Data Valorization:
Point Estimation

Lionel Fillatre

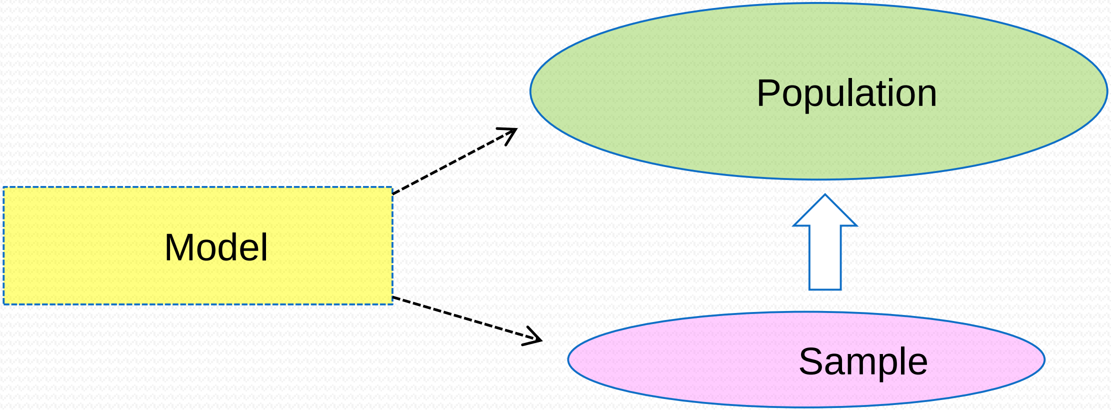
fillatre@unice.fr

Topics

- Statistical Inference
- Point Estimation
- Likelihood Method
- Quality of Estimation
- Cramer-Rao Bound
- Conclusion

1 Statistical Inference

Statistical inference in general



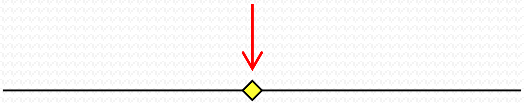
Conclusions about the population is drawn from the sample with assistance from a specified model

Estimation

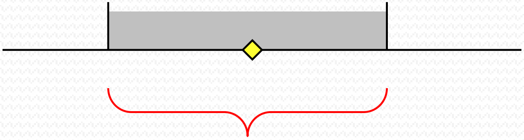
- The objective of estimation is to determine the approximate value of a population parameter on the basis of a sample statistic.

- There are two types of estimators:

  - Point Estimator

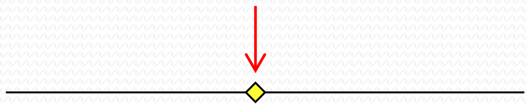


  - Interval Estimator



Point Estimator

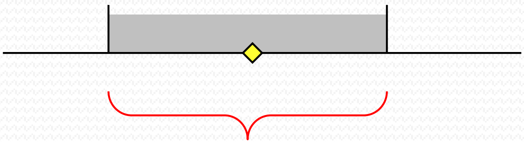
- A ***point estimator*** draws inferences about a population by estimating the value of an unknown parameter using a single value or point.



- Point probabilities in continuous distributions are virtually zero. Likewise, we expect that the point estimator gets closer to the parameter value with an increased sample size.

Interval Estimator

- An ***interval estimator*** draws inferences about a population by estimating the value of an unknown parameter using an interval.



- That is we say (with some ___% certainty) that the population parameter of interest is between some lower and upper bounds.

Point and Interval Estimation

- For example, suppose we want to estimate the mean summer income of a class of $n = 25$ business students

- Point estimate: $\bar{x}$ is calculated to be 300 €/week.

- Interval estimate: the mean income is between 280 and 320 €/week.

