University of BRISTOL

Foundations of statistical estimation: Consistency, bias and variance

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.

What will we cover today?

- We will view sample statistics as estimators of parameters of interest.
- We will discuss the concept of statistical consistency.
- We will also introduce the ideas of statistical bias and the bias-variance decomposition.
- We will also consider the concept of a minimum variance unbiased-estimator.

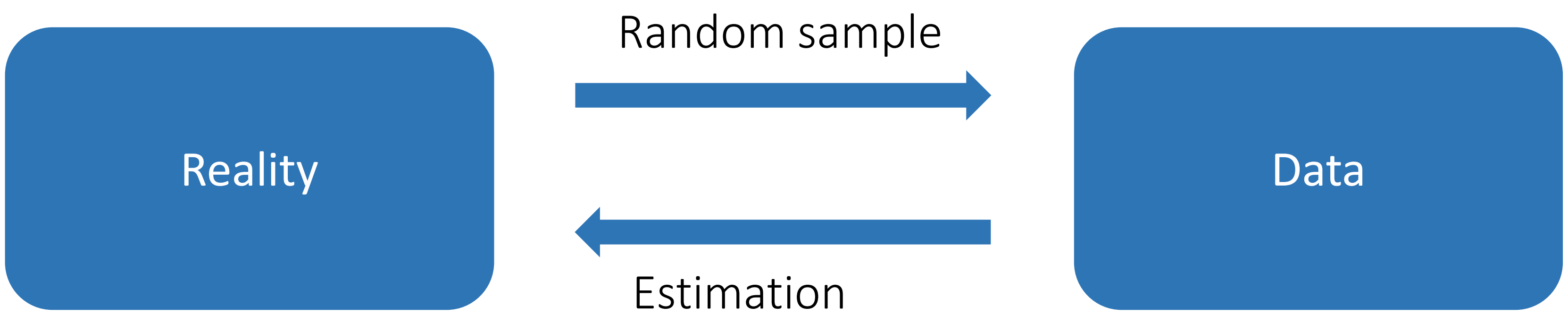
Samples and populations

We attempt to understand populations of penguins by looking at random samples.

Population

Sample

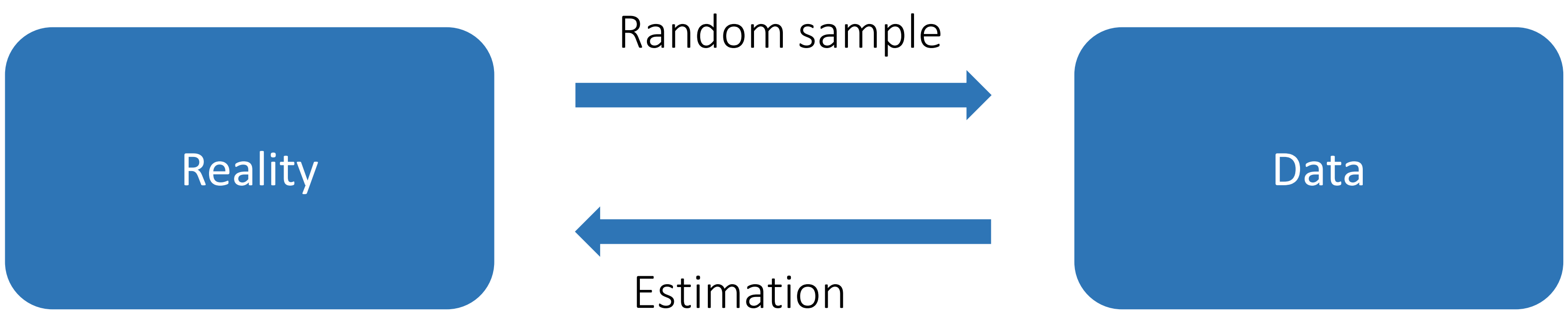
Statistical estimation



Population statistic

Sample statistic

Statistical estimation



Population statistic

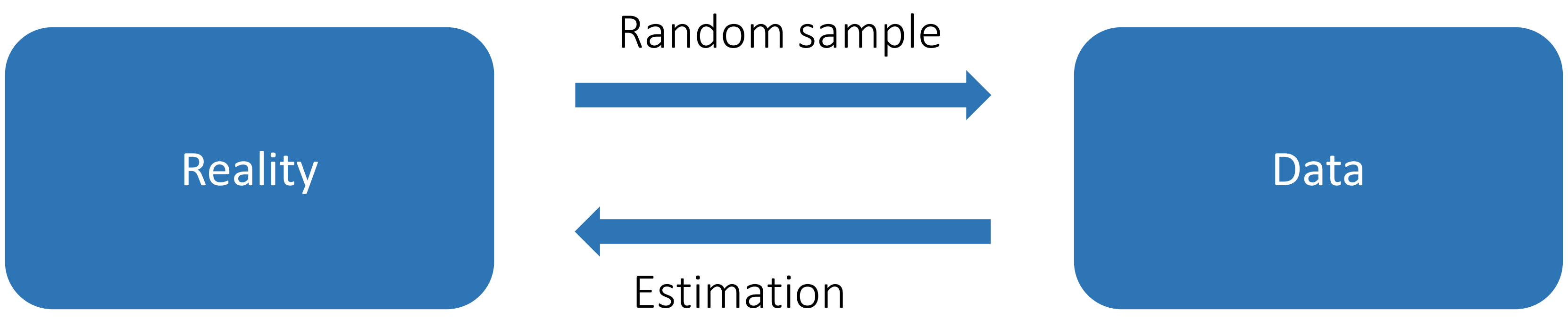
Sample statistic

Mean

$$\mu = \mathbb{E}[X_i]$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Statistical estimation



	Population statistic	Sample statistic
Mean	$\mu = \mathbb{E}[X_i]$	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Variance	$\sigma^2 = \mathbb{E}\left[(X_i - \mu)^2\right]$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

Probabilistic modelling

It is often useful to model our data as being generated by a **probabilistic model** $\mathbb{P}_\theta$

Probabilistic modelling

It is often useful to model our data as being generated by a **probabilistic model** $\mathbb{P}_\theta$

Examples

1. Suppose we have a sequence $(X_i)_{i=1}^n$ in $\{0,1\}^n$ corresponding to pass or fail for a driving test.

   We can model $(X_i)_{i=1}^n$ as a sequence of independent and identically distributed Bernoulli RVs

$$X_1, \cdots, X_n \sim \mathcal{B}(q) \qquad\qquad \theta = q$$

Probabilistic modelling

It is often useful to model our data as being generated by a **probabilistic model** $\mathbb{P}_\theta$

Examples

1. Suppose we have a sequence $(X_i)_{i=1}^n$ in $\{0,1\}^n$ corresponding to pass or fail for a driving test.

   We can model $(X_i)_{i=1}^n$ as a sequence of independent and identically distributed Bernoulli RVs

   $$X_1, \cdots, X_n \sim \mathcal{B}(q) \qquad\qquad \theta = q$$

2. Suppose we have a sequence $(X_i)_{i=1}^n$ in $\mathbb{R}^n$ corresponding to the height of a penguin.

   We can model $(X_i)_{i=1}^n$ as a sequence of independent and identically distributed Gaussian RVs

   $$X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2) \qquad\qquad \theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$$

Probabilistic modelling

It is often useful to model our data as being generated by a **probabilistic model** $\mathbb{P}_\theta$

We often use prior knowledge to choose the form of our model e.g. Gaussian, Bernoulli, ….

Probabilistic modelling

It is often useful to model our data as being generated by a **probabilistic model** $\mathbb{P}_\theta$

We often use prior knowledge to choose the form of our model e.g. Gaussian, Bernoulli, ….

We need to estimate the parameters θ in our model based upon a sample $X_1, \cdots, X_n \sim \mathbb{P}_\theta$

Probabilistic modelling

It is often useful to model our data as being generated by a **probabilistic model** $\mathbb{P}_\theta$

We often use prior knowledge to choose the form of our model e.g. Gaussian, Bernoulli, ….

We need to estimate the parameters θ in our model based upon a sample $X_1, \cdots, X_n \sim \mathbb{P}_\theta$

We estimate our parameters based upon **sample statistics**:

Functions of your sample $\hat{\theta} = g(X_1, \cdots, X_n)$ that don't depend on θ.

Probabilistic modelling

It is often useful to model our data as being generated by a **probabilistic model** $\mathbb{P}_\theta$

We often use prior knowledge to choose the form of our model e.g. Gaussian, Bernoulli, ….

We need to estimate the parameters θ in our model based upon a sample $X_1, \cdots, X_n \sim \mathbb{P}_\theta$

We estimate our parameters based upon **sample statistics**:

Functions of your sample $\hat{\theta} = g(X_1, \cdots, X_n)$ that don't depend on θ.

Note that sample statistics depend on the sample, so are themselves random variables.

Probabilistic modelling

It is often useful to model our data as being generated by a **probabilistic model** $\mathbb{P}_\theta$

We often use prior knowledge to choose the form of our model e.g. Gaussian, Bernoulli, ….

We need to estimate the parameters θ in our model based upon a sample $X_1, \cdots, X_n \sim \mathbb{P}_\theta$

We estimate our parameters based upon **sample statistics**:

Functions of your sample $\hat{\theta} = g(X_1, \cdots, X_n)$ that don't depend on θ.

