Simple Random Samples: Sections 7.2, 7.3.1, 7.3.2

Motivating Example (Section 7.2, Example A)

The population is 393 short-stay hospitals.

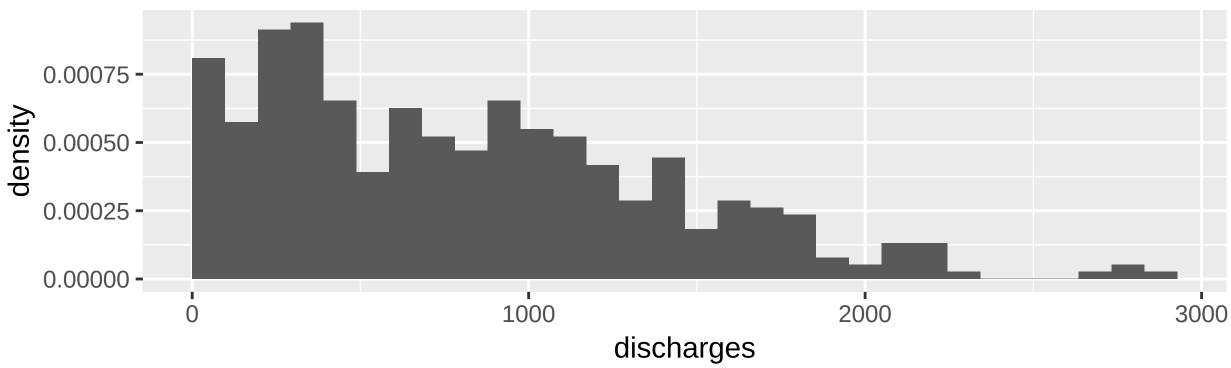
Let c_i denote the number of patients discharged from the ith hospital during January 1968 (our book is old).

```r
library(ggplot2) # for making plots

# read in data and look at first few rows
hospitals <- read.csv("http://www.evanlray.com/data/rice/Chapter%207/hospitals.txt")
head(hospitals)
```

```
##   discharges beds
## 1         57   10
## 2         35   16
## 3         23   20
## 4        120   24
## 5         92   25
## 6         98   26
```

```r
# make a histogram
ggplot(data = hospitals, mapping = aes(x = discharges, y = ..density..)) +
  geom_histogram(boundary = 0, bins = 30)
```



Suppose we want to estimate the average number of patients discharged across this population of all hospitals, based on a sample from that population.

Population

- N is the population size ($N = 393$)
- $c_1, c_2, \dots, c_N$ are distinct values in the population ($c_1 = 57$, $c_2 = 35$, ...)
  - For this chapter, we regard these as N **fixed constant** values (hence c_i); not random!

Population Parameters

- A **population parameter** is a number describing the population. Examples:
  - **population mean**: $\mu = \frac{1}{N}\sum_{i=1}^N c_i$
  - **population total**: $\tau = \sum_{i=1}^N c_i = N\mu$
  - **population variance**: $\sigma^2 = \frac{1}{N}\sum_{i=1}^N (c_i - \mu)^2$
- More generally, a **parameter** is a constant in a probability model, such as the mean and variance of a Normal distribution.