2 Point Estimation

The univariate population/sample model

- The population to be investigated is such that the values that comes out in a sample $x_1, x_2, \ldots$ are governed by a probability distribution

- The probability distribution is represented by a probability density (or mass) function $f(x)$

- The sample values can be seen as the outcomes of independent random variables $X_1, X_2, \ldots$ all with probability density (or mass) function $f(x)$

$$f(x) \xrightarrow{\text{governs}} X_1, X_2, X_3, \ldots \xrightarrow{\text{realized as}} x_1, x_2, x_3, \ldots$$

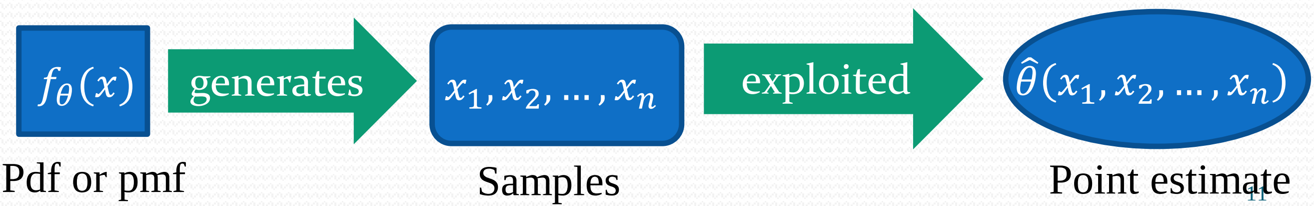
Pdf or pmf | Random variables | Samples

Point estimation (frequentistic paradigm)

- We have a sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ from a population
- The population depends on an unknown parameter θ
- The probability density or mass function of the distribution is known but it depends on the unknown θ, denoted by $f(x;\theta)$ or $f_\theta(x)$
- A point estimate of θ is a function of the sample values

$$\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n) = \hat{\theta}(\boldsymbol{x})$$

such that its values should be close to the unknown θ.



“Standard” point estimates

- The sample mean $\overline{\boldsymbol{x}}$ is a point estimate of the population mean μ

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}(x_1, \dots, x_n)$$

- The sample variance s^2 is a point estimate of the population variance σ^2

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{\boldsymbol{x}})^2 = \hat{\sigma}^2(x_1, \dots, x_n)$$

- The sample proportion p of a specific event (success/failure, positive/negative, etc.) is a point estimate of the corresponding population proportion π

$$p = \frac{\#\{x_i \colon \text{event is satisfied}\}}{n} = \hat{\pi}(x_1, \dots, x_n)$$

Assessing a point estimate

- A point estimate has a sampling distribution:
  - Replace the sample observations $x_1, \dots, x_n$ with their corresponding random variables $X_1, \dots, X_n$ in the functional expression:

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$$

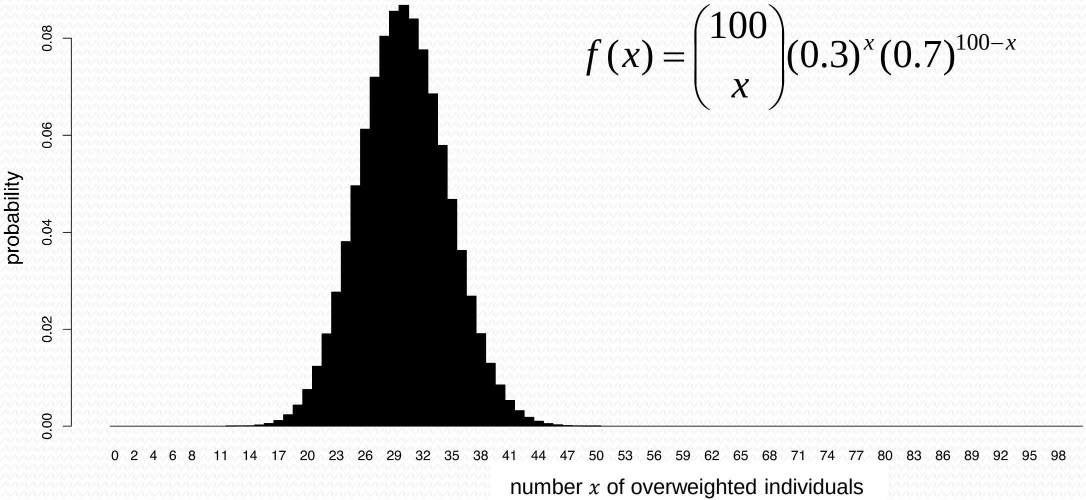
  - **The point estimate is the realization of a random variable (point estimator) that is observed in the sample**
  - As a random variable, the point estimator must have a probability distribution than can be deduced from $f(x; \theta)$
- The point estimator/estimate is assessed by investigating its sampling distribution, in particular the mean and the variance.

