AUA CS108, Statistics, Fall 2020

Lecture 26

Michael Poghosyan

26 Oct 2020

Contents

Example

Example: Assume we want to model the daily number of car accidents in some city.

Example

Example: Assume we want to model the daily number of car accidents in some city. Let X be that daily number of car accidents. Of course, X is a r.v.

Example

Example: Assume we want to model the daily number of car accidents in some city. Let X be that daily number of car accidents. Of course, X is a r.v. An appropriate Distribution for X will be

$$X \sim$$

Example

Example: Assume we want to model the daily number of car accidents in some city. Let X be that daily number of car accidents. Of course, X is a r.v. An appropriate Distribution for X will be

$$X \sim Pois(\lambda),$$

for some λ to be estimated.

Example

Example: Assume we want to model the daily number of car accidents in some city. Let X be that daily number of car accidents. Of course, X is a r.v. An appropriate Distribution for X will be

$$X \sim Pois(\lambda),$$

for some λ to be estimated.

Now, if we will collect data for some n days, we will get the Random Sample

$$X_1, X_2, ..., X_n \sim$$

Example

Example: Assume we want to model the daily number of car accidents in some city. Let X be that daily number of car accidents. Of course, X is a r.v. An appropriate Distribution for X will be

$$X \sim Pois(\lambda),$$

for some λ to be estimated.

Now, if we will collect data for some n days, we will get the Random Sample

$$X_1, X_2, ..., X_n \sim Pois(\lambda).$$

Example

Example: Assume we want to model the daily number of car accidents in some city. Let X be that daily number of car accidents. Of course, X is a r.v. An appropriate Distribution for X will be

$$X \sim Pois(\lambda),$$

for some λ to be estimated.

Now, if we will collect data for some n days, we will get the Random Sample

$$X_1, X_2, ..., X_n \sim Pois(\lambda).$$

After collecting that data, we will get the Dataset $x_1, x_2, ..., x_n$ of the daily number of car accidents for day $1, 2, ..., n$.

Example

Example: Assume we want to model the daily number of car accidents in some city. Let X be that daily number of car accidents. Of course, X is a r.v. An appropriate Distribution for X will be

$$X \sim Pois(\lambda),$$

for some λ to be estimated.

Now, if we will collect data for some n days, we will get the Random Sample

$$X_1, X_2, ..., X_n \sim Pois(\lambda).$$

After collecting that data, we will get the Dataset $x_1, x_2, ..., x_n$ of the daily number of car accidents for day $1, 2, ..., n$.

- Here our Parameter is $\lambda = \theta$; it is 1D;

Example

Example: Assume we want to model the daily number of car accidents in some city. Let X be that daily number of car accidents. Of course, X is a r.v. An appropriate Distribution for X will be

$$X \sim Pois(\lambda),$$

for some λ to be estimated.

Now, if we will collect data for some n days, we will get the Random Sample

$$X_1, X_2, ..., X_n \sim Pois(\lambda).$$

After collecting that data, we will get the Dataset $x_1, x_2, ..., x_n$ of the daily number of car accidents for day $1, 2, ..., n$.

- Here our Parameter is $\lambda = \theta$; it is 1D;
- The set of Parameters is $(0, +\infty) = \Theta \subset \mathbb{R}$;

Example

Example: Assume we want to model the daily number of car accidents in some city. Let X be that daily number of car accidents. Of course, X is a r.v. An appropriate Distribution for X will be

$$X \sim Pois(\lambda),$$

for some λ to be estimated.

Now, if we will collect data for some n days, we will get the Random Sample

$$X_1, X_2, ..., X_n \sim Pois(\lambda).$$

After collecting that data, we will get the Dataset $x_1, x_2, ..., x_n$ of the daily number of car accidents for day $1, 2, ..., n$.

- Here our Parameter is $\lambda = \theta$; it is 1D;
- The set of Parameters is $(0, +\infty) = \Theta \subset \mathbb{R}$;
- The Parametric Family of Distributions is $Pois(\lambda)$.

Example

Example: Assume we want to model the daily number of car accidents in some city. Let X be that daily number of car accidents. Of course, X is a r.v. An appropriate Distribution for X will be

$$X \sim Pois(\lambda),$$

for some λ to be estimated.

