- Inferential Statistics: Parametric Models
- Statistics v3, Estimators
- Properties of Estimators: MSE
- Bias and Unbiasedness
- Asymptotic Unbiasedness

Parametric Statistics: General Problem

One of the general Problems of Statistics is the following: we have a Sample, a Dataset $x: \; x_1, ..., x_n$, and our aim is to get an insight from these numbers, to get an information about the Population, about the *process* generating that Dataset.

We can, of course, calculate the Sample Mean and the Sample Variance of our Dataset. Or, we can plot the Histogram or KDE. But will this give an info about the Population or the process generating the Dataset? Well, no, in general.

Parametric Statistics: Modeling

Again, assume we have a Dataset $x: \ x_1, ..., x_n$. For a Statistical analysis, statisticians assume that this Dataset is a realization of a finite sequence of r.v.s $X_1, X_2, ..., X_n$.

Why is this? Because statisticians think about the Dataset x as being **one of the possible realizations** in the universe of all possible scenarios. Say, if we consider the obervations in the `cars` Dataset:

```
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

then we think about this as a Sample, and we realize that if we would make another Sampling, we'd get other numbers. Even with the same cars!

Parametric Statistics: Modeling

So we will think about our Dataset $x_1, ..., x_n$ as being one possible realization (possible values) of the r.v. Sample $X_1, X_2, ..., X_n$. Then we can think about the sequence $X_1, ..., X_n$ as the **process generating samples** $x_1, ..., x_n$. And we can think about X_1 (or, generic, X) as the process generating numbers in our Dataset.

Example: If we consider the weights (in Kg) of 10 persons:

$$69.5, 77.1, ..., 109,$$

then we make the following model: let X_1 be the weight of the first person (say, the first person we will meet when performing the experiment), X_2 be the weight of the second person,. . . , X_{10} be the weight of the 10-th person. Before making the observations, $X_1, ..., X_{10}$ are random. And, in generic, let X be the r.v. showing the weight of a random person.

Our Dataset of weights is just one of the possible realizations of $X_1, ..., X_{10}$.

Example

Example: Let me make a simulation: say, I want to have a model for the height of a 21 year male person. To that end I will use a Sample of size 6. Instead of randomly asking 6 persons, I will use computer to get that Sample. Now I start from the other end - assume I *already know* the solution: I know beforehand that the heights are Normally Distributed with the mean $155cm$ and variance $30cm^2$ (well, in fact, this is what we want to obtain). Then I can run

```
rnorm(6, mean = 155, sd = sqrt(30))
```

```
## [1] 152.2921 160.6513 157.5400 156.5840 150.1584 147.83
```

This is my Sample. If I will run the code again (in some sense, ask another 6 random persons), I will get, say,

```
rnorm(6, mean = 155, sd = sqrt(30))
```

```
## [1] 157.5197 152.0858 154.2360 164.5723 145.2901 153.82
```

And so on.

Example, Cont'd

Now let's start from the initial point of Parametric Statistics: assume that the heights are Normally Distributed, but I do not know the parameters - the Mean and Variance. But I just have one of the above Samples as an observation of heights. And our task is not to get an information about that specific observation, but *the total process* generating heights, i.e., information about the Distribution of heights.

So, again, having a Dataset $x_1, ..., x_n$, statisticians work with a r.v.s $X_1, X_2, ..., X_n$ to work not only with a particular Sample, but with **all possible samples** from the Distribution (Process) behind the phenomenon.

Parametric Statistics: Modeling

And we give a

Definition: We will call the collection of IID r.v.s $X_1, X_2, ..., X_n$ a **Random Sample** of size n.

In the rest of our story, we will work with Random Samples.

Now, if we have a Random Sample $X_1, ..., X_n$, then, because they are IID, we will have that all X_k-s are coming from the same Distribution:

$$X_1, X_2, ..., X_n \sim \mathcal{F}$$

Our general Problem will be: get an information about $\mathcal{F}$, estimate/recover $\mathcal{F}$.

This problem is very general, and hence, not so much can be said about $\mathcal{F}$. So, we need to impose some conditions about $\mathcal{F}$ to be able to get some more information about it.

Parametric Statistics: Modeling

In Parametric Statistics, we assume that we have a Random Sample

$$X_1, X_2, \dots, X_n \sim \mathcal{F},$$

and $\mathcal{F}$ is a member of the Parametric Familiy of Distributions:

$$\mathcal{F} = \mathcal{F}_\theta, \qquad \theta \in \Theta \subset \mathbb{R}^m.$$

Here

- θ is our Parameter, it is **fixed, but unknown**, it is, in general, m-dimensional;
- $\Theta \subset \mathbb{R}^m$ is our Parameter Set, the set of all possible values of our Parameter θ;
- $\mathcal{F}_\theta$ is our Parametric Distribution.

We will consider one of the main Problems of the Parametric Statistics: **Using the observations from our Random Sample, estimate the value of the Parameter θ.**

Page 38 / 140

Open with Google Docs

Example

Example: Assume we have a coin, and we are tossing it n times, and let $x_1, x_2, ..., x_n$ be the result of that n tosses: $x_k = 1$, of the k-the toss resulted in Heads, and $x_k = 0$ otherwise. Then we can model this experiment as a realization of a Random Sample $X_1, X_2, ..., X_n$, where

$$X_1, X_2, ..., X_n \sim Bernoulli(p), \qquad p \in [0, 1],$$

where p is the Probability of Heads for our coin. In this case,

- our Parameter is $p = \theta$;
- the set $[0, 1] = \Theta \subset \mathbb{R}$ is our Parameters set;
- the Parametric family of Distributions is the family $Bernoulli(p)$, $p \in [0, 1]$.

And a generic $X \sim Bernoulli(p)$ will be the result of our coin toss.

In this problem, p is **fixed, but unknown**. And our aim will be to estimate p, using our observations $x_1, ..., x_n$.

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;
- 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. 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.

Motivating Example $\ddot\smile$

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