3 Likelihood Method

Example: binomial distribution

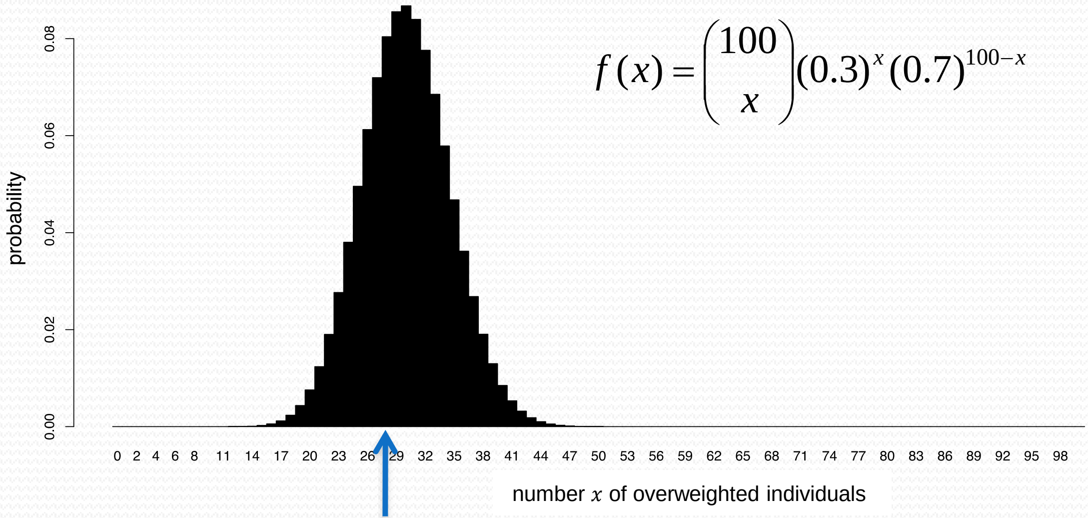
In a population of 1,000,000 people with a true prevalence of 30%, the probability distribution of number x of overweighted individuals if 100 individuals are sample is

$$f(x) = \binom{100}{x}(0.3)^x(0.7)^{100-x}$$



barplot(dbinom(x = 0:100, size = 100, prob = .3), names.arg = 0:100)

