# Parameter estimation

" **Estimation**": Using **low accurate** measuring tools (using data collected in a **very limited sample** of population) to determine as precisely as possible value of a certain parameter (of all population).

An opinion or judgment of the worth, extent, or quantity of anything, formed **without using precise data**; as, estimations of distance, magnitude, mount, or moral qualities.

# Parameter Estimation

* Estimation methods

* Distribution of estimated parameters

* Comparing distribution of estimated parameter wit Normal distribution

* Confidence Interval of estimation (Interval Estimation)

# Estimation of rate (proportion, probability)

**Example: - Tossing a coin: What is possibility to get "figure side" ?**

**- Tossing a dice: What is probability to get the side with six points ?**

**-Tobacco smoking study: How large is smoking rate in elderly people (over 60) ?**

**- Proportion of rural households using rain water?**

Normally it is very hard to determine exactly the real value of concerned parameter. The one must estimate the value by using some suitable method

→ Meet with some error in estimation

→ Need to evaluate accuracy of estimation: with a given precise level the estimation result is acceptable or not?

To determine possible accuracy of estimation with given presice level, we need to know **distribution** of the estimation

# Distribution of variable

**The set of values of a set of data, possibly grouped into classes, together with their frequencies or relative frequencies**

**Distribution of variable: the set of possible values with their probability**

**Example:**

- **Tossing a coin:** Possibility to get "figure side" = 1/2 → uniform distribution of two values "figure side" and "number side"

- **Tossing a dice:** Probability to get the side with six points = 1/6 → uniform distribution of 6 values * , ** , *** , **** , ***** and ******

- **Tossing 6 dices:** Non-uniform distribution of 36 values 6* , 7* , 8* , … , 35* and 36*

# Concept of probability distribution

*** Discrete distributions:** *Variable X with*

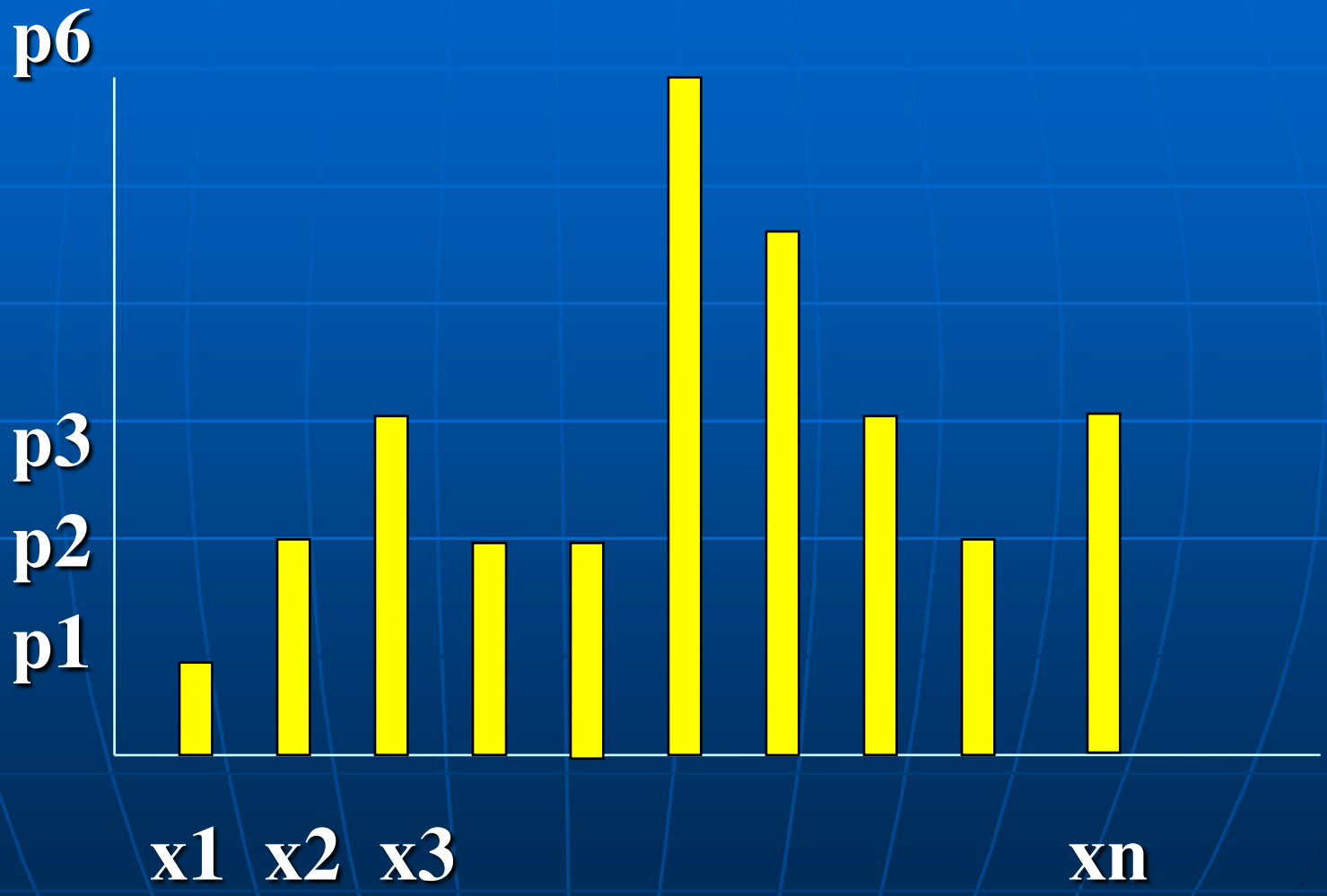
| Value: | X1 | X2 | X3 | . . . | Xn |
|---|---|---|---|---|---|
| | \| | \| | \| | | \| |
| Probability: | p1 | p2 | p3 | | pn |