```
##         [,1]   [,2]   [,3]   [,4]   [,5]    [,6]
## [1,] -1.03  5.767 -1.645 -6.438  0.737  -0.436
## [2,]  2.45 -0.749 -5.037 -5.257  4.370  -1.238
## [3,] -2.87 -6.943 -3.761 -5.560 -1.515  -5.367
## [4,] -1.94  0.732 -0.563 -5.461 -0.505  -8.492
## [5,] -4.98 -2.744  3.812 -0.785  4.176 -15.160
```

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.

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:

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

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.

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

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

The value of the Estimator at our observations, $g(x_1, x_2, ..., x_n)$, is called an **Estimate** for θ, and it is again (unfortunately) denoted by $\hat{\theta} = \hat{\theta}_n$.

Example

Example: Say, we want to estimate the parameter λ in the

$$\{Exp(\lambda): \ \lambda > 0\}$$

model, using the Random Sample

$$X_1, X_2, ..., X_n \overset{IID}{\sim} Exp(\lambda).$$

Then the followings are **Esimators**:

- $\hat{\lambda} = \dfrac{X_1}{3}$;
- $\hat{\lambda} = \dfrac{X_1 + X_2 + ... + X_n}{n}$;
- $\hat{\lambda} = (X_1 + X_2 + ... + X_n)^2$;

And the following is not an estimator:

- $\hat{\lambda} = \dfrac{\lambda}{X_1 + X_n}$, since it depends on λ - the unkown parameter value.

Estimators and Estimates

Note: We require our Estimator to be independent of the Parameter θ, since we want to be able to calculate the value of the Estimator on the observation, i.e., we want to be able to calculate the Estimate, we want to make our Estimator *observable*. Since θ is unknown, then we cannot use it in the Estimator - we will not be able to calculate the Estimate.

Note: Again,

- **Estimator** is a function of our Random Sample, it is a r.v.. It doesn't depend on the unknown Parameter. Plugging the values of our observation into the Estimator we will get the Estimate
- **Estimate** is a number, it is the result of plugging the observation into the Estimator.

Example

Example: Assume we want to estimate the probability of a birth (birthrate) of a girl child in Armenia. To that end, say, we ask a maternity hospital to provide the birth data for a day. And say, our data is:

$$b, g, g, b, b, g, b$$

First we need to build a Model. To that end, we encode our data by

$$x: \;\; 0, 1, 1, 0, 0, 1, 0,$$

where $b = 0$ and $g = 1$: this is to be able to use one of our standard Distributions. Next, from a Dataset we pass, for a generalization, to a Random Sample

$$X_1, X_2, X_3, X_4, X_5, X_6, X_7,$$

where X_k is the gender of the k-th child *before the observation was made* ($X_k = 1$ if the child will be a girl, and 0 otherwise).

Example, cont'd

Then we will have

$$X_1, X_2, X_3, X_4, X_5, X_6, X_7 \sim Bernoulli(p),$$

where p is the Probability of having a girl child. So our Model will be

$$\{Bernoulli(p) : p \in [0, 1]\},$$

and our task will be to estimate p based on our Sample x.

To estimate p, let us take the following **Estimator**:

$$\hat{p} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7}{7}.$$

This is a r.v. . The **Estimate** for p, using our Observation, will be

$$\hat{p} = \frac{0 + 1 + 1 + 0 + 0 + 1 + 0}{7} = \frac{3}{7}.$$

Point Estimation Setup

In the Inferential Statistics part, for the Point Estimation of the Parameter(s), we usually do the following steps, after describing the problem (and, maybe after obtaining Data):

- First we specify the Model and unknown Parameter(s);
- Then we find Estimators for our Parameter(s);
- Prove that the Estimators found are good ones;
- Using Estimators and Data, calculate the Estimate for the Parameter(s).

In the next few lectures, we will consider what it means that an Estimator is a good one. Later, we will consider some general methods to find good Estimators.

Example