Now, if we will collect data for some n days, we will get the Random Sample

$$X_1, X_2, ..., X_n \sim Pois(\lambda).$$

After collecting that data, we will get the Dataset $x_1, x_2, ..., x_n$ of the daily number of car accidents for day $1, 2, ..., n$.

- Here our Parameter is $\lambda = \theta$; it is 1D;
- The set of Parameters is $(0, +\infty) = \Theta \subset \mathbb{R}$;
- The Parametric Family of Distributions is $Pois(\lambda)$.

And our problem here will be to estimate our unknown λ, using the realizations $x_1, x_2, ..., x_n$.

Example

Example: Assume we want to model the height of a 20 year old person, X.

Example

Example: Assume we want to model the height of a 20 year old person, X. Of course, X is a r.v.

Example

Example: Assume we want to model the height of a 20 year old person, X. Of course, X is a r.v. An appropriate model for X is

$$X \sim$$

Example

Example: Assume we want to model the height of a 20 year old person, X. Of course, X is a r.v. An appropriate model for X is

$$X \sim \mathcal{N}(\mu, \sigma^2), \qquad \text{for some } \ \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

Example

Example: Assume we want to model the height of a 20 year old person, X. Of course, X is a r.v. An appropriate model for X is

$$X \sim \mathcal{N}(\mu, \sigma^2), \qquad \text{for some} \ \ \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

We will consider a Random Sample (heights of n persons of age 20, but before getting the actual data)

$$X_1, X_2, ..., X_n \sim$$

Example

Example: Assume we want to model the height of a 20 year old person, X. Of course, X is a r.v. An appropriate model for X is

$$X \sim \mathcal{N}(\mu, \sigma^2), \qquad \text{for some} \;\; \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

We will consider a Random Sample (heights of n persons of age 20, but before getting the actual data)

$$X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

and do our analysis based on this Random Sample.

Example

Example: Assume we want to model the height of a 20 year old person, X. Of course, X is a r.v. An appropriate model for X is

$$X \sim \mathcal{N}(\mu, \sigma^2), \qquad \text{for some} \ \ \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

We will consider a Random Sample (heights of n persons of age 20, but before getting the actual data)

$$X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

and do our analysis based on this Random Sample. Then, we will collect data, and obtain a Sample $x_1, x_2, ..., x_n$, a realization of $X_1, X_2, ..., X_n$.

Example

Example: Assume we want to model the height of a 20 year old person, X. Of course, X is a r.v. An appropriate model for X is

$$X \sim \mathcal{N}(\mu, \sigma^2), \qquad \text{for some } \ \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

We will consider a Random Sample (heights of n persons of age 20, but before getting the actual data)

$$X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

and do our analysis based on this Random Sample. Then, we will collect data, and obtain a Sample $x_1, x_2, ..., x_n$, a realization of $X_1, X_2, ..., X_n$. Here

- Our Parameter is $(\mu, \sigma^2) = \theta$ (or $(\mu, \sigma) = \theta$), which is 2D;

Example

Example: Assume we want to model the height of a 20 year old person, X. Of course, X is a r.v. An appropriate model for X is

$$X \sim \mathcal{N}(\mu, \sigma^2), \qquad \text{for some} \quad \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

We will consider a Random Sample (heights of n persons of age 20, but before getting the actual data)

$$X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

and do our analysis based on this Random Sample. Then, we will collect data, and obtain a Sample $x_1, x_2, ..., x_n$, a realization of $X_1, X_2, ..., X_n$. Here

- Our Parameter is $(\mu, \sigma^2) = \theta$ (or $(\mu, \sigma) = \theta$), which is 2D;
- The Parameter Set is $\mathbb{R} \times [0, +\infty) = \Theta \subset \mathbb{R}^2$;

Example

Example: Assume we want to model the height of a 20 year old person, X. Of course, X is a r.v. An appropriate model for X is

$$X \sim \mathcal{N}(\mu, \sigma^2), \qquad \text{for some} \quad \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

We will consider a Random Sample (heights of n persons of age 20, but before getting the actual data)

$$X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

and do our analysis based on this Random Sample. Then, we will collect data, and obtain a Sample $x_1, x_2, ..., x_n$, a realization of $X_1, X_2, ..., X_n$. Here

- Our Parameter is $(\mu, \sigma^2) = \theta$ (or $(\mu, \sigma) = \theta$), which is 2D;
- The Parameter Set is $\mathbb{R} \times [0, +\infty) = \Theta \subset \mathbb{R}^2$;
- The Parametric Family of Distributions is $\mathcal{N}(\mu, \sigma^2)$.