$$P\{X=X1\} = p1 >= 0$$
$$P\{X=X2\} = p2 >= 0$$
$$. \quad . \quad .$$
$$P\{X=Xn\} = pn >= 0$$

$$p1 + p2 + . \ . \ . + pn = 1\ (100\%)$$

# Concept of probability distribution

* Discrete distributions:



# Concept of probability distribution

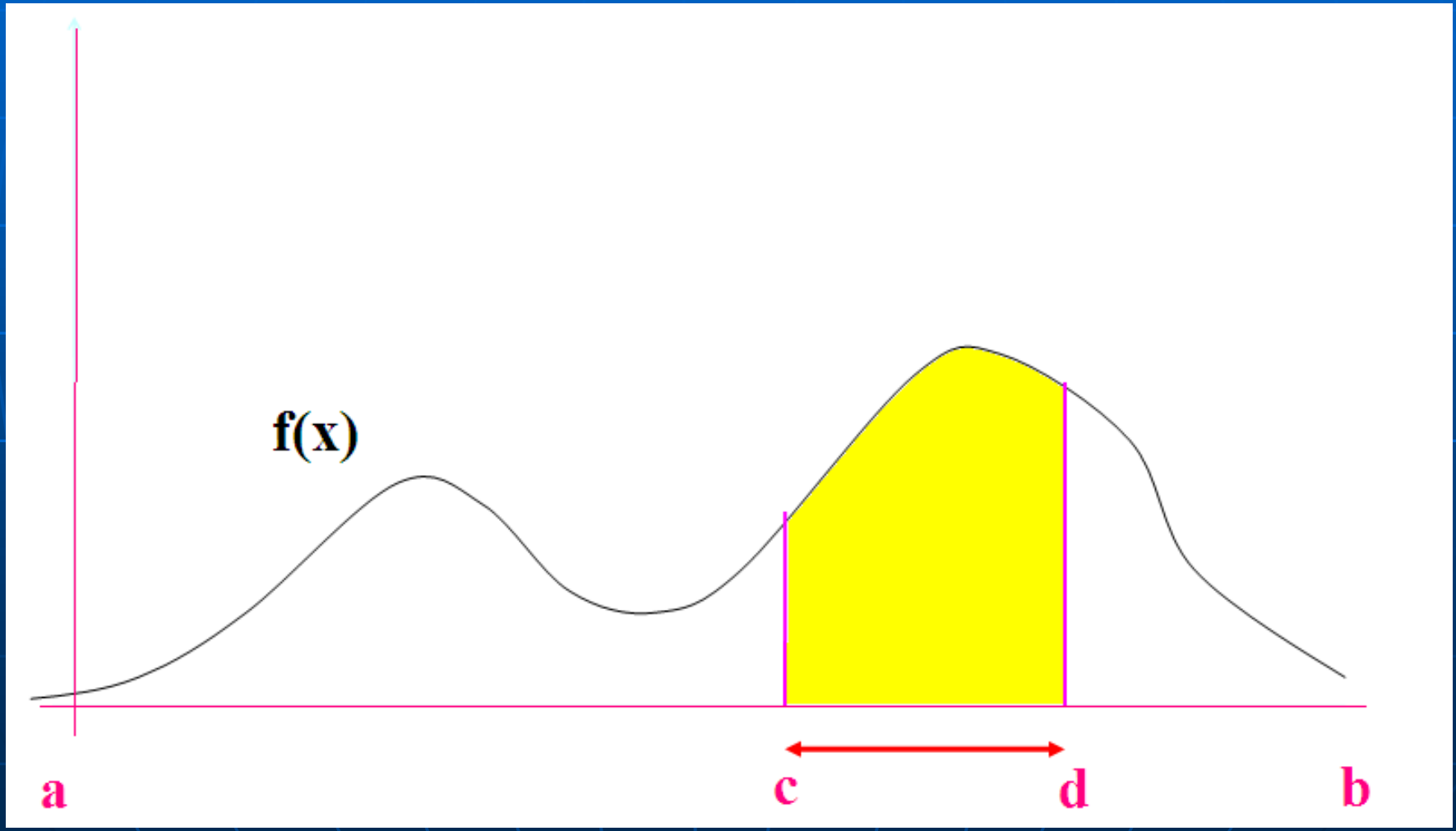
*** Continuous distributions:** *Variable X taken value x inside interval (a;b) with density function f(x) >= 0*