Example: binomial distribution

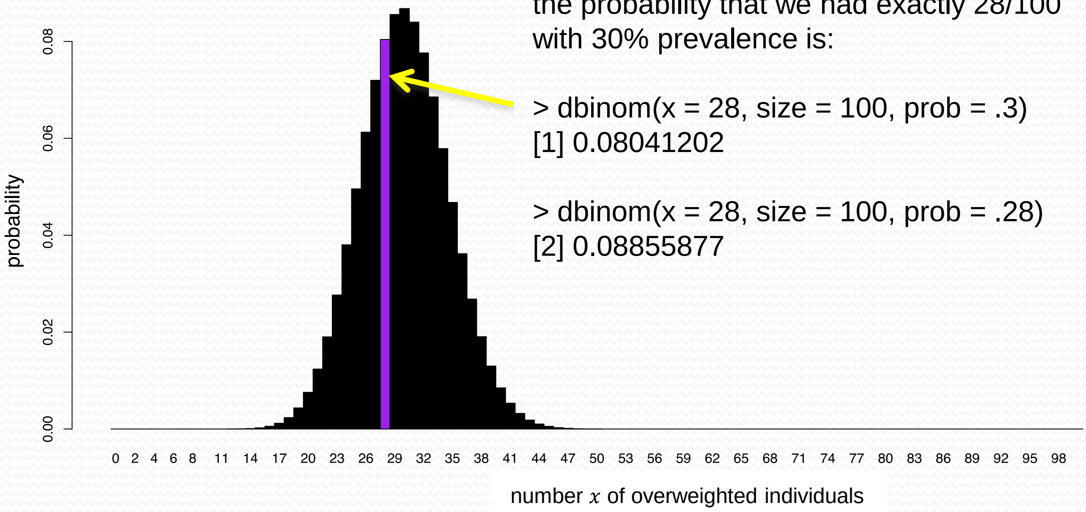


$$f(x) = \binom{100}{x} (0.3)^x (0.7)^{100-x}$$

We sample 100 people once and 28 are positive:

```
> rbinom(n = 1, size = 100, prob = .3)
[1] 28
```

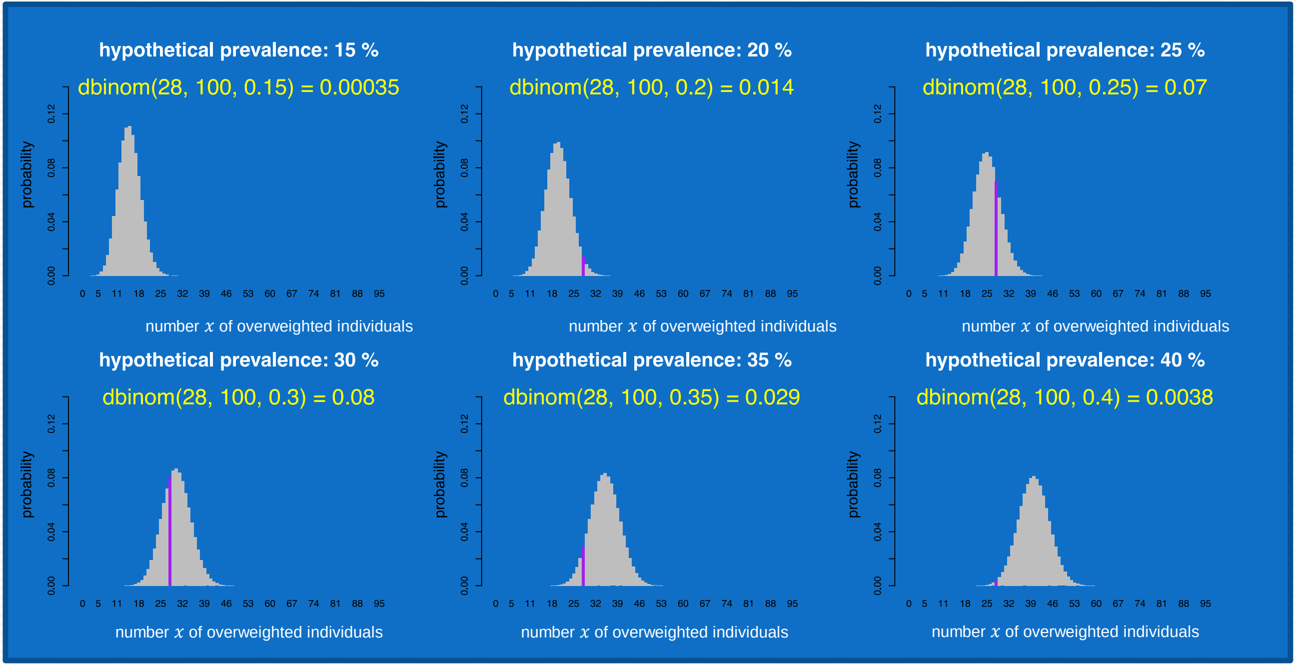
How to estimate the true prevalence?