Note that sample statistics depend on the sample, so are themselves random variables.

Our goal is to find statistics $\hat{\theta} = g(X_1, \cdots, X_n)$ approximating θ or coordinates θ_j

Statistical estimation

Our goal is to find statistics $\hat{\theta} = g(X_1, \cdots, X_n)$ approximating θ or coordinates θ_j.

Statistical estimation

Our goal is to find statistics $\hat{\theta} = g(X_1, \cdots, X_n)$ approximating θ or coordinates θ_j .

Examples

1. Suppose $X_1, \cdots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$

   We estimate q with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Statistical estimation

Our goal is to find statistics $\hat{\theta} = g(X_1, \cdots, X_n)$ approximating θ or coordinates θ_j.

Examples

1. Suppose $X_1, \cdots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$

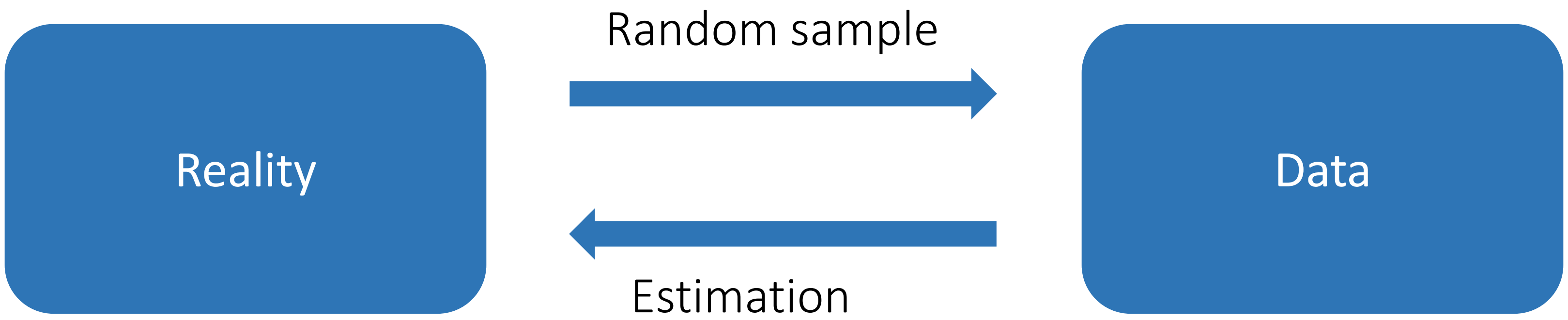
   We estimate q with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

2. Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. with parameters $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

   We estimate μ with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

   Similarly, we estimate σ^2 with the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$

Statistical estimation



Population

Sample

$$\theta$$

$$\hat{\theta} = g(X_1, \cdots, X_n)$$

Sample statistics are computed from a random sample so are themselves random variables.

The goal is to find sample statistics which closely approximate corresponding population statistics.

Examples of statistical estimation

Example 1

Suppose $X_1, \ldots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$.

We estimate estimate $q = \mathbb{E}(X_i)$ with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Examples of statistical estimation

Example 1

Suppose $X_1, \ldots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$.

We estimate estimate $q = \mathbb{E}(X_i)$ with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

```
set.seed(0)
num_trials<-1000
sample_size<-30
q<-0.3
```

Examples of statistical estimation

Example 1

Suppose $X_1, \dots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$.

We estimate estimate $q = \mathbb{E}(X_i)$ with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

```
set.seed(0)
num_trials<-1000
sample_size<-30
q<-0.3


simulation_df<-data.frame(trial=seq(num_trials))%>%
  mutate(simulation=map(.x=trial,.f=~rbinom(sample_size,1,q)))%>%
  # simulate sequences of Bernoulli random variables
```

Examples of statistical estimation

Example 1

Suppose $X_1, \ldots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$.

We estimate estimate $q = \mathbb{E}(X_i)$ with the sample mean $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

```
set.seed(0)
num_trials<-1000
sample_size<-30
q<-0.3



simulation_df<-data.frame(trial=seq(num_trials))%>%
  mutate(simulation=map(.x=trial,.f=~rbinom(sample_size,1,q)))%>%
  # simulate sequences of Bernoulli random variables
  mutate(sample_mean=map_dbl(.x=simulation,.f=mean))
  # compute the sample means
```

Examples of statistical estimation

Example 1

Suppose $X_1, \ldots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$.

We estimate estimate $q = \mathbb{E}(X_i)$ with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

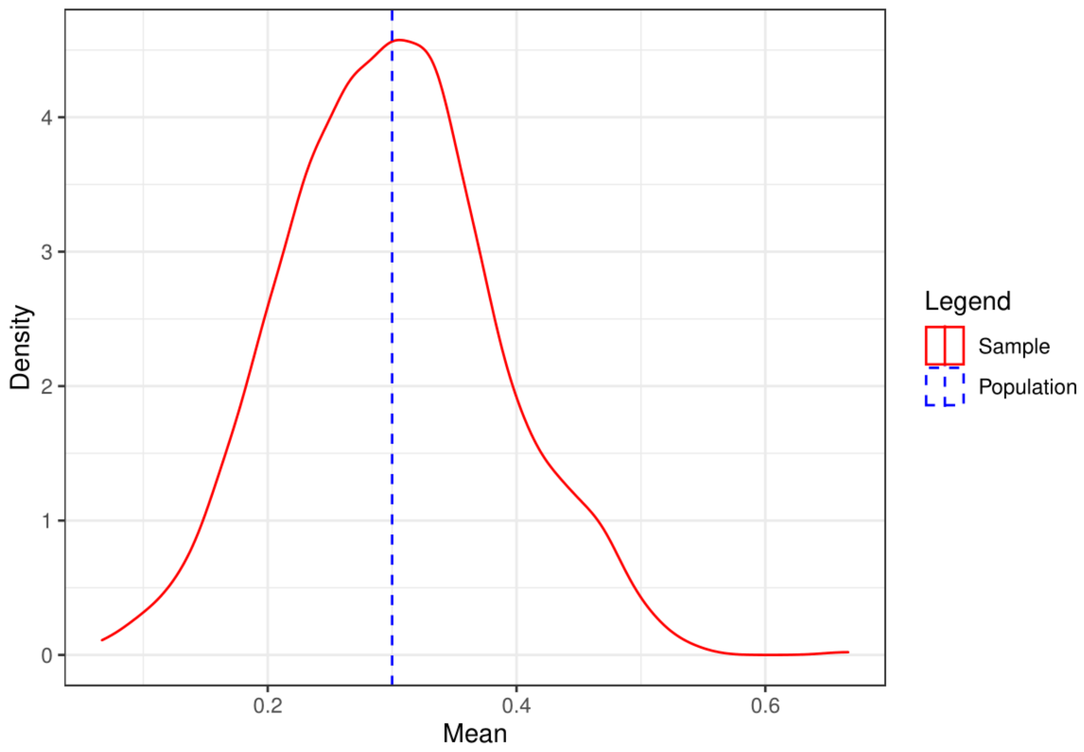
```
ggplot()+labs(x="Mean",y="Density")+theme_bw()+
  geom_density(data=simulation_df,
                aes(x=sample_mean,color="Sample",
                    linetype="Sample"))+
  # kernel density plot of sample means
  geom_vline(aes(xintercept=q,color="Population",
               linetype="Population"))+
  # vertical line displaying population mean
  scale_color_manual(name = "Legend",
                    values=c("Sample"="red", "Population"="blue"))+
  scale_linetype_manual(name="Legend",
                       values=c("Sample"="solid", "Population"="dashed"))
```

Examples of statistical estimation

Example 1

Suppose $X_1, \ldots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$.

We estimate estimate $q = \mathbb{E}(X_i)$ with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.



Examples of statistical estimation

Example 1

Suppose $X_1, \ldots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$.

We estimate estimate $q = \mathbb{E}(X_i)$ with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

```
set.seed(0)
num_trials_per_sample_size<-10
max_sample_size<-10000
q<-0.3
```

Examples of statistical estimation

Example 1

Suppose $X_1, \ldots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$.

We estimate estimate $q = \mathbb{E}(X_i)$ with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

```
set.seed(0)
num_trials_per_sample_size<-10
max_sample_size<-10000
q<-0.3


sim_by_n_df<-crossing(trial=seq(num_trials_per_sample_size),
                      sample_size=seq(to=sqrt(max_sample_size),by=0.1)**2)%>%
  # create data frame of all pairs of sample_size and trial
```

Examples of statistical estimation

Example 1

Suppose $X_1, \ldots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$.

