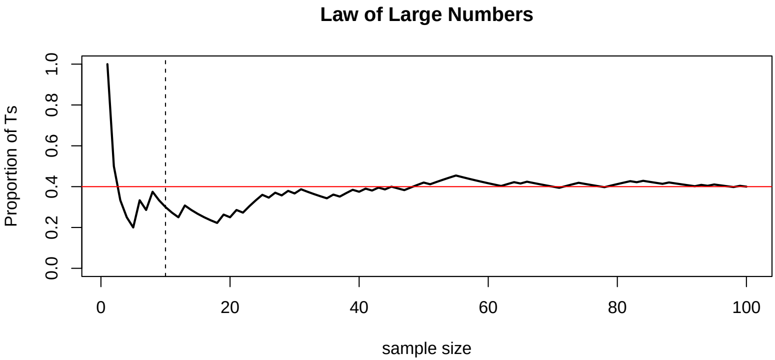
- $\binom{10}{5} = 252$
- $(0.4)^5 = 0.01024$
- $(0.6)^5 = 0.07776$
- $P(X = 5) = 252 \times 0.01024 \times 0.07776 = 0.2006$

Result: The probability of getting exactly 5 "T"s is 20.06%.

Law of Large Numbers: connection parameter & sample

As sample size increases:

- Sample proportion of Ts converges to population proportion.



This is the **Law of Large Numbers**.

Binomial as a sum

Let

$$X_i = \begin{cases} 1 & \text{voter } i \text{ in sample votes for T} \\ 0 & \text{voter } i \text{ in sample votes for H} \end{cases}$$

Then

$$\begin{aligned} X &= \text{number of people in sample voting for T} \\ &= \sum_{i=1}^{10} X_i \end{aligned}$$

and sample proportion $\hat{p}$:

$$\hat{p} = \frac{1}{10} \sum_{i=1}^{10} X_i$$

Expected Value of a Binomial Random Variable

Expected Value:

Expected value or **mean** of X is defined as:

$$E[X] = \sum_{k \in D_X} k \cdot p_X(k)$$

Expections are linear:

$$E[aX + bY] = aEX + bEY$$

Example. For each X_i we have:

$$E[X_i] = 1 \times p + 0 \times (1 - p) = p$$

Therefore, expected value of binomial random variable X is:

$$E[X] = \sum_{i=1}^{n} p = n \cdot p$$

Law of Large Numbers (precisely)

Law of Large Numbers

Let X_i have expected value $EX_i = \mu$ (with bounded variance), then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu$$

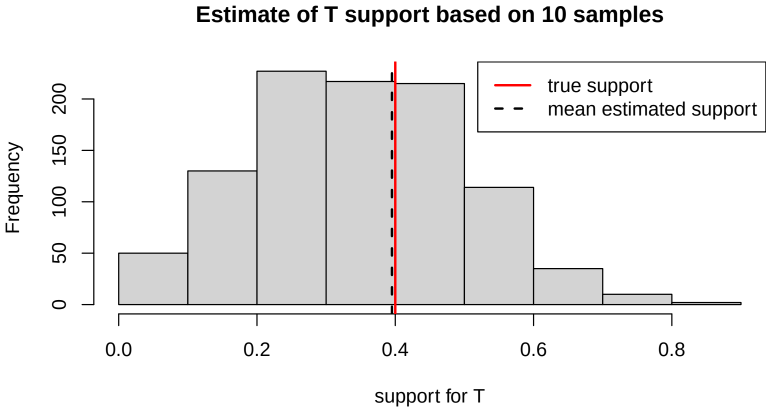
Example. In particular for Binomial in voter example,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} 0.4.$$

Key take-away: $\frac{X}{10}$ should be a good approximation for 0.4.

Estimating T support based on SRS of 10 voters

We repeat 1000 times taking a sample of size 10:



NOTE: Estimates are quite variable (next lecture!)