We don't know the true prevalence, but the probability that we had exactly 28/100 with 30% prevalence is:

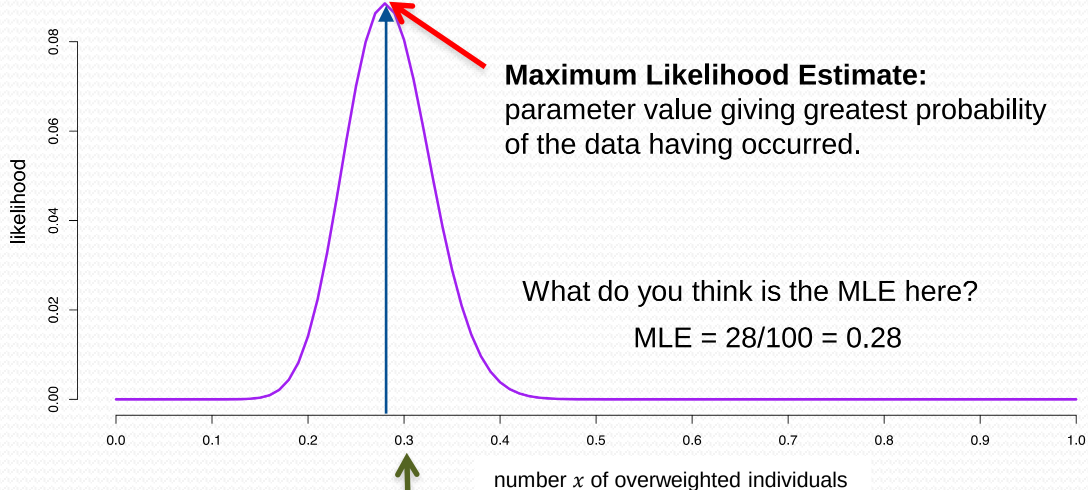
```
> dbinom(x = 28, size = 100, prob = .3)
[1] 0.08041202

> dbinom(x = 28, size = 100, prob = .28)
[2] 0.08855877
```

Which is most likely given our data?



Informal definition

P(our data given the prevalence)=likelihood



Maximum Likelihood Estimate: parameter value giving greatest probability of the data having occurred.

What do you think is the MLE here?

MLE = 28/100 = 0.28

true unknown value = 0.30

Defining likelihood

- L(parameter | data) = p(data | parameter)

function of x

PDF: $f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}$

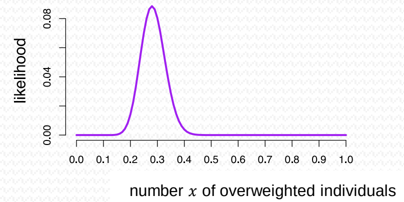
LIKELIHOOD: $L(p \mid x) = \binom{n}{x} p^x (1-p)^{n-x}$

function of p

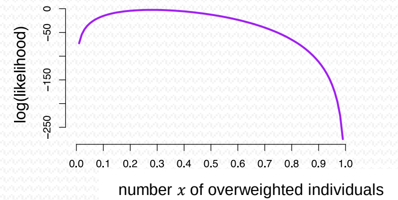
- The likelihood function is a function of the unknown parameter (the samples are fixed and known)
- Not a probability distribution of the parameter p!
- It measures all of the evidence in a sample relevant to p

Deriving the Maximum Likelihood Estimate

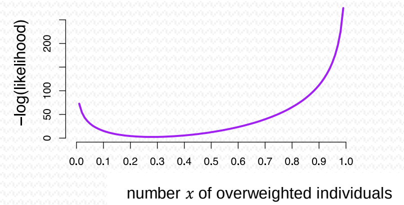
$$\overbrace{L(p)}^{\text{maximize}} = \binom{n}{x} p^{x}(1-p)^{n-x}$$



$$l(p) = \overbrace{\log(L(p)}^{\text{maximize}} = \log\left[\binom{n}{x} p^{x}(1-p)^{n-x}\right]$$



$$\overbrace{-l(p)}^{\text{minimize}} = -\log\left[\binom{n}{x} p^{x}(1-p)^{n-x}\right]$$