We estimate estimate $q = \mathbb{E}(X_i)$ with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

```r
set.seed(0)
num_trials_per_sample_size<-10
max_sample_size<-10000
q<-0.3


sim_by_n_df<-crossing(trial=seq(num_trials_per_sample_size),
                      sample_size=seq(to=sqrt(max_sample_size),by=0.1)**2)%>%
  # create data frame of all pairs of sample_size and trial
  mutate(simulation=pmap(.l=list(trial,sample_size),.f=~rbinom(.y,1,q)))%>%
  # simulate sequences of Bernoulli random variables
  mutate(sample_mean=map_dbl(.x=simulation,.f=mean))%>%
  # compute the sample means
```

Examples of statistical estimation

Example 1

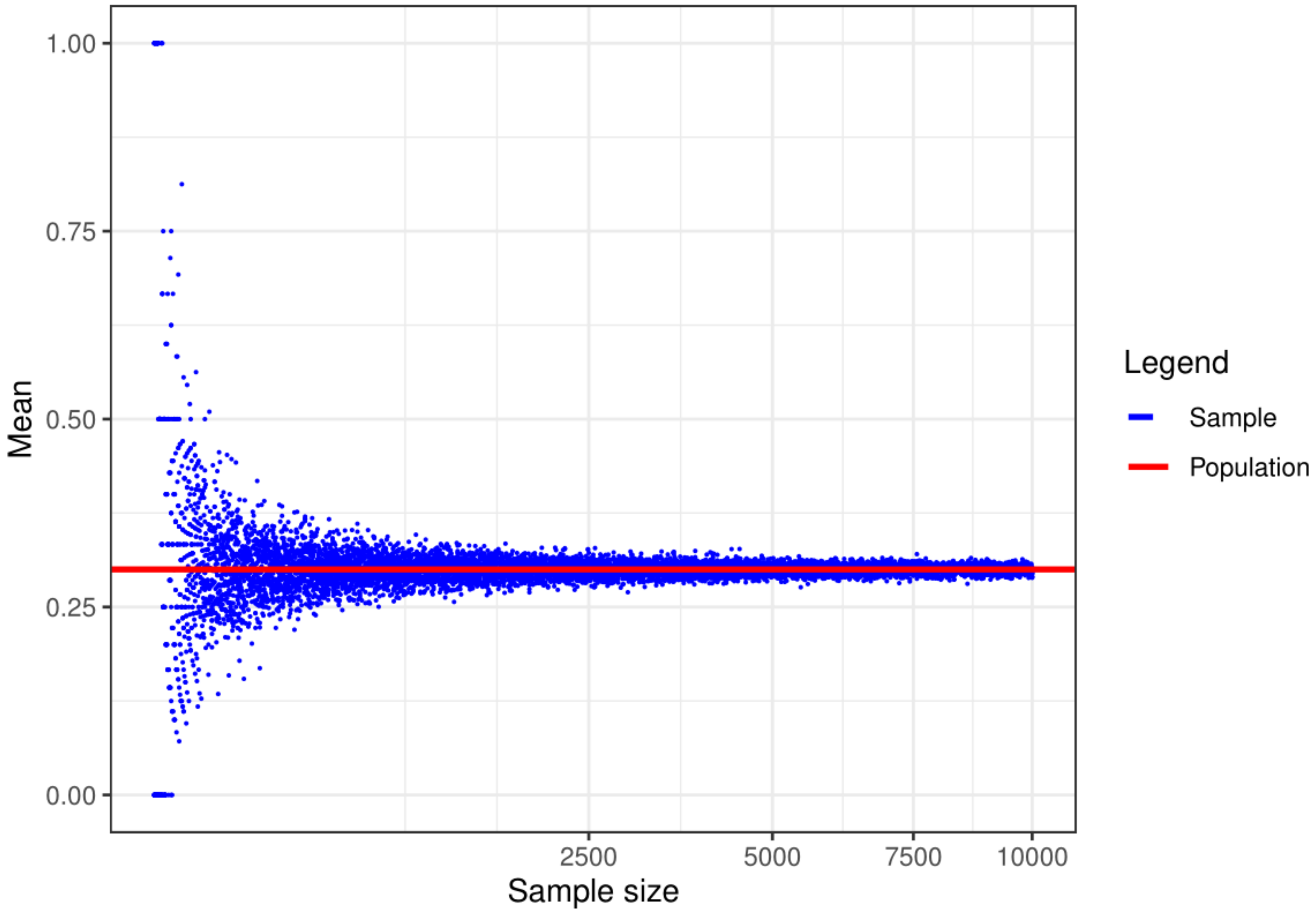
Suppose $X_1, \ldots, X_n \sim \mathcal{B}(q)$ are i.i.d. with a single parameter $\theta = q$.

We estimate estimate $q = \mathbb{E}(X_i)$ with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

```
ggplot()+labs(x="Sample size",y="Mean")+theme_bw()+
  geom_point(data=sim_by_n_df,
                aes(x=sample_size,y=sample_mean,color="Sample",
                    linetype="Sample"),size=0.1)+
  # scatter plot of sample means
  geom_hline(aes(yintercept=q,color="Population",
                linetype="Population"),size=1)+
  # horizontal line displaying population mean
  scale_color_manual(name = "Legend",
                    values=c("Sample"="blue", "Population"="red"))+
  scale_linetype_manual(name="Legend",
                    values=c("Sample"="dashed", "Population"="solid"))+
  scale_x_sqrt()
```

Examples of statistical estimation

Example 1



Examples of statistical estimation

Example 2

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. with parameters $\theta = (\mu, \sigma^2)$.

We estimate estimate $\sigma^2 = \mathrm{Var}(X_i)$ with the sample variance $\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$.

Examples of statistical estimation

Example 2

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. with parameters $\theta = (\mu, \sigma^2)$.

We estimate estimate $\sigma^2 = \mathrm{Var}(X_i)$ with the sample variance $\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$.

```r
set.seed(0)
num_trials<-1000
sample_size<-30
mu<-1
sigma_sqr<-3

simulation_df<-data.frame(trial=seq(num_trials))%>%
  mutate(simulation=map(.x=trial,
                        .f=~rnorm(sample_size,mean=mu,sd=sqrt(sigma_sqr))))%>%
  # simulate sequences of Gaussian random variables
  mutate(sample_var=map_dbl(.x=simulation,.f=var))
  # compute the sample variances
```

Examples of statistical estimation

Example 2

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. with parameters $\theta = (\mu, \sigma^2)$.

We estimate estimate $\sigma^2 = \mathrm{Var}(X_i)$ with the sample variance $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

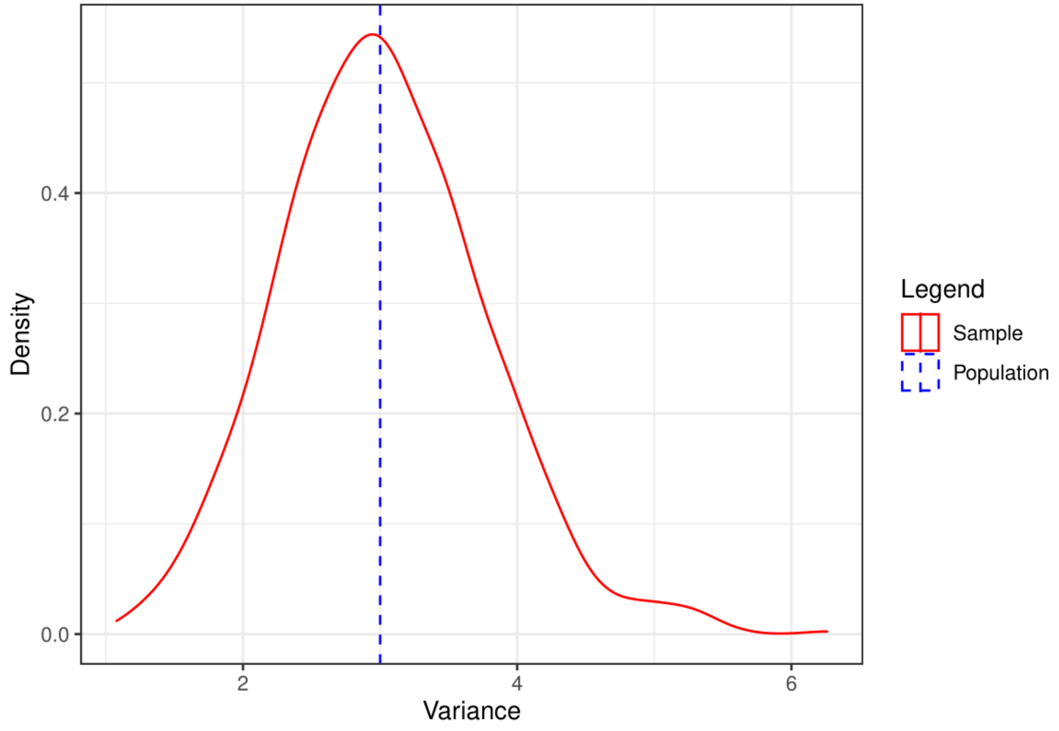
```
ggplot()+labs(x="Variance",y="Density")+theme_bw()+
  geom_density(data=simulation_df,
                aes(x=sample_var,color="Sample",
                    linetype="Sample"))+
  # kernel density plot of sample variances
  geom_vline(aes(xintercept=sigma_sqr,color="Population",
                 linetype="Population"))+
  # vertical line displaying population mean
  scale_color_manual(name = "Legend",
                     values=c("Sample"="red", "Population"="blue"))+
  scale_linetype_manual(name="Legend",
                        values=c("Sample"="solid", "Population"="dashed"))
```

Examples of statistical estimation

Example 2

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. with parameters $\theta = (\mu, \sigma^2)$.

We estimate estimate $\sigma^2 = \mathrm{Var}(X_i)$ with the sample variance $\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$.



