Point estimators

Chrysafis Vogiatzis

Lecture 15-16

Learning objectives

After these lectures, we will be able to:

- Describe point estimation.
- Explain the difference between bias and variance of a point estimator.
- Evaluate the bias, variance, mean square error of a point estimator.
- Compare two point estimators and pick the better one.

Motivation: Inferring parameters

In most applications, we have a good enough idea on the distribution that we need to follow. When a company makes vehicles, we could pick some of them to check for their quality and count how many are of high quality: this can be modeled as a binomial or a hypergeometric distribution. When a student takes an exam, they will expect to do similarly to their previous exams plus or minus some points if they prepare in a similar manner: their score can be modeled as a normal distribution.

However, one of the questions we need to answer is: what are the parameters of the distributions? What is the mean and variance of that normal distribution? What is p in a binomial distribution? So far, we have been given this as part of our data. What happens when we are given data and need to infer their values, based on real-life observations?

Motivation: Predicting an election

Before an election takes place, we see many polls. Some of them appear to better resemble the final result (after the election); others fail to capture reality. Given this data based on a *sample* of the whole *population*, what can we say about the election? What can we say about the probabilities of one candidate versus another?

Statistical inference

Statistical inference takes us from the sample to the whole. For example, consider any of the next scenarios:

- We interviewed 50 people about the next election. What do the results imply for the general election?
- We picked a sample of 10 cars and performed a crash test. What do the observations imply for the whole production line?
- We collected exit interview data from 100 alumni. What do their answers imply for the starting salary of our alumni?

Right away, we can make some intuitive observations:

1. Checking a sample, rather than the whole, saves us time and effort.
2. Checking a sample, rather than the whole, comes with a loss of information.
3. Checking a sample, rather than the whole, we want to recreate the whole.

Statistical inference theme

The general theme for this part of the class can be summarized pretty well in the following Figure 1.

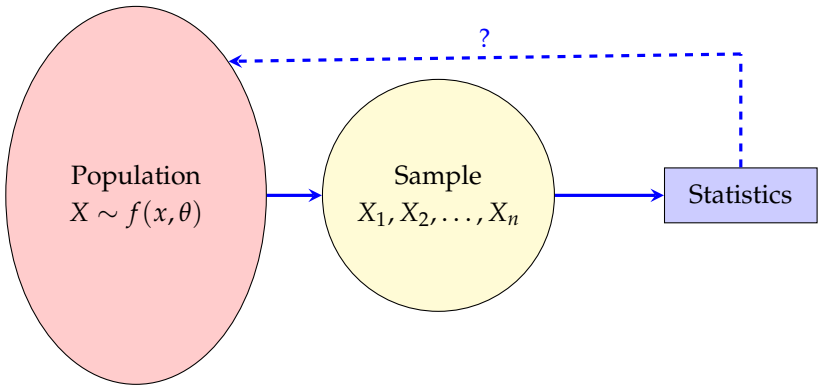


Figure 1: The theme of statistical inference. We collect a **sample** of the whole **population** that we analyze to get some **statistic**, which we then use to infer information about the population.

Sample averages and variances

Assume a large population X: you decide to collect only 5 random variables X_1, X_2, X_3, X_4, X_5. Then, we may calculate the sample average $\frac{X_1+X_2+X_3+X_4+X_5}{5}$ and the sample variance and use those in lieu of the population mean (unknown) and the population variance (unknown).

For example, say I want to figure out the average height of every Chicago resident, I could (i) travel to Chicago, (ii) ask 10 people about their height, and (iii) calculate the average of these 10 people. Is this the true average?

Statistics

We proceed with some preliminaries that will be used throughout the next few classes.

Definition 1 (Statistic) *A statistic is any value obtained by random data.*

Well, this is not very useful. The only recurring theme here is *random* data. In essence, the definition claims that any value that is different for differently obtained data can be considered a statistic!

Height statistics

Assume that the heights of the 5 people in the leadership team of a student chapter are: 60, 67, 72, 63, 60. Then, the height of the second person picked (67) is a statistic. The average height is another statistic. Finally, getting the height of the first person and multiplying it by 3 and adding to it the height of the last person is *also* a statistic.

Clearly, some statistics are more useful than others. For example, in our earlier discussion the average is more useful than the last statistic. To verify, answer the following questions.

Are the following statistics? True or False.