Deriving the Maximum Likelihood Estimate

$$-l(p) = -\log(L(p)) = -\log\left[\binom{n}{x} p^x (1-p)^{n-x}\right]$$

$$-l(p) = -\log\binom{n}{x} - \log(p^x) - \log((1-p)^{n-x})$$

$$-l(p) = -\log\binom{n}{x} - x\log(p) - (n-x)\log(1-p)$$

Deriving the Maximum Likelihood Estimate

$$-l(p) = -\log\binom{n}{x} - x\log(p) - (n-x)\log(1-p)$$

$$-\frac{dl(p)}{dp} = 0 - \frac{x}{p} - \frac{-(n-x)}{1-p}$$

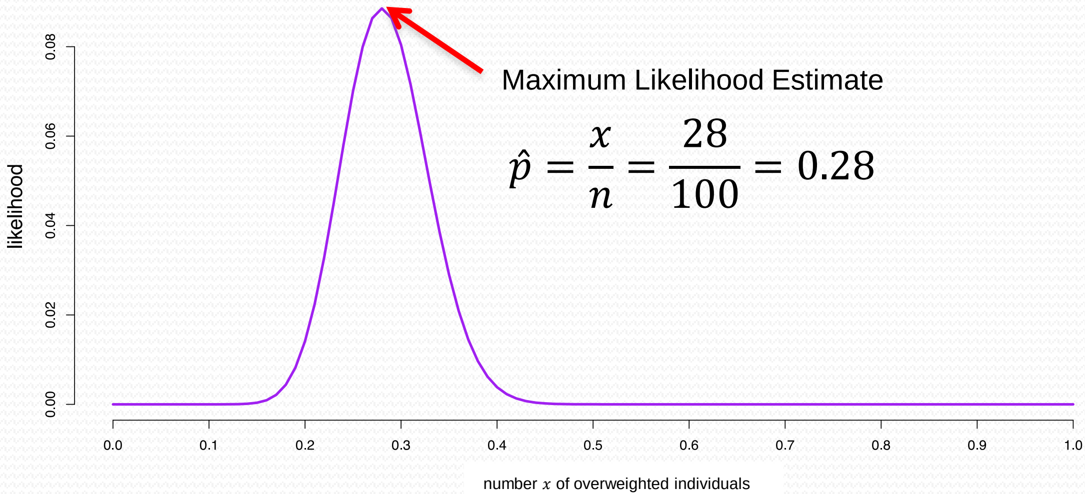
$$0 = -\frac{x}{\hat{p}} + \frac{n-x}{1-\hat{p}}$$

$$0 = \frac{-x(1-\hat{p}) + \hat{p}(n-x)}{\hat{p}(1-\hat{p})}$$

$$0 = -x + \hat{p}x + \hat{p}n - \hat{p}x \qquad \Rightarrow \hat{p} = \frac{x}{n}$$

: the proportion of positives!

Maximum Likelihood Estimate



$$\hat{p} = \frac{x}{n} = \frac{28}{100} = 0.28$$

Likelihood function: general definition

- For a sample $\boldsymbol{x} = (x_1, \dots, x_n)$
  - The likelihood function for a parameter θ is defined as

$$L(\theta; \boldsymbol{x}) = f(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

  - The log-likelihood function is

$$l(\theta; \boldsymbol{x}) = \ln\big(L(\theta; \boldsymbol{x})\big) = \sum_{i=1}^{n} \ln f(x_i; \theta)$$

  - It measures how likely (or expected) the sample is with respect to θ
  - We maximize it with respect to θ to obtain the likelihood estimate $\hat{\theta}$

Advantages of likelihood

- Practical method for estimating parameters
- Easily adaptable to different probability distributions
- It is often a good estimate

4 Quality of estimation

Unbiasedness

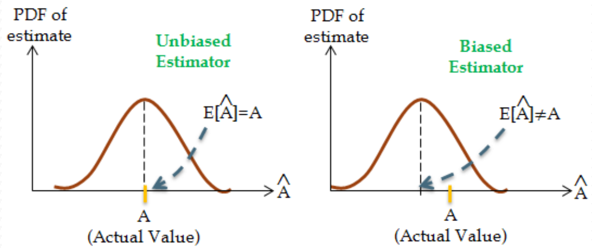
- A point estimator is **unbiased** for θ if the mean of its sampling distribution is equal to θ

$$E(\hat{\theta}) = E\left(\hat{\theta}(X_1, \dots, X_n)\right) = \theta$$

- The **bias** of a point estimate for θ is

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- Thus, a point estimate with *bias* = 0 is **unbiased**, otherwise it is **biased**