Examples of statistical estimation

Example 2

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. with parameters $\theta = (\mu, \sigma^2)$.

We estimate estimate $\sigma^2 = \mathrm{Var}(X_i)$ with the sample variance $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

```r
set.seed(0)
num_trials_per_sample_size<-10
max_sample_size<-10000
mu<-1
sigma_sqr<-3


sim_by_n_df<-crossing(trial=seq(num_trials_per_sample_size),
                      sample_size=seq(to=sqrt(max_sample_size),by=0.1)**2)%>%
  # create data frame of all pairs of sample_size and trial
  mutate(simulation=pmap(.l=list(trial,sample_size),
                         .f=~rnorm(.y,mean=mu,sd=sqrt(sigma_sqr))))%>%
  # simulate sequences of Gaussian random variables
  mutate(sample_var=map_dbl(.x=simulation,.f=var))
  # compute the sample variances
```

Examples of statistical estimation

Example 2

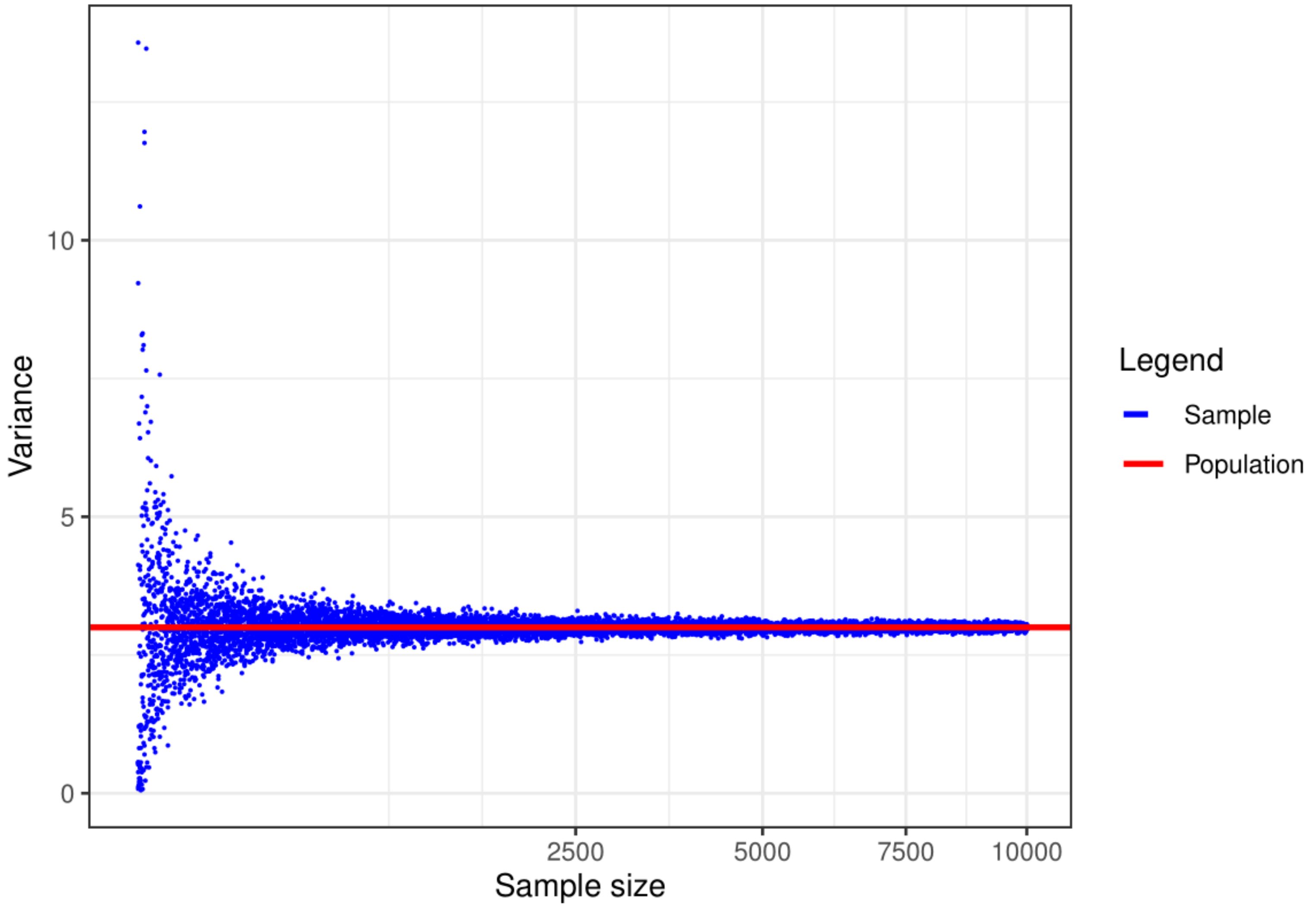
Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. with parameters $\theta = (\mu, \sigma^2)$.

We estimate estimate $\sigma^2 = \mathrm{Var}(X_i)$ with the sample variance $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

```
ggplot()+labs(x="Sample size",y="Variance")+theme_bw()+
  geom_point(data=sim_by_n_df,
                  aes(x=sample_size,y=sample_var,color="Sample",
                      linetype="Sample"),size=0.1)+
  # scatter plot of sample variances
  geom_hline(aes(yintercept=sigma_sqr,color="Population",
                 linetype="Population"),size=1)+
  # horizontal line displaying population variance
  scale_color_manual(name = "Legend",
                     values=c("Sample"="blue", "Population"="red"))+
  scale_linetype_manual(name="Legend",
                        values=c("Sample"="dashed", "Population"="solid"))+
  scale_x_sqrt()
```

Examples of statistical estimation

Example 2



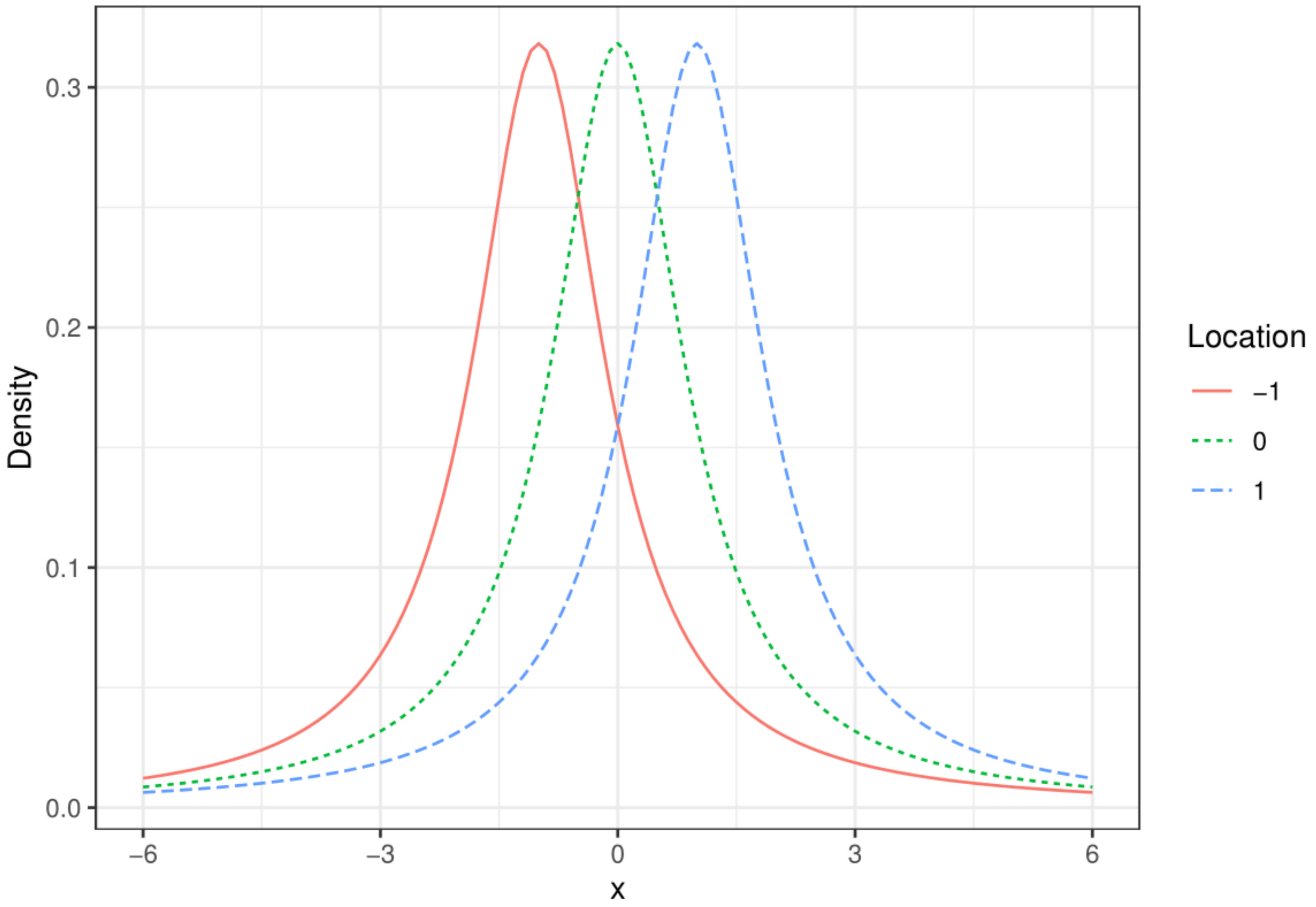
The Cauchy distribution

A random variable X has a Cauchy distribution with location parameter θ if its density is

$$f_\theta(x) := \frac{1}{\pi\left\{1 + (x-\theta)^2\right\}}.$$

The Cauchy distribution

A Cauchy random variable with location parameter θ has density $f_\theta(x) := \frac{1}{\pi\{1+(x-\theta)^2\}}$.