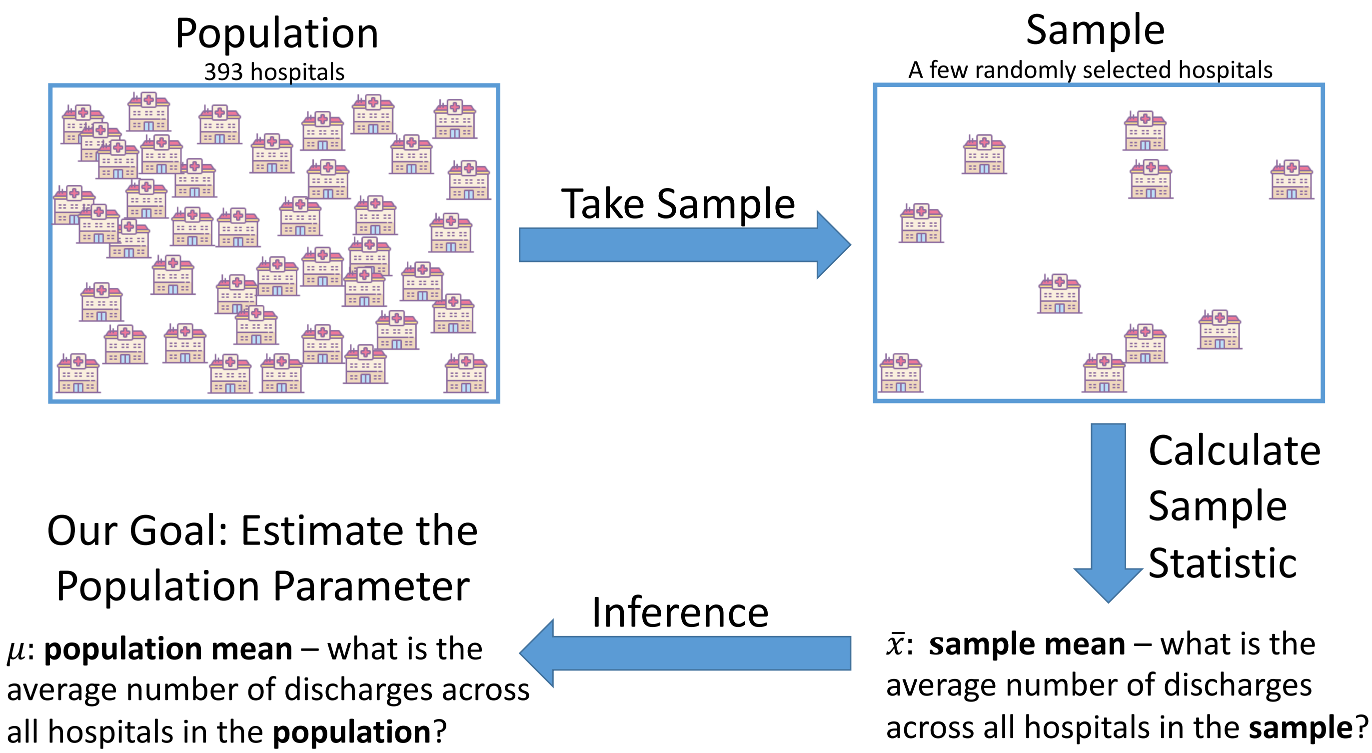
Samples

- **Sample**: n observations from the population that we will use to estimate population parameters of interest.
- **Simple random sample**: A sample obtained from a procedure where every sample of size n has the same probability of being drawn from the population
  - For this chapter, **the random mechanism is in how the sample is selected** from the population of fixed values.

Statistics and Estimators

- **Statistic**: A number describing the sample. Depending on context, this may refer to either:
  - A random variable: each sample will result in a different value of the statistic.
    - Example: $\bar{X} = \sum_{i=1}^{n} X_i$ (average number of discharges across hospitals in a sample that has not yet been selected; different samples will result in different realized values).
  - The realized value of this random variable for a particular sample.
    - Example: $\bar{x} = \sum_{i=1}^{n} x_i$ (average number of discharges across hospitals in a sample we have taken; this is a number).
- **Estimator**: A statistic (as a random variable) that is used to estimate a population parameter.
  - The random variable $\bar{X}$ may be used as an estimator for the population mean μ.
- **Estimate**: The realized value of an estimator for a particular sample.
  - For a particular sample, the realized value $\bar{x}$ may be used as an estimate for the population mean μ.

Summary in a Picture



Note: Confusingly, our book labels the observations in the population as $x_1, \ldots, x_N$ and in the sample as $x_1, \ldots, x_n$ – but the first sample item value, x_1, may not be the same as the first population item value c_1.