Examples

- The sample mean is always unbiased for estimating the population mean μ:

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \mu$$

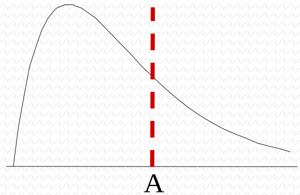
- Why do we divide by $n-1$ in the sample variance (and not by n)?

$$\begin{aligned} E((n-1)s^2) &= E\left(\sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})^2\right) = \sum_{i=1}^{n} E\left(x_i^2\right) - n\,E(\bar{x}^2) \\ &= n(\sigma^2 + \mu^2) - n(Var(\bar{x}) + E(\bar{x})^2) = n\left(\sigma^2 - Var(\bar{x})\right) \\ &= n\left(\sigma^2 - \frac{1}{n}\sigma^2\right) = (n-1)\sigma^2 \end{aligned}$$

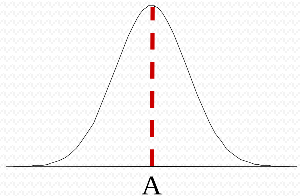
It follows that $E(s^2) = \sigma^2$.

Density of the estimate

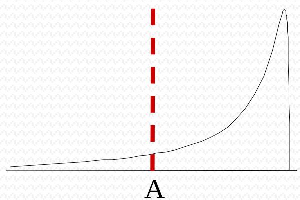
- Assume A is being estimate



Negative bias
Under estimate
High variability



Unbiased
On target estimate



Positive bias
Over estimate
Low variability

Consistency

- A point estimator is (weakly) **consistent** if

$$\Pr\left(\left|\hat{\theta}-\theta\right| > \varepsilon\right) \to 0 \text{ as } n \to \infty \text{ for any } \varepsilon > 0$$

Thus, a consistent point estimator should converge in probability to θ

- **Theorem:** A point estimator is consistent if

$$bias(\hat{\theta}) \to 0 \text{ and } Var(\hat{\theta}) \to 0 \text{ as } n \to \infty$$

Proof: Use Chebyshev's inequality in terms of the mean squared error

$$E\left(\left(\hat{\theta}-\theta\right)^2\right) = Var(\hat{\theta}) + \left(bias(\hat{\theta})\right)^2$$

Examples

- The sample mean is a consistent estimator of the population mean for any distribution with finite mean and finite variance.
  - How to prove it?

Efficiency

- The notations $E_\theta(\hat{\theta})$ and $Var_\theta(\hat{\theta})$ underlines that the true parameter is θ, hence the distribution of $\hat{\theta}$ depends on θ

- Assume we have two **unbiased** estimators of $\theta \in \Theta$, where Θ is a set of possible values, i.e.

$$\hat{\theta}^{(1)}, \hat{\theta}^{(2)}: E_\theta\left(\hat{\theta}^{(1)}\right) = E_\theta\left(\hat{\theta}^{(2)}\right) = \theta$$

- If $Var_\theta(\hat{\theta}^{(1)}) \leq Var_\theta(\hat{\theta}^{(2)})$ with strict inequality for at least one value of θ, then $\hat{\theta}^{(1)}$ is said to be **more efficient** than $\hat{\theta}^{(2)}$

Example with $E(x_i) = \mu$ and $Var(x_i) = \sigma^2$

- $\hat{\mu}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} x_i$ or $\hat{\mu}^{(2)} = \frac{x_1 + x_n}{2}$ for $n > 2$?