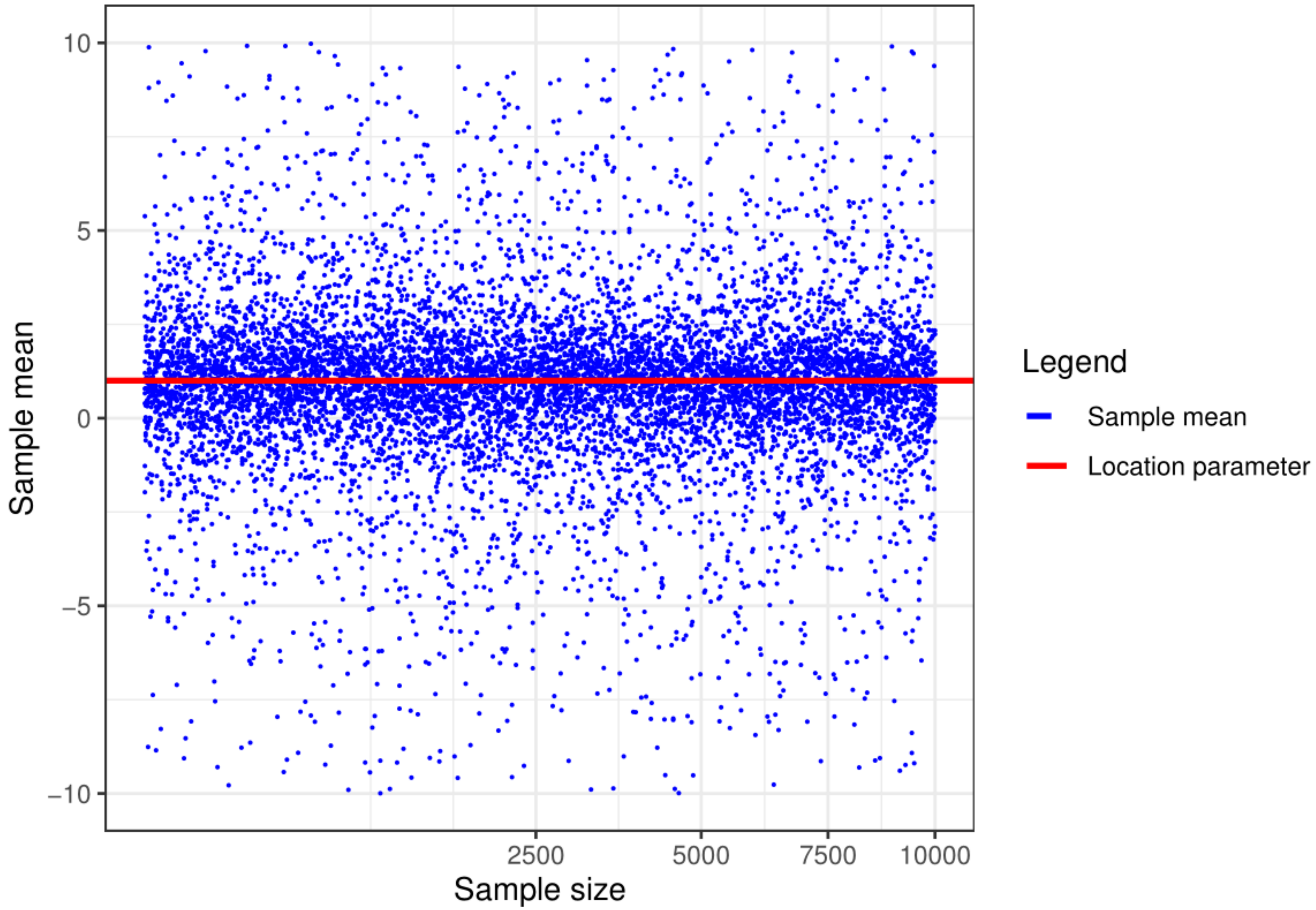
The Cauchy distribution

A Cauchy random variable with location parameter θ has density $f_\theta(x) := \frac{1}{\pi\left\{1+(x-\theta)^2\right\}}$.

```
set.seed(0)
num_trials_per_sample_size<-10
max_sample_size<-10000
theta<-1


sim_by_n_df<-crossing(trial=seq(num_trials_per_sample_size),
                      sample_size=seq(to=sqrt(max_sample_size),by=0.1)**2)%>%
  # create data frame of all pairs of sample_size and trial
  mutate(simulation=pmap(.l=list(trial,sample_size),.f=~rcauchy(.y,location=theta)))%>%
  # simulate sequences of Cauchy random variables
  mutate(sample_mean=map_dbl(.x=simulation,.f=mean))
  # compute the sample means
```

The Cauchy distribution

A Cauchy random variable with location parameter θ has density $f_\theta(x) := \frac{1}{\pi\{1+(x-\theta)^2\}}$.



The Cauchy distribution

A Cauchy random variable with location parameter θ has density $f_\theta(x) := \frac{1}{\pi\left\{1+(x-\theta)^2\right\}}$.

A Cauchy random variable has cumulative distribution function

$$F_\theta(x) = \mathbb{P}\left(X \leq x\right) = \int_{-\infty}^{x} f_\theta(z) = \frac{1}{\pi}\arctan\left(x-\theta\right) + \frac{1}{2}.$$

The Cauchy distribution

A Cauchy random variable with location parameter θ has density $f_\theta(x) := \frac{1}{\pi\left\{1+(x-\theta)^2\right\}}$.

A Cauchy random variable has cumulative distribution function

$$F_\theta(x) = \mathbb{P}\left(X \leq x\right) = \int_{-\infty}^{x} f_\theta(z) = \frac{1}{\pi} \arctan\left(x - \theta\right) + \frac{1}{2}.$$

The population median of the Cauchy random variable is

$$F_\theta^{-1}\left(\frac{1}{2}\right) = \inf\left\{x \in \mathbb{R} \;:\; F_\theta(x) \geq \frac{1}{2}\right\} = \theta.$$

The Cauchy distribution

A Cauchy random variable with location parameter θ has density $f_\theta(x) := \frac{1}{\pi\left\{1+(x-\theta)^2\right\}}$.

A Cauchy random variable has cumulative distribution function

$$F_\theta(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_\theta(z) = \frac{1}{\pi}\arctan(x-\theta) + \frac{1}{2}.$$

The population median of the Cauchy random variable is

$$F_\theta^{-1}\left(\frac{1}{2}\right) = \inf\left\{x \in \mathbb{R} \ : \ F_\theta(x) \geq \frac{1}{2}\right\} = \theta.$$

Suppose we have i.i.d. data $X_1, \ldots, X_n$ with Cauchy density f_θ.

A natural estimator for θ is the sample median $\hat{\theta} = \text{Median}(X_1, \ldots, X_n)$.

The Cauchy distribution

Example 3

Suppose we have i.i.d. data $X_1, \ldots, X_n$ with Cauchy density f_θ.

A natural estimator for θ is the sample median $\hat{\theta} = \text{Median}\,(X_1, \ldots, X_n)$.

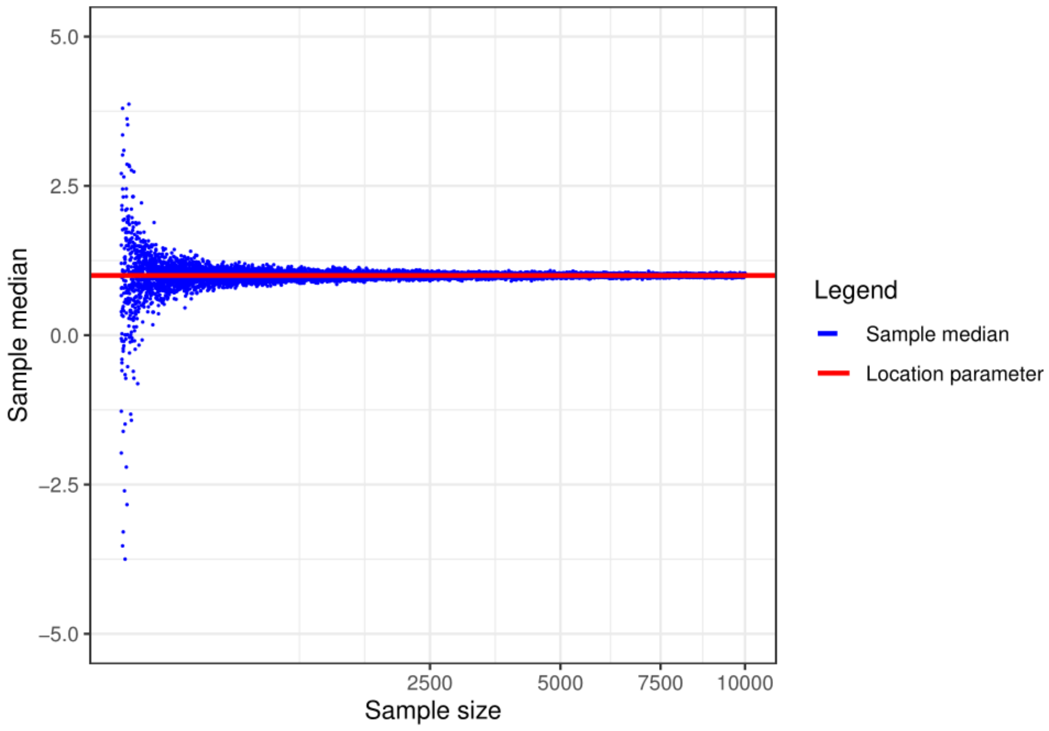
```r
set.seed(0)
num_trials_per_sample_size<-10
max_sample_size<-10000
theta<-1


sim_by_n_df<-crossing(trial=seq(num_trials_per_sample_size),
                      sample_size=seq(to=sqrt(max_sample_size),by=0.1)**2)%>%
  # create data frame of all pairs of sample_size and trial
  mutate(simulation=pmap(.l=list(trial,sample_size),.f=~rcauchy(.y,location=theta)))%>%
  # simulate sequences of Cauchy random variables
  mutate(sample_median=map_dbl(.x=simulation,.f=median))
  # compute the sample median
```