Example: Assume we work with the Bernoulli Model: we have a Random Sample

$$X_1, X_2, ..., X_n \sim Bernoulli(p),$$

and we want to estimate the Parameter p.

Question: Which Estimator to use? Say, is

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

good enough to estimate the unknown p?

Or, maybe,

$$\hat{p} = \frac{X_1 + X_n}{2}$$

is better?

Or, maybe

$$\hat{p} = \frac{X_{(1)} + X_{(n)}}{2} \qquad or \qquad \hat{p} = Median(X_1, ..., X_n)?$$

Example

Example: Assume we work with the Gaussian Model: we have a Random Sample

$$X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2),$$

and we want to estimate the Parameters μ and σ^2.

Question: Which Estimator to use? Say, is

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

good enough to estimate the unknown μ?

And what about estimating σ^2? Can you suggest Estimators? Say, which one to choose:

$$\widehat{\sigma^2} = \left(\frac{\sum_{k=1}^n |X_k - \overline{X}_n|}{n}\right)^2 \quad \textit{or} \quad \widehat{\sigma^2} = \frac{\sum_{k=1}^n (X_k - \overline{X}_n)^2}{n} \quad \textit{or}$$

$$\widehat{\sigma^2} = \frac{\sum_{k=1}^n (X_k - \overline{X}_n)^2}{n-1} \quad \textit{or} \quad \widehat{\sigma^2} = \text{other Estimator?}$$

Example

Example: Assume we work with the Exponential Model: we have a Random Sample

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

and we want to estimate the Parameter λ.

Question: Which Estimator to use? Say, is

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

good enough to estimate the unknown λ?

Notation

Before starting, few Notations. Let us consider the Model

$$\{\mathcal{F}_\theta : \theta \in \Theta\}.$$

Assume $F_\theta(x) = F(x|\theta)$ is the CDF of the Distribution $\mathcal{F}_\theta$ and $f_\theta(x) = f(x|\theta)$ is the PD(M)F of $\mathcal{F}_\theta$.

If $X \sim \mathcal{F}_\theta$, then we will write

$$\mathbb{E}_\theta(X)$$

the Expected value of X: since $X \sim \mathcal{F}_\theta$, then the Expected value depends on θ. Say, if $X \sim \mathcal{F}_\theta$ is continuous, then

$$\mathbb{E}_\theta(X) = \int_{-\infty}^{+\infty} x \cdot f(x|\theta)dx.$$

And we will use $Var_\theta(X)$ for the Variance of X.

Risk, Mean Squared Error of the Estimator

Assume we have a Random Sample

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

and we use the Estimator $\hat{\theta}$ to estimate θ.

Definition: The **Mean Squared Error** or the **Quadratic Risk** of the estimator $\hat{\theta}$ of θ is

$$MSE(\hat{\theta}, \theta) = Risk(\hat{\theta}, \theta) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2].$$

Note: MSE calculates how close are, in the Quadratic Mean sense, possible values of the Estimator $\hat{\theta}$ to the actual (unknown) value of θ. The smaller the value of MSE, the better, of course.

Definition: We say that the estimator $\hat{\theta}^1$ of θ is **preferable** to $\hat{\theta}^2$, another estimator of θ, if

$$MSE(\hat{\theta}^1, \theta) \leq MSE(\hat{\theta}^2, \theta), \qquad \forall \theta \in \Theta,$$

and there exists a θ s.t. $MSE(\hat{\theta}^1, \theta) < MSE(\hat{\theta}^2, \theta)$.

Useful Identity

A very useful identity for calculating $\mathbb{E}(X^2)$ for a r.v. X is the following: because

$$Var(X) = \mathbb{E}(X^2) - \left[\mathbb{E}(X)\right]^2,$$

then

$$\mathbb{E}(X^2) = Var(X) + \left[\mathbb{E}(X)\right]^2.$$

Example

Example: Assume we have a Random Sample

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

and we want to Estimate the unknown λ. We want to use one of the following Estimators:

$$\hat{\lambda}_1 = X_1 \qquad or \qquad \hat{\lambda}_2 = \frac{X_1 + X_n}{2}.$$

Question: Calculate the Risks of $\hat{\lambda}_1$ and $\hat{\lambda}_2$. Compare these two Estimators: which of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ is preferable?

Solution: OTB: Use the fact that for $X \sim Pois(\lambda)$,

$$\mathbb{E}(X) = \lambda = Var(X).$$