$$\int_a^b f(x)dx = 1 \; ; \quad -\infty \leq a < b \leq +\infty$$

$$P\{X \in (c;d)\} = \int_c^d f(x)dx \quad \text{for } a \leq c < d \leq b$$

# Concept of probability distribution

* Continuous distributions:



# Estimation of rate (proportion, probability)

In study population let's consider a binary variable $X$ with 2 values $0$ and $1$

Suppose $X$ takes value $1$ with rate (proportion, probability) $p$ and value $0$ with rate $1 - p$, where $p$ is unknown ($0 < p < 1$)

Usually we estimate the rate $p$ by taking a sample of the variable $X$ with $n$ observations $x(1), x(2), \ldots, x(n)$. Then determine the number $m(p)$ of values $1$ among the $n$ observations and perform the proportion

$$m(p) / n$$

as an estimated value of the rate $p$.

That way of estimation is "reasonable" or not?

THEOREM (**LOUVLIER**). The proportion

**m(p) / n**

tends to **p** when **n** tens to infinity (is very large).

- The theorem proved mathematically shows the taking the proportion m(p) / n for estimation of the rate p is completely "reasonable": we can get the "true" rate when the sample size is very large.

# Distribution of sample rate (proportion)

Let X be a binary variable taken value 1 with unknown probability p and taken value 0 with probability 1 – p (Bernoulli's distribution).

Estimating p : perform a sample x(1), x(2), … , x(n) of X and take m(p) / n as an estimation of p (m(p) = number of 1's appeared in the sample).

Quantity m(p) / n should take values

0/n , 1/n , 2/n , … , (n-1) / n , n/n ,

each with certain "possibility" (probability)

# Distribution of sample rate (proportion)

Quantity m(p) / n is a random variable with binomial distribution with parameters p and n .

# Binomial Distribution

$$P\ m(p) = k\ = C_n^k\ p^k\ 1 - p\ ^{n-k};\ k = 0,1,2,\ldots,n$$


Parameters of binomial distribution are the rate p and number n of experiments

# Distribution of sample rate (proportion)

- Binomial distribution can be used to evaluate error in estimating p by m(p) / n

- For small n, calculation with binomial distribution is practicable