The Cauchy distribution

Example 3

Suppose we have i.i.d. data $X_1, \ldots, X_n$ with Cauchy density f_θ.

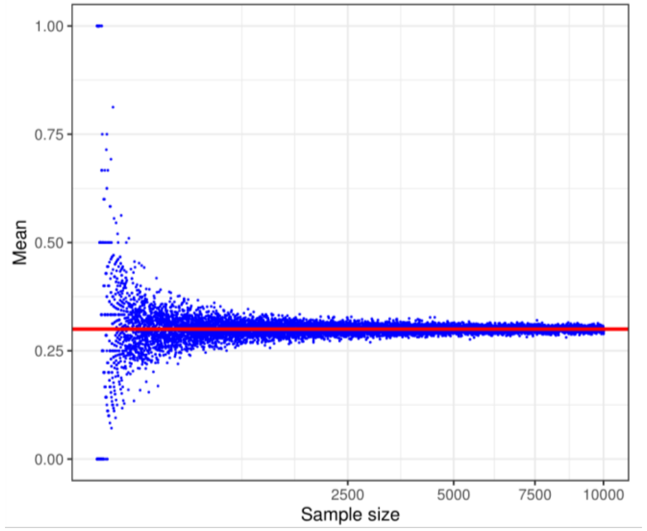
A natural estimator for θ is the sample median $\hat{\theta} = \text{Median}\,(X_1, \ldots, X_n)$.



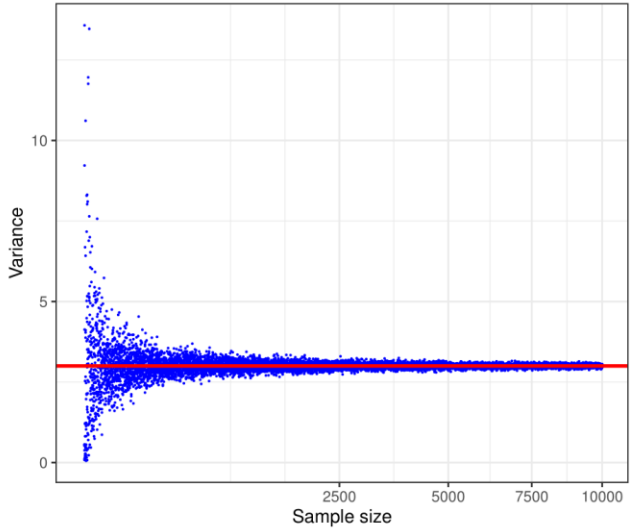
Consistency

We are interested in statistical estimators $\hat{\theta}$ which tend towards θ as $n \to \infty$.

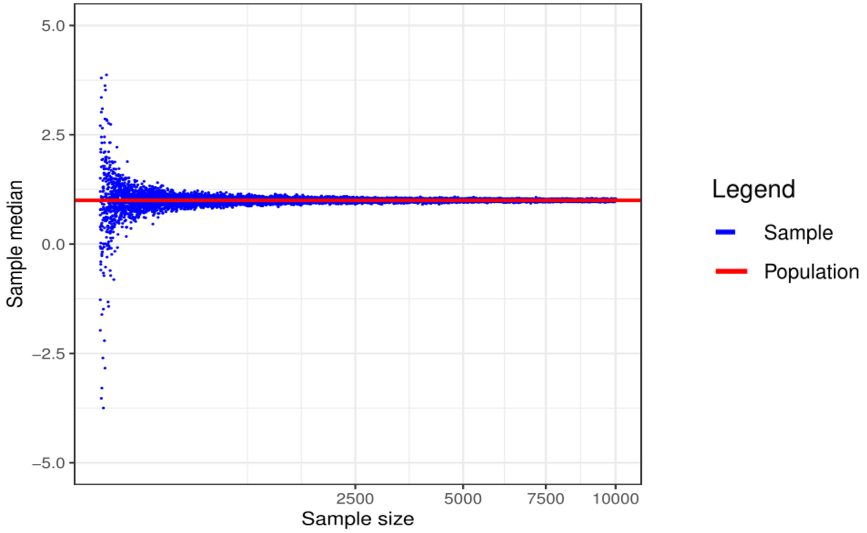
Example 1



Example 2



Example 3



We refer to such $\hat{\theta}$ as **consistent** estimators of θ.

Consistent estimators

A sample statistic $\hat{\theta} = g(X_1, \cdots, X_n)$ of a population parameter θ is **consistent** if

$$\hat{\theta} = g(X_1, \cdots, X_n) \quad \to \quad \theta \quad \text{as} \quad n \to \infty$$

More precisely, $\hat{\theta} = g(X_1, \cdots, X_n)$ is a consistent estimator of θ if for all $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left[|g(X_1, \cdots, X_n) - \theta| > \epsilon\right] = 0.$$

Consistency and the law of large numbers

A sample statistic $\hat{\theta}_n$ is a **consistent** estimator of θ if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \epsilon\right) = 0$.

Theorem (Bernoulli, circa. 1700). *Let $X : \Omega \to \mathbb{R}$ be a random variable with a well-defined expectation $\mu = \mathbb{E}(X)$. Let $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ be a sequence of independent copies of X. Then for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) = 0.$$

Consistency and the law of large numbers

A sample statistic $\hat{\theta}_n$ is a **consistent** estimator of θ if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \epsilon\right) = 0$.

Theorem (Bernoulli, circa. 1700). *Let $X : \Omega \to \mathbb{R}$ be a random variable with a well-defined expectation $\mu = \mathbb{E}(X)$. Let $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ be a sequence of independent copies of X. Then for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) = 0.$$

Example 1

Suppose that $X_1, \ldots, X_n$ are i.i.d. with expectation $\mu = \mathbb{E}(X_i) \in \mathbb{R}$.

Then the sample mean $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ is a consistent estimator of μ.

Consistency and the law of large numbers

A sample statistic $\hat{\theta}_n$ is a **consistent** estimator of θ if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \epsilon\right) = 0$.

Theorem (Bernoulli, circa. 1700). *Let $X : \Omega \to \mathbb{R}$ be a random variable with a well-defined expectation $\mu = \mathbb{E}(X)$. Let $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ be a sequence of independent copies of X. Then for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) = 0.$$

Example 2

Suppose that $X_1, \ldots, X_n$ are i.i.d. with expectation $\mu = \mathbb{E}(X_i) \in \mathbb{R}$ and variance $\sigma^2 = \mathrm{Var}(X_i) \in \mathbb{R}$.

Then the sample variance $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is a consistent estimator of σ^2.

Consistency of sample statistics

A sample statistic $\hat{\theta}_n$ is a **consistent** estimator of θ if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \epsilon\right) = 0$.

Examples

Suppose that $X_1, \ldots, X_n$ is a well-behaved sequence of independent and identically distributed data.

1. The sample mean $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ is a consistent estimator of μ.
2. The sample variance $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is a consistent estimator of σ^2.

Consistency of sample statistics

A sample statistic $\hat{\theta}_n$ is a **consistent** estimator of θ if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \epsilon\right) = 0$.

Examples

Suppose that $X_1, \ldots, X_n$ is a well-behaved sequence of independent and identically distributed data.

1. The sample mean $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ is a consistent estimator of μ.
2. The sample variance $\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ is a consistent estimator of σ^2.
3. The sample standard deviation $\sqrt{\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}$ is a consistent estimator of σ.

Consistency of sample statistics

A sample statistic $\hat{\theta}_n$ is a **consistent** estimator of θ if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \epsilon\right) = 0$.

Examples

Suppose that $X_1, \ldots, X_n$ is a well-behaved sequence of independent and identically distributed data.

1. The sample mean $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ is a consistent estimator of μ.
2. The sample variance $\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ is a consistent estimator of σ^2.
3. The sample standard deviation $\sqrt{\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}$ is a consistent estimator of σ.
4. Suppose that $X_1, \ldots, X_n$ are continuous random variables with continuous density f.