- The sample variance.

  a) True b) False

- The value of the first element of a sample.

  a) True b) False

- The value of the first element of a sample minus 3.

  a) True b) False

Some basic properties of statistics:

1. **Statistics depend on the sample selected**: the value a statistic gets will be different depending on the sample selected. If I select 5 students in the class and report their average exam score (a statistic), it will be different depending on the 5 students selected.
2. **Statistics are functions of the sample selected**: we can use the same "formula" or follow the same "approach" to estimate the value of a statistic given a sample, no matter the sample.
3. **Statistics are random variables**: statistics are distributed as random variables. Certain values may appear more often than others, we can define expectations for statistics, etc.

Polling with few people

For a poll, we are asked to find 10 people and ask them a question on a Likert scale (that is from 1 to 5). Then, we take their answers and add them up: we say that if the score is ≥ 30 then the people agree with that statement on average.

Unfortunately you were only able to find 5 people, who provided the following answers X_i: 2, 1, 3, 3, 3. You then decide to use $Y = 2 \cdot \sum X_i$ as the statistic you report back. In this case, you'd report $Y = 2 \cdot (2+1+3+3+3) = 24$.

- **Is this a statistic?** Yes.
- **Does it depend on the sample selected?** Yes. Change the sample asked to obtain a different number.
- **Is it a function of the sample selected?** Yes. We always add up the answers and multiply by 2.
- **Is this a random variable?** Yes. We can calculate an expectation and a variance, and we can estimate probabilities!

Sampling distribution

So, if a statistic is a random variable, what is its distribution? The distribution of a statistic, called the **sampling distribution** depends on three things:

1. The distribution of the whole population. Of course we should expect that the distribution of the population will be reflected when looking at a sample!
2. The size of the sample. Once again, it should make sense that the bigger the sample we pick the more accurately we will reflect the population distribution.
3. The way the sample was selected. We will not devote a lot of time in this: but, picking a sample in a non-random way will affect the distribution we see.

   Confused? Don't be! We have done that already..

Back to the normal distribution

Assume you have a population where each individual is distributed following a normal distribution with mean μ and variance σ^2. You pick, at random, a set of n individuals X_i. What is the average distributed as?

We have seen that the average $Y = \frac{\sum X_i}{n}$ is also normally distributed with the same mean μ and variance $\frac{\sigma^2}{n}$. This normal distribution $\mathcal{N}(\mu, \sigma^2/n)$ is the sampling distribution.

Point estimators

Let's put everything formally. Let X be a population distributed with some pdf $f(x, \theta)$, where θ is some unknown parameter. By the way, you may treat θ as a vector of multiple parameters. [1]

Furthermore, let $X_1, X_2, \ldots, X_n$ be a series of random elements picked from the population. They form the sample of size n that was selected. By definition, seeing as $X_1, X_2, \ldots, X_n$ all come from the same place, they are identically distributed and independent random variables.

[1] Recall that every distribution we have seen had some parameters that were required to define it. For example, the binomial distribution needed $n > 0$ and $p \geq 0$, whereas the exponential distribution or the Poisson distribution needed $\lambda > 0$.

Definition 2 (Point estimators) *We define point estimator(s) $\hat{\Theta}$ as a statistic that is used to approximate the unknown parameter(s) θ.*

By definition, $\hat{\Theta}$ is a function of the sample selected $(X_1, X_2, \ldots, X_n)$, hence we may say that $\hat{\Theta}$ is a random variable depending on the sample ($\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$, where $h(\cdot)$ is some function).

Of course, once we have picked a sample then $\hat{\Theta}$ can be calculated and assigned a value. This value is called the **point estimate** $\hat{\theta}$. To summarize, $\hat{\Theta}$ is the general statistic used (e.g., the point estimator can be found if we take the average and add 2) whereas $\hat{\theta}$ is the value the estimator receives for a specific sample (e.g., for our sample, the average is 7 so the point estimate is 9). To summarize, $\hat{\Theta}$ is typically a "formula" or an "expression", whereas $\hat{\theta}$ is an actual number.

Common point estimators

Such a topic (statistical inference) is so broad and useful that we definitely already have some estimators that are typically used. Are you looking for the (unknown) mean of a population? Collect a sample and report its average. Are you looking for the (unknown) population variance? Collect a sample and report its sample variance. We differentiate between single and two populations.