Unequal Sampling Probabilities

In real-world scenarios, sampling probabilities are often **not equal**:

Some groups may be **over- or under-represented** due to

- practical constraints
- sampling bias
- non-response bias

This results in biased estimates unless adjustments are made.

Example: In election example, assume that

- T voters are **twice as likely** to be sampled as H voters.
- As before 10 voters are sampled with replacement and

$$X = \text{number of T voters in sample of 10}$$

Probability that single draw is T

Let's consider a single draw from population:

- This is one of $\{T_1, \ldots, T_{40}, H_1, \ldots, H_{60}\}$.

Probability that single draw is T

Let's consider a single draw from population:

- This is one of $\{T_1, \ldots, T_{40}, H_1, \ldots, H_{60}\}$.
- so $\sum_{i=1}^{40} P(T_i) + \sum_{j=1}^{60} P(H_j) = 1$.

Probability that single draw is T

Let's consider a single draw from population:

- This is one of $\{T_1, \ldots, T_{40}, H_1, \ldots, H_{60}\}$.
- so $\sum_{i=1}^{40} P(T_i) + \sum_{j=1}^{60} P(H_j) = 1$.
- we also know: Ts are twice as likely to be sampled

Probability that single draw is T

Let's consider a single draw from population:

- This is one of $\{T_1, \ldots, T_{40}, H_1, \ldots, H_{60}\}$.
- so $\sum_{i=1}^{40} P(T_i) + \sum_{j=1}^{60} P(H_j) = 1$.
- we also know: Ts are twice as likely to be sampled

$$P(T_1) = 2P(H_1).$$

so, $\sum_{i=1}^{40} 2P(H_1) + \sum_{j=1}^{60} P(H_1) = 1$ so,

Probability that single draw is T

Let's consider a single draw from population:

- This is one of $\{T_1, \ldots, T_{40}, H_1, \ldots, H_{60}\}$.
- so $\sum_{i=1}^{40} P(T_i) + \sum_{j=1}^{60} P(H_j) = 1$.
- we also know: Ts are twice as likely to be sampled

$$P(T_1) = 2P(H_1).$$

so, $\sum_{i=1}^{40} 2P(H_1) + \sum_{j=1}^{60} P(H_1) = 1$ so,

$$140P(H_1) = 1 \qquad \Rightarrow \qquad P(H_1) = 1/140$$

and, therefore, probability that single draw is T

$$P(T) =$$

Probability that single draw is T

Let's consider a single draw from population:

- This is one of $\{T_1, \ldots, T_{40}, H_1, \ldots, H_{60}\}$.
- so $\sum_{i=1}^{40} P(T_i) + \sum_{j=1}^{60} P(H_j) = 1$.
- we also know: Ts are twice as likely to be sampled

$$P(T_1) = 2P(H_1).$$

so, $\sum_{i=1}^{40} 2P(H_1) + \sum_{j=1}^{60} P(H_1) = 1$ so,

$$140P(H_1) = 1 \qquad \Rightarrow \qquad P(H_1) = 1/140$$

and, therefore, probability that single draw is T

$$P(T) = 2 \times 40 \times \frac{1}{140} = \frac{4}{7}$$

Probability that single draw is T

Let's consider a single draw from population:

- This is one of $\{T_1, \ldots, T_{40}, H_1, \ldots, H_{60}\}$.
- so $\sum_{i=1}^{40} P(T_i) + \sum_{j=1}^{60} P(H_j) = 1$.
- we also know: Ts are twice as likely to be sampled

$$P(T_1) = 2P(H_1).$$

so, $\sum_{i=1}^{40} 2P(H_1) + \sum_{j=1}^{60} P(H_1) = 1$ so,

$$140P(H_1) = 1 \qquad \Rightarrow \qquad P(H_1) = 1/140$$

and, therefore, probability that single draw is T

$$P(T) = 2 \times 40 \times \frac{1}{140} = \frac{4}{7}$$

PS. Just note that that

$$P(T) = \frac{2p}{1+p}$$

where $p = 0.4$ is true fraction of T supporters.