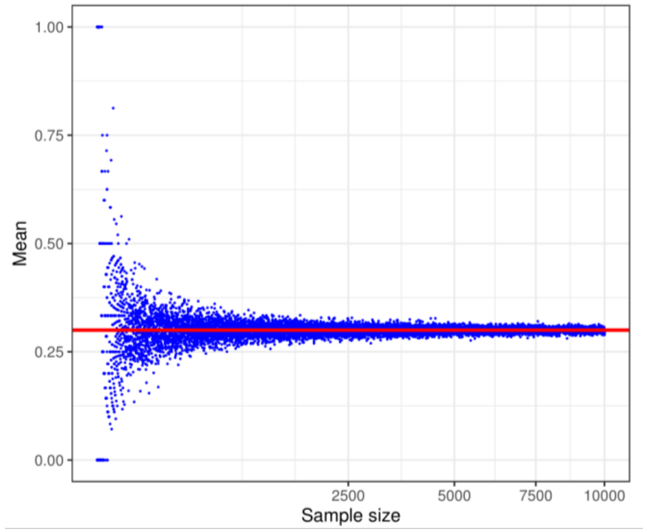
   Given $q \in (0,1)$, let $x_q \in \mathbb{R}$ be the population q-quantile and suppose $f(x_q) > 0$.

   Then the sample q-quantile is a consistent estimator or the population q-quantile x_q.

Consistency of sample statistics

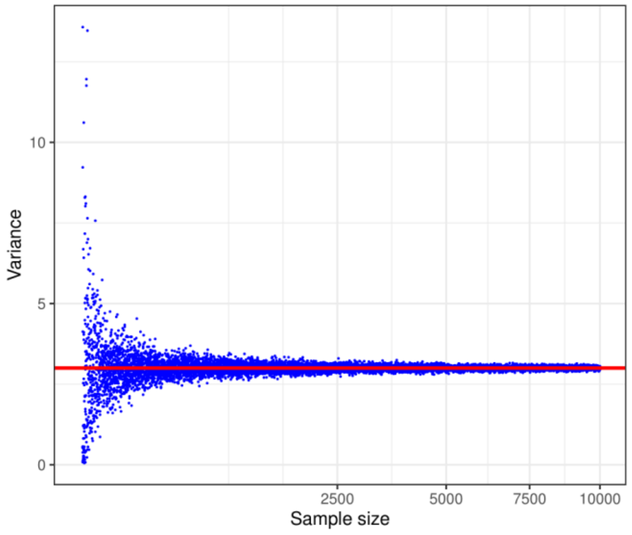
A sample statistic $\hat{\theta}_n$ is a **consistent** estimator of θ if for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \epsilon\right) = 0$.

Example 1



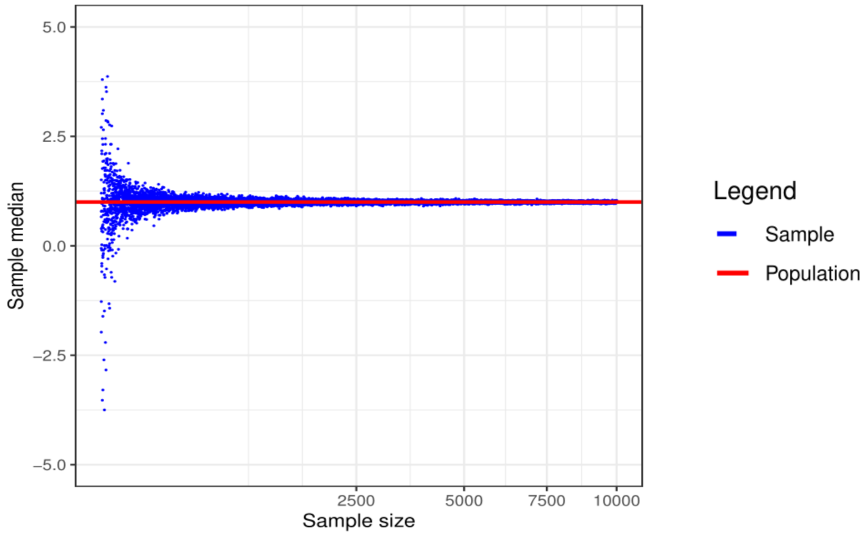
Sample mean

Example 2



Sample variance

Example 3



Sample median

Statistical bias

The **bias** of an estimator $\hat{\theta} = g(X_1, \cdots, X_n)$ of a population parameter θ is

$$\mathsf{Bias}(\hat{\theta}) := \mathbb{E}(\hat{\theta}) - \theta.$$

The estimator is said to be **unbiased** if $\mathrm{Bias}(\hat{\theta}) = 0.$

Statistical bias

The **bias** of an estimator $\hat{\theta} = g(X_1, \cdots, X_n)$ of a population parameter θ is

$$\text{Bias}(\hat{\theta}) := \mathbb{E}(\hat{\theta}) - \theta.$$

The estimator is said to be **unbiased** if $\mathrm{Bias}(\hat{\theta}) = 0.$

Given an independent and identically distributed sample $X_1, \cdots, X_n$ we have

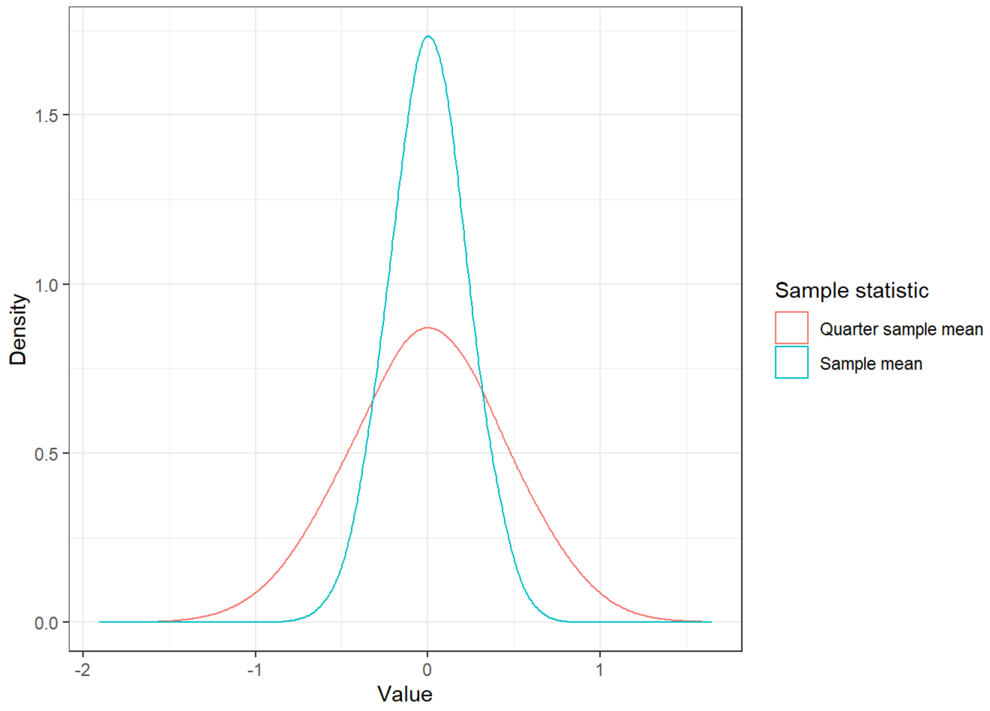
$$\mathrm{Bias}(\overline{X}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] - \mu = 0.$$

$$\mathrm{Bias}(s^2) = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right] - \sigma^2 = 0.$$

The variance of an estimator

The sample mean is an unbiased estimator for the population mean….

…. but so is the sample mean computed from just the first quarter of the data!



There are many unbiased estimators… however, many such estimators are very high variance!

Bias and variance of an estimator

Suppose that $\hat{\theta}$ is an estimator of a population parameter θ .