Single population. For a single population:

Parameters	Point estimators
Population mean μ	Sample average $\hat{\Theta} = \overline{x}$
Population variance σ^2	Sample variance $\hat{\Theta} = s^2$
Population proportion p	Sample proportion $\frac{\hat{n}}{n}$

Single population proportions

Assume a population that we want to ask whether they agree or disagree with a new policy. Should we enact it? If it is difficult or impossible to collect feedback from all, we may pick a sample and ask them if they agree or not. Let n be the sample size and $\hat{n}$ be the number of people who agree.

Finally, we may report that $\frac{\hat{n}}{n}$ is the point estimator for the unknown proportion.

Two populations. For two populations:

Parameters	Point estimators
Difference in population means $\mu_1 - \mu_2$	Difference in sample averages $\hat{\Theta} = \overline{x}_1 - \overline{x}_2$
Ratio in population variances $\frac{\sigma_1^2}{\sigma_2^2}$	Ratio in sample variance $\frac{s_1^2}{s_2^2}$
Difference in population proportions $p_1 - p_2$	Difference in sample proportions $\hat{\Theta} = \frac{\hat{n}_1}{n_1} - \frac{\hat{n}_2}{n_2}$

Two population proportions

Assume a population that we want to ask whether they agree or disagree with a new policy. However, we are also aware of the existence of two populations: say, for example, people who make more that $100,000 and people who make less. Is the policy more preferred to people of one category versus the other? Let n_1 be the sample size from the first population and $\hat{n}_1$ be the number of people who agree from that population. Similarly, define n_2 and $\hat{n}_2$.

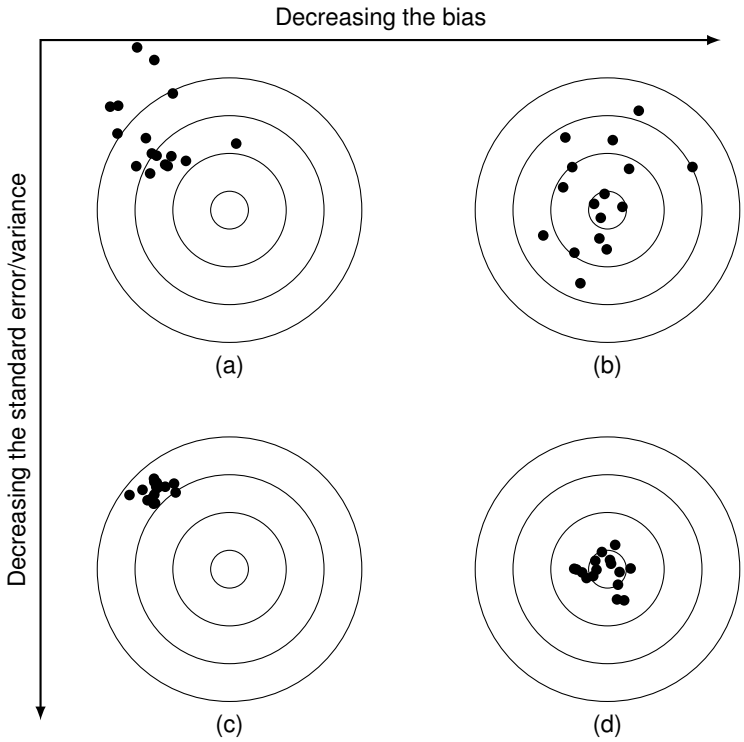
Finally, we may report that the difference between the two proportions is $\frac{\hat{n}_1}{n_1} - \frac{\hat{n}_2}{n_2}$ as the point estimator for the unknown proportion difference.

Before moving to the next section, take a moment to summarize what we have seen. We have a population distributed with some

probability density function ($f(x)$) but with unknown parameters θ. We then come up with a plan: select a sample, and calculate a point estimator $\hat{\Theta}$. For a given sample, you obtain a point estimate (actual value) $\hat{\theta}$. You use that to estimate the unknown parameter. Additionally, recall that $\hat{\Theta}$ is a random variable, so it should come as no surprise that we can analyze it as such!

What makes a good estimator?

Every estimator has two main items we want to evaluate it by: **accuracy** and **precision**. In statistics terms, we refer to them as **bias** and **standard error** or **variance**. We present the effect that each of the two would have to our estimation process: decreasing the bias would lead to better results *on average*; decreasing the standard error/variance would lead to smaller dispersion.



A "good" estimator should have zero bias and zero variance! However, this is practically impossible: hence, we settle for zero bias and minimum variance. Let us proceed with the definitions.