What is distribution of new X?

Let

$$X = \text{number of T voters in sample of 10}$$

As before: Binomial distribution but with different probability:

$$X \sim \text{Bin}(10, \frac{4}{7})$$

What is distribution of new X?

Let

$$X = \text{number of T voters in sample of 10}$$

As before: Binomial distribution but with different probability:

$$X \sim \text{Bin}(10, \frac{4}{7})$$

With Law of Large Numbers,

$$\frac{X}{10} \approx$$

What is distribution of new X?

Let

$$X = \text{number of T voters in sample of 10}$$

As before: Binomial distribution but with different probability:

$$X \sim \text{Bin}(10, \frac{4}{7})$$

With Law of Large Numbers,

$$\frac{X}{10} \approx \frac{4}{7} =$$

What is distribution of new X?

Let

$$X = \text{number of T voters in sample of 10}$$

As before: Binomial distribution but with different probability:

$$X \sim \text{Bin}(10, \frac{4}{7})$$

With Law of Large Numbers,

$$\frac{X}{10} \approx \frac{4}{7} = \frac{2p}{1+p}$$

Let's use this to estimate support for T, i.e., p:

What is distribution of new X?

Let

$$X = \text{number of T voters in sample of 10}$$

As before: Binomial distribution but with different probability:

$$X \sim \text{Bin}(10, \frac{4}{7})$$

With Law of Large Numbers,

$$\frac{X}{10} \approx \frac{4}{7} = \frac{2p}{1+p}$$

Let's use this to estimate support for T, i.e., p:

$$(1+p)X \approx 20p$$

What is distribution of new X?

Let

$$X = \text{number of T voters in sample of 10}$$

As before: Binomial distribution but with different probability:

$$X \sim \text{Bin}(10, \frac{4}{7})$$

With Law of Large Numbers,

$$\frac{X}{10} \approx \frac{4}{7} = \frac{2p}{1+p}$$

Let's use this to estimate support for T, i.e., p:

$$(1+p)X \approx 20p$$

$$X \approx (20 - X)p$$

What is distribution of new X?

Let

$$X = \text{number of T voters in sample of 10}$$

As before: Binomial distribution but with different probability:

$$X \sim \text{Bin}(10, \frac{4}{7})$$

With Law of Large Numbers,

$$\frac{X}{10} \approx \frac{4}{7} = \frac{2p}{1+p}$$

Let's use this to estimate support for T, i.e., p:

$$(1+p)X \approx 20p$$

$$X \approx (20 - X)p$$

$$\hat{p} = \frac{X}{20 - X}$$

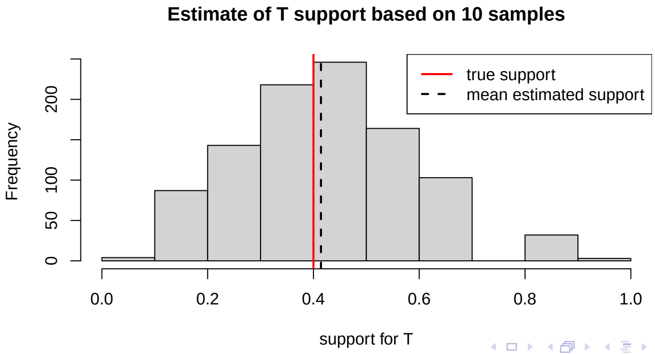
Estimating T support with unequal sampling probabilities

We sample 10 voters and find

$$X = 5$$

then estimated T support is:

$$\hat{p} = \frac{5}{20-5} = 0.33$$



Summary of Lecture 4

Explored **simple random sampling with replacement**:

- Binomial distribution:
  - Probabilities for different outcomes
  - Expected values
- Introduced Law of Large Numbers:
  - connection between sample average and population mean
- Extend Simple Random Sampling to Unequal Probabilities.