The **bias** of the estimator is defined by

$$\mathsf{Bias}(\hat{\theta}) := \mathbb{E}(\hat{\theta}) - \theta.$$

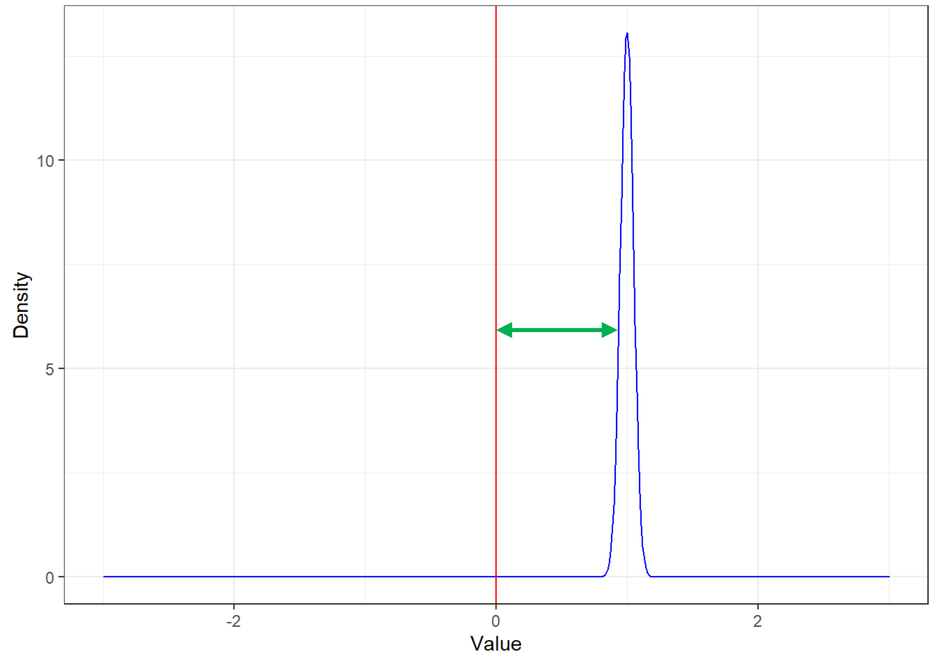
The **variance** of the estimator is defined by

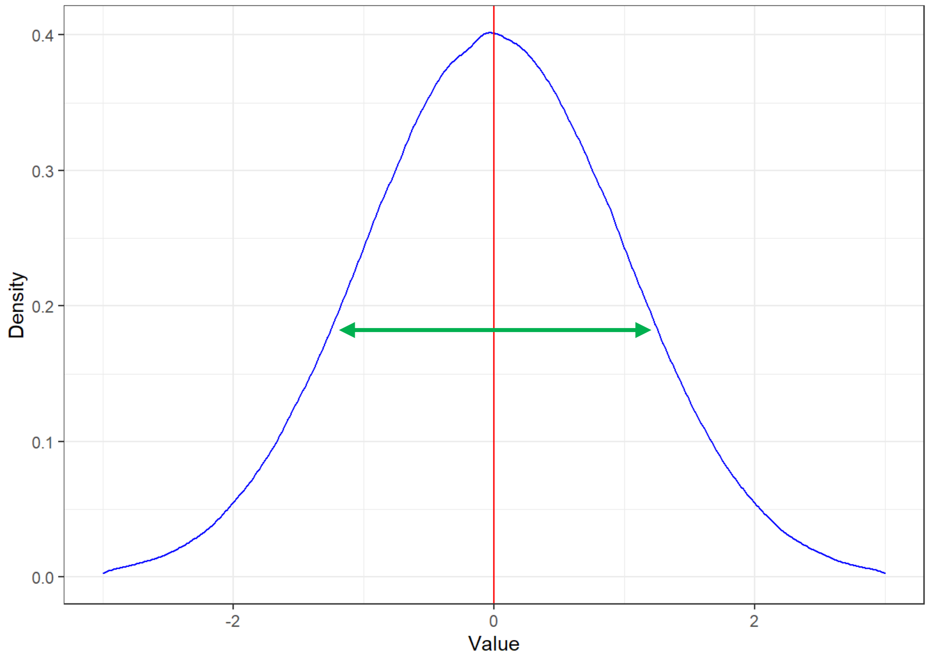
$$\mathsf{Var}(\hat{\theta}) := \mathbb{E}\left[\left\{\hat{\theta} - \mathbb{E}(\hat{\theta})\right\}^2\right]$$

Bias and variance of an estimator

$$\text{Bias}(\hat{\theta}) := \mathbb{E}(\hat{\theta}) - \theta.$$

$$\text{Var}(\hat{\theta}) := \mathbb{E}\left[\left\{\hat{\theta} - \mathbb{E}(\hat{\theta})\right\}^2\right]$$





Mean squared error

Suppose that $g(X_1, \cdots, X_n)$ is an estimator of a population parameter θ

The **bias** of the estimator is $\mathsf{Bias}(\hat{\theta}) := \mathbb{E}(\hat{\theta}) - \theta.$

The **variance** of the estimator is $\mathsf{Var}(\hat{\theta}) := \mathbb{E}\{\hat{\theta} - \mathbb{E}(\hat{\theta})\}^2$

The **mean square error** of an estimator is $\mathsf{MSE}(\hat{\theta}) := \mathbb{E}\{(\hat{\theta} - \theta)^2\}.$

Mean squared error

Suppose that $g(X_1, \cdots, X_n)$ is an estimator of a population parameter θ

The **bias** of the estimator is $\text{Bias}(\hat{\theta}) := \mathbb{E}(\hat{\theta}) - \theta.$

The **variance** of the estimator is $\text{Var}(\hat{\theta}) := \mathbb{E}\{\hat{\theta} - \mathbb{E}(\hat{\theta})\}^2$

The **mean square error** of an estimator is $\text{MSE}(\hat{\theta}) := \mathbb{E}\{(\hat{\theta} - \theta)^2\}.$

Theorem (Bias-variance decomposition). *Suppose that $\hat{\theta}$ is an estimator of a parameter θ. Then, $MSE(\hat{\theta}) = Bias(\hat{\theta})^2 + Var(\hat{\theta})$.*

Mean squared error

Theorem (Bias-variance decomposition). *Suppose that $\hat{\theta}$ is an estimator of a parameter θ. Then, $MSE(\hat{\theta}) = Bias(\hat{\theta})^2 + Var(\hat{\theta})$.*

Proof.

$$\begin{aligned}
\mathsf{MSE}(\hat{\theta}) &:= \mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left\{\left(\hat{\theta}-\mathbb{E}[\hat{\theta}]\right)+\left(\mathbb{E}[\hat{\theta}]-\theta\right)\right\}^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}[\hat{\theta}]\right)^2+\left(\mathbb{E}[\hat{\theta}]-\theta\right)^2+2\left(\hat{\theta}-\mathbb{E}[\hat{\theta}]\right)\left(\mathbb{E}[\hat{\theta}]-\theta\right)\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}[\hat{\theta}]\right)^2\right]+\left(\mathbb{E}[\hat{\theta}]-\theta\right)^2 \\
&= \mathsf{Var}(\hat{\theta})+\mathsf{Bias}(\hat{\theta})^2.
\end{aligned}$$

$\square$

Minimum variance unbiased estimator

Suppose that $\hat{\theta}$ is a statistical estimator of a population parameter θ.

We say that $\hat{\theta}$ is **unbiased** if $\mathbb{E}(\hat{\theta}) = \theta$.

We say that $\hat{\theta}$ is a **minimum variance unbiased estimator** (MVUE) if

1. $\hat{\theta}$ is unbiased i.e. $\mathbb{E}(\hat{\theta}) = \theta$;
2. $\hat{\theta}$ is minimum variance i.e. we have $\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\tilde{\theta})$, for all $\tilde{\theta}$ with $\mathbb{E}(\tilde{\theta}) = \theta$.

Remark: A MVUE also has minimal mean-square error over all unbiased estimators.

This is a consequence of the bias-variance decomposition $\mathsf{MSE}(\hat{\theta}) = \mathsf{Bias}(\hat{\theta})^2 + \mathsf{Var}(\hat{\theta})$.

Minimum variance unbiased estimator

A minimum variance unbiased estimator (MVUE) has minimum variance over all unbiased estimators.

<u>Example 1</u>

Suppose $X_1, \cdots, X_n \sim \mathcal{B}(q)$ are independent and identically distributed. Then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ is a MVUE of q

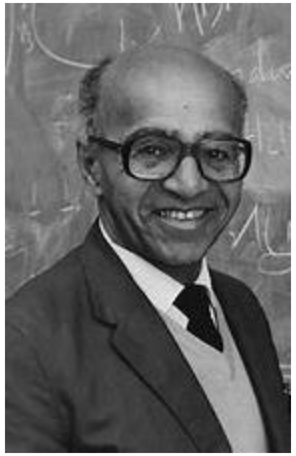
Minimum variance unbiased estimator

A minimum variance unbiased estimator (MVUE) has minimum variance over all unbiased estimators.

<u>Example 1</u>

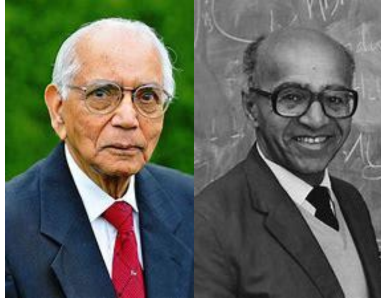
Suppose $X_1, \cdots, X_n \sim \mathcal{B}(q)$ are independent and identically distributed. Then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{is a MVUE of} \quad q$$



This is a consequence of the Rao-Blackwell theorem.

Minimum variance unbiased estimator



Example 2

Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are independent and identically distributed. Then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{is a MVUE of} \quad \mu = \mathbb{E}[X_i]$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \quad \text{is a MVUE of} \quad \sigma^2 = \mathbb{E}\left[(X_i - \mu)^2\right]$$

This is also a consequence of the Rao-Blackwell theorem.

What have we covered?

- We considered sample statistics as estimators of parameters of interest.
- We introduced the concept of statistical consistency.
- We also considered the idea of statistical bias and the bias-variance decomposition.
- We will discussed the concept of a minimum variance unbiased-estimator.
- In the next lecture we will consider a general-purpose approach to deriving statistical estimators.

University of BRISTOL

Thanks for listening!

Henry W J Reeve

henry.reeve@bristol.ac.uk

Include EMATM0061 in the subject of your email.

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.