Example

Example: Assume we want to model the height of a 20 year old person, X. Of course, X is a r.v. An appropriate model for X is

$$X \sim \mathcal{N}(\mu, \sigma^2), \qquad \text{for some} \quad \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

We will consider a Random Sample (heights of n persons of age 20, but before getting the actual data)

$$X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

and do our analysis based on this Random Sample. Then, we will collect data, and obtain a Sample $x_1, x_2, ..., x_n$, a realization of $X_1, X_2, ..., X_n$. Here

- Our Parameter is $(\mu, \sigma^2) = \theta$ (or $(\mu, \sigma) = \theta$), which is 2D;
- The Parameter Set is $\mathbb{R} \times [0, +\infty) = \Theta \subset \mathbb{R}^2$;
- The Parametric Family of Distributions is $\mathcal{N}(\mu, \sigma^2)$.

Our Problem here is, using the observation $x_1, x_2, ..., x_n$, to estimate μ and σ^2.

Point Estimates

Motivating Example $\ddot{\smile}$

Motivating Example ⌣̈

Example: I have generated the following Data from a Normal Distribution:

```
##          [,1]  [,2]    [,3]   [,4]     [,5]   [,6]
## [1,] -0.0733 -2.14  -0.366 -1.950 -10.4956 -7.266
## [2,]  0.0756 -3.56  -2.657 -1.824   0.4723  3.393
## [3,] -0.1541 -6.94  -3.666 -0.968  -5.9566  0.123
## [4,] -2.7044 -9.00  -4.847 -0.746  -1.3706 -1.196
## [5,] -6.6370 -6.10 -10.580  2.783  -0.0354  4.852
```

Motivating Example ⌣̈

Example: I have generated the following Data from a Normal Distribution:

```
##          [,1]  [,2]    [,3]   [,4]     [,5]   [,6]
## [1,] -0.0733 -2.14  -0.366 -1.950 -10.4956 -7.266
## [2,]  0.0756 -3.56  -2.657 -1.824   0.4723  3.393
## [3,] -0.1541 -6.94  -3.666 -0.968  -5.9566  0.123
## [4,] -2.7044 -9.00  -4.847 -0.746  -1.3706 -1.196
## [5,] -6.6370 -6.10 -10.580  2.783  -0.0354  4.852
```

Question: Find/Estimate the Parameter values I was using.

Motivating Example $\ddot\smile$

Example: I have generated the following Data from a Normal Distribution:

```
##            [,1]  [,2]    [,3]   [,4]     [,5]   [,6]
## [1,] -0.0733 -2.14  -0.366 -1.950 -10.4956 -7.266
## [2,]  0.0756 -3.56  -2.657 -1.824   0.4723  3.393
## [3,] -0.1541 -6.94  -3.666 -0.968  -5.9566  0.123
## [4,] -2.7044 -9.00  -4.847 -0.746  -1.3706 -1.196
## [5,] -6.6370 -6.10 -10.580  2.783  -0.0354  4.852
```

Question: Find/Estimate the Parameter values I was using.

Moral: Statistics is like a Detective Story: you need to find the Unknown (murderer?) using some (small?) amount of Observations, Data you have $\ddot\smile$

Statistics, Estimator and Estimate

Let us recall what is our Problem: assume we have a Dataset $x_1, ..., x_n$. We assume that this is a realization of a Random Sample $X_1, ..., X_n$, coming from one of the Distributions from some Parametric Family:

$$X_1, X_2, ..., X_n \sim \mathcal{F}_\theta, \qquad \theta \in \Theta.$$

And our problem is to estimate the value of our unknown Parameter θ.