Definition 3 (Bias) *We define the **bias** of a point estimator as the difference between its expectation and the parameter itself.*

$$bias\left[\hat{\Theta}\right] = E\left[\hat{\Theta}\right] - \theta.$$

An estimator with zero bias is referred to as unbiased.

Bias example

Assume a population with mean μ and variance σ^2. As the mean is unknown you decide to use the following three approaches to estimate it:

1. Get the average from a sample of 3 randomly picked observations.
2. Get a sample of 3 randomly picked observations and calculate $\frac{2 \cdot X_1 + X_2 - X_3}{2}$.
3. Get a sample of 3 randomly picked observations and calculate $2X_1 + X_2 - X_3$.

What are the biases of each of the three point estimators?

1. $\hat{\Theta}_1 = \frac{X_1 + X_2 + X_3}{3}$:

$$E\left[\hat{\Theta}_1\right] = E\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{3}\left(\underbrace{E\left[X_1\right]}_{\mu} + \underbrace{E\left[X_2\right]}_{\mu} + \underbrace{E\left[X_3\right]}_{\mu}\right) =$$
$$= \frac{1}{3}\left(\mu + \mu + \mu\right) = \mu \implies$$
$$\implies bias(\hat{\Theta}_1) = 0.$$

2. $\hat{\Theta}_2 = \frac{2 \cdot X_1 + X_2 - X_3}{2}$:

$$E\left[\hat{\Theta}_2\right] = E\left[\frac{2 \cdot X_1 + X_2 - X_3}{2}\right] = \frac{2\mu + \mu - \mu}{2} = \mu \implies$$
$$\implies bias\left(\hat{\Theta}_2\right) = 0.$$

3. $\hat{\Theta}_3 = 2 \cdot X_1 + X_2 - X_3$:

$$E\left[\hat{\Theta}\right] = E\left[2 \cdot X_1 + X_2 - X_3\right] = 2\mu + \mu - \mu = 2\mu \implies$$
$$\implies bias\left(\hat{\Theta}_3\right) = \mu.$$

So, the first two estimators will be unbiased (zero bias)! The last one is biased and its bias is as big as the unknown mean.

Definition 4 (Standard error and variance) *We define the **standard error** of a point estimator as the square root of its **variance**.*

$$SE\left[\hat{\Theta}\right] = \sqrt{Var\left[\hat{\Theta}\right]}.$$

We want this to be minimum. A point estimator with minimum variance and zero bias is called a minimum variance unbiased estimator.

Variances example

Assume the same population with unknown mean μ and variance σ^2. We use again the three estimators from before (referred to as $\hat{\Theta}_1, \hat{\Theta}_2, \hat{\Theta}_3$). What are the variances of each of the three point estimators?

1. $\hat{\Theta}_1 = \frac{X_1+X_2+X_3}{3}$:

$$Var\left[\hat{\Theta}_1\right] = Var\left[\frac{X_1 + X_2 + X_3}{3}\right] =$$
$$= \frac{1}{9}\left(\underbrace{Var\left[X_1\right]}_{\sigma^2} + \underbrace{Var\left[X_2\right]}_{\sigma^2} + \underbrace{Var\left[X_3\right]}_{\sigma^2}\right) =$$
$$= \frac{1}{9}3\sigma^2 = \frac{\sigma^2}{3}.$$

2. $\hat{\Theta}_2 = \frac{2\cdot X_1+X_2-X_3}{2}$:

$$Var\left[\hat{\Theta}_2\right] = Var\left[\frac{2\cdot X_1 + X_2 - X_3}{2}\right] =$$
$$= Var\left[X_1\right] + \frac{1}{4}Var\left[X_2\right] + \frac{1}{4}Var\left[X_3\right] =$$
$$= \sigma^2 + \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 \implies Var\left[\hat{\Theta}_2\right] = \frac{3}{2}\sigma^2.$$

3. $\hat{\Theta}_3 = 2\cdot X_1 + X_2 - X_3$:

$$Var\left[\hat{\Theta}_3\right] = Var\left[2\cdot X_1 + X_2 - X_3\right] =$$
$$= 4\sigma^2 + \sigma^2 + \sigma^2 \implies Var\left[\hat{\Theta}_3\right] = 6\sigma^2.$$

Comparing, the first estimator has a significantly smaller variance than the other two. Among the three options, $\hat{\Theta}_1$ is the minimum variance unbiased estimator.