- Both estimators are unbiased:
  - $E\left(\hat{\mu}^{(1)}\right) = \mu$ and $E\left(\hat{\mu}^{(2)}\right) = \frac{E(x_1)+E(x_n)}{2} = \mu$

- Variance of the estimators:
  - $Var\left(\hat{\mu}^{(1)}\right) = \frac{\sigma^2}{n}$
  - $Var\left(\hat{\mu}^{(2)}\right) = \frac{\sigma^2+\sigma^2}{4} = \frac{\sigma^2}{2}$
  - Hence, $Var\left(\hat{\mu}^{(2)}\right) > Var\left(\hat{\mu}^{(1)}\right)$ if $n > 2$

- Conclusion: $\hat{\mu}^{(1)}$ is more efficient than $\hat{\mu}^{(2)}$

5 Cramer-Rao Bound

Fisher information

- Likelihood of random variables: $l(\theta;\boldsymbol{X}) = \ln L(\theta;\boldsymbol{X}) = \ln \prod_{i=1}^{n} f(X_i;\theta)$
- The Fisher Information about θ contained in a sample **x** is defined as

$$I(\theta) = E\left(\left(\frac{\partial}{\partial\theta}\{l(\theta;\boldsymbol{X})\}\right)^2\right) = E\left(\left(\frac{\partial}{\partial\theta}\{l(\theta;X_1,\ldots,X_n)\}\right)^2\right)$$

- **Theorem:** Under some regularity conditions (interchangeability of integration and differentiation), we get

$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\{l(\theta;\boldsymbol{X})\}\right)$$

Why is it measure of information for θ

- $L(\theta; \boldsymbol{X})$ and $l(\theta; \boldsymbol{X})$ is related to the probability of the sample

- The change of the probability with respect to θ is given by $\frac{\partial l(\theta;\boldsymbol{X})}{\partial \theta}$:
  - If $\frac{\partial l(\theta;\boldsymbol{X})}{\partial \theta}$ is close to 0, the probability is not affected by a slight modification of θ
  - If $\frac{\partial l(\theta;\boldsymbol{X})}{\partial \theta}$ is largely positive or negative, the probability changes a lot if θ changes slightly

- $\left(\frac{\partial l(\theta;\boldsymbol{X})}{\partial \theta}\right)^2$ measures the amount of information about θ in the sample $\boldsymbol{X}$

- $E\left(\frac{\partial l(\theta;\boldsymbol{X})}{\partial \theta}\right)^2$ measures **generally** the amount of information about θ in a sample from the current distribution

Example

- $X\sim\text{Exp}(\mu)$ follows an exponential distribution

$$L(\mu;\boldsymbol{x})=\prod_1^n f(x_i;\theta)=\prod_1^n (1/\mu)e^{-x_i/\mu}=\frac{1}{\mu^n}e^{-\frac{1}{\mu}\sum_1^n x_i}$$

$$l(\mu;\boldsymbol{x})=\ln(L(\mu;\boldsymbol{x}))=-n\ln\mu-\frac{1}{\mu}\sum_1^n x_i$$

$$\frac{\partial l}{\partial\mu}=-\frac{n}{\mu}+\frac{1}{\mu^2}\sum_1^n x_i\text{ ; the distribution fulfills the regularity conditions}$$

$$\frac{\partial^2 l}{\partial\mu^2}=\frac{n}{\mu^2}-\frac{2}{\mu^3}\sum_1^n x_i\Rightarrow I(\mu)=-E\left(\frac{\partial^2 l}{\partial\mu^2}\right)=-\left(\frac{n}{\mu^2}-\frac{2}{\mu^3}\sum_1^n E(X_i)\right)=-\frac{n}{\mu^2}+\frac{2}{\mu^3}\cdot n\mu=\frac{n}{\mu^2}$$

Cramér-Rao inequality

- Under the same regularity conditions as for the previous theorem, the following inequality holds for any **unbiased** estimator:

$$Var_\theta(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

- If an unbiased estimator attains this lower bound, it is **efficient**.

- Example: $X \sim \mathrm{Exp}(\mu)$ follows an exponential distribution

$$Var_\mu(\hat{\mu}) \geq \frac{\mu^2}{n}$$

for any unbiased estimate $\hat{\mu}$

6 Conclusion

Conclusion

- Estimation is essential to infer the distribution of data
- Maximum likelihood method is the most famous method!
  - Implemented in many softwares and languages
  - Well studied in practice and in theory
- The quality of an estimator must be analyzed
- The Cramer-Rao bound is an useful tool to establish the efficiency of an estimator