Statistics, Estimator and Estimate

Let us recall what is our Problem: assume we have a Dataset $x_1, ..., x_n$. We assume that this is a realization of a Random Sample $X_1, ..., X_n$, coming from one of the Distributions from some Parametric Family:

$$X_1, X_2, ..., X_n \sim \mathcal{F}_\theta, \qquad \theta \in \Theta.$$

And our problem is to estimate the value of our unknown Parameter θ.

To estimate θ, we will use only $X_1, ..., X_n$ (or, $x_1, x_2, ..., x_n$), since we do not have any other thing.

Statistics, Estimator and Estimate

Let us recall what is our Problem: assume we have a Dataset $x_1, ..., x_n$. We assume that this is a realization of a Random Sample $X_1, ..., X_n$, coming from one of the Distributions from some Parametric Family:

$$X_1, X_2, ..., X_n \sim \mathcal{F}_\theta, \qquad \theta \in \Theta.$$

And our problem is to estimate the value of our unknown Parameter θ.

To estimate θ, we will use only $X_1, ..., X_n$ (or, $x_1, x_2, ..., x_n$), since we do not have any other thing. Now,

Definition: Any (measurable) function of the Random Sample $X_1, X_2, ..., X_n$ is called a **Statistics**. So Statistics is a r.v. of the form

$$g(X_1, X_2, ..., X_n).$$

Statistics, Estimator and Estimate

Let us recall what is our Problem: assume we have a Dataset $x_1, ..., x_n$. We assume that this is a realization of a Random Sample $X_1, ..., X_n$, coming from one of the Distributions from some Parametric Family:

$$X_1, X_2, ..., X_n \sim \mathcal{F}_\theta, \qquad \theta \in \Theta.$$

And our problem is to estimate the value of our unknown Parameter θ.

To estimate θ, we will use only $X_1, ..., X_n$ (or, $x_1, x_2, ..., x_n$), since we do not have any other thing. Now,

Definition: Any (measurable) function of the Random Sample $X_1, X_2, ..., X_n$ is called a **Statistics**. So Statistics is a r.v. of the form

$$g(X_1, X_2, ..., X_n).$$

This is our third meaning of the term *Statistics*.

Example

Example: For example, the followings are Statistics:

$$X_1, \qquad \frac{X_1 + X_n}{3}, \qquad \sin(X_1 \cdot X_2 + X_3 + ... + X_n).$$

Example

Example: For example, the followings are Statistics:

$$X_1, \qquad \frac{X_1 + X_n}{3}, \qquad \sin(X_1 \cdot X_2 + X_3 + ... + X_n).$$

Definition: The Distribution of the Statistics $g(X_1, X_2, ..., X_n)$ is called a **Sampling Distribution**.

Example

Example: For example, the followings are Statistics:

$$X_1, \qquad \frac{X_1 + X_n}{3}, \qquad \sin(X_1 \cdot X_2 + X_3 + ... + X_n).$$

Definition: The Distribution of the Statistics $g(X_1, X_2, ..., X_n)$ is called a **Sampling Distribution**.

Example: The Sampling Distribution of the Statistics

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}$$

is

Example

Example: For example, the followings are Statistics:

$$X_1, \qquad \frac{X_1 + X_n}{3}, \qquad \sin(X_1 \cdot X_2 + X_3 + ... + X_n).$$

Definition: The Distribution of the Statistics $g(X_1, X_2, ..., X_n)$ is called a **Sampling Distribution**.

Example: The Sampling Distribution of the Statistics

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}$$

is almost Normal, for large n, by the CLT.

Statistics, Estimator and Estimate

Assume we want to estimate the value of the parameter $\theta \in \Theta$, and we will use the Statistics $g(X_1, ..., X_n)$ for that.

Statistics, Estimator and Estimate

Assume we want to estimate the value of the parameter $\theta \in \Theta$, and we will use the Statistics $g(X_1, ..., X_n)$ for that.

Definition: If

- $g : \mathbb{R}^n \to \Theta$;
- g doesn't depend on the unknown θ;

then the Statistics $g(X_1, X_2, ..., X_n)$ is called an **Estimator** for θ, and it is usually denoted by

$$\hat{\theta} = \hat{\theta}_n = g(X_1, X_2, ..., X_n).$$