Definition 5 (Mean square error) *We define the **mean square error** of a point estimator as the expected value of the square error $\left(\hat{\Theta} - \theta\right)^2$:*

$$MSE = E\left[\left(\hat{\Theta} - \theta\right)^2\right].$$

We can use this to derive the fact that the mean square error is equal to

the summation of the variance plus the square of the bias:

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

By definition, the MSE tries to capture both bias and variance at the same time. Hence, we typically say that one estimator is better than another if its MSE is smaller. We may also define the **relative efficiency** as the ratio of two estimator mean square errors:

$$\text{Relative efficiency} = \frac{MSE\left(\hat{\Theta}_1\right)}{MSE\left(\hat{\Theta}_2\right)}.$$

If the relative efficiency is less than 1, then we say that point estimator $\hat{\Theta}_1$ is preferred to point estimator $\hat{\Theta}_2$.

Mean square errors example

Assume the same population with unknown mean μ and variance σ^2. We use for one last time the three estimators $\hat{\Theta}_1, \hat{\Theta}_2$, and $\hat{\Theta}_3$. What are the mean square errors of each of the three point estimators? Which one would we prefer? What are the relative efficiencies of $\hat{\Theta}_1, \hat{\Theta}_2$, $\hat{\Theta}_1, \hat{\Theta}_3$, $\hat{\Theta}_2, \hat{\Theta}_3$?

1. $\hat{\Theta}_1 = \frac{X_1+X_2+X_3}{3}$: $MSE(\hat{\Theta}_1) = \frac{\sigma^2}{3} + 0 = \frac{\sigma^2}{3}$.
2. $\hat{\Theta}_2 = \frac{2\cdot X_1+X_2-X_3}{2}$: $MSE(\hat{\Theta}_2) = \frac{3\sigma^2}{2} + 0 = \frac{3\sigma^2}{2}$.
3. $\hat{\Theta}_3 = 2\cdot X_1 + X_2 - X_3$: $MSE(\hat{\Theta}_3) = 6\sigma^2 + \mu^2$.

$\hat{\Theta}_1$ has the smallest MSE (as expected), followed by $\hat{\Theta}_2$. The relative efficiencies can be found as:

1. $\hat{\Theta}_1, \hat{\Theta}_2$: $\frac{MSE\left(\hat{\Theta}_1\right)}{MSE\left(\hat{\Theta}_2\right)} = \frac{\frac{\sigma^2}{3}}{\frac{3\sigma^2}{2}} = \frac{2}{9} < 1$, so $\hat{\Theta}_1$ is preferred.
2. $\hat{\Theta}_1, \hat{\Theta}_3$: $\frac{MSE\left(\hat{\Theta}_1\right)}{MSE\left(\hat{\Theta}_3\right)} = \frac{\frac{\sigma^2}{3}}{6\sigma^2+\mu^2} < 1$, so $\hat{\Theta}_1$ is preferred.
3. $\hat{\Theta}_2, \hat{\Theta}_3$: $\frac{MSE\left(\hat{\Theta}_2\right)}{MSE\left(\hat{\Theta}_3\right)} = \frac{\frac{3\sigma^2}{2}}{6\sigma^2+\mu^2} < 1$, so $\hat{\Theta}_2$ is preferred.

Review

Let us review very quickly the notions we have seen in this lecture:

- **Population**: X, where each element in the population is distributed with the same distribution (assume pdf $f(x)$) and with potentially unknown parameter(s) θ.
- **Random sample**: $X_1, X_2, \ldots, X_n$ each independent and from the same population with mean μ and variance σ^2.
  - $E[X_i] = E[X] = \mu$.
  - $Var[X_i] = Var[X] = \sigma^2$.
- **Statistic**: any function of a random variable.
- **Sampling distribution**: the distribution of a statistic.
- **Parameter**: (potentially unknown) information necessary to fully define the distribution of the population.
- **Point estimator** $\hat{\Theta}$: a statistic to estimate or approximate an unknown parameter θ.
- **Bias**: $E\left[\hat{\Theta}\right] - \theta$. we want this to be zero.
- **Standard error**: $\sqrt{Var\left[\hat{\Theta}\right]}$. we want this to be small.
- **Minimum variance unbiased estimator**: an estimator $\hat{\Theta}$ with zero bias and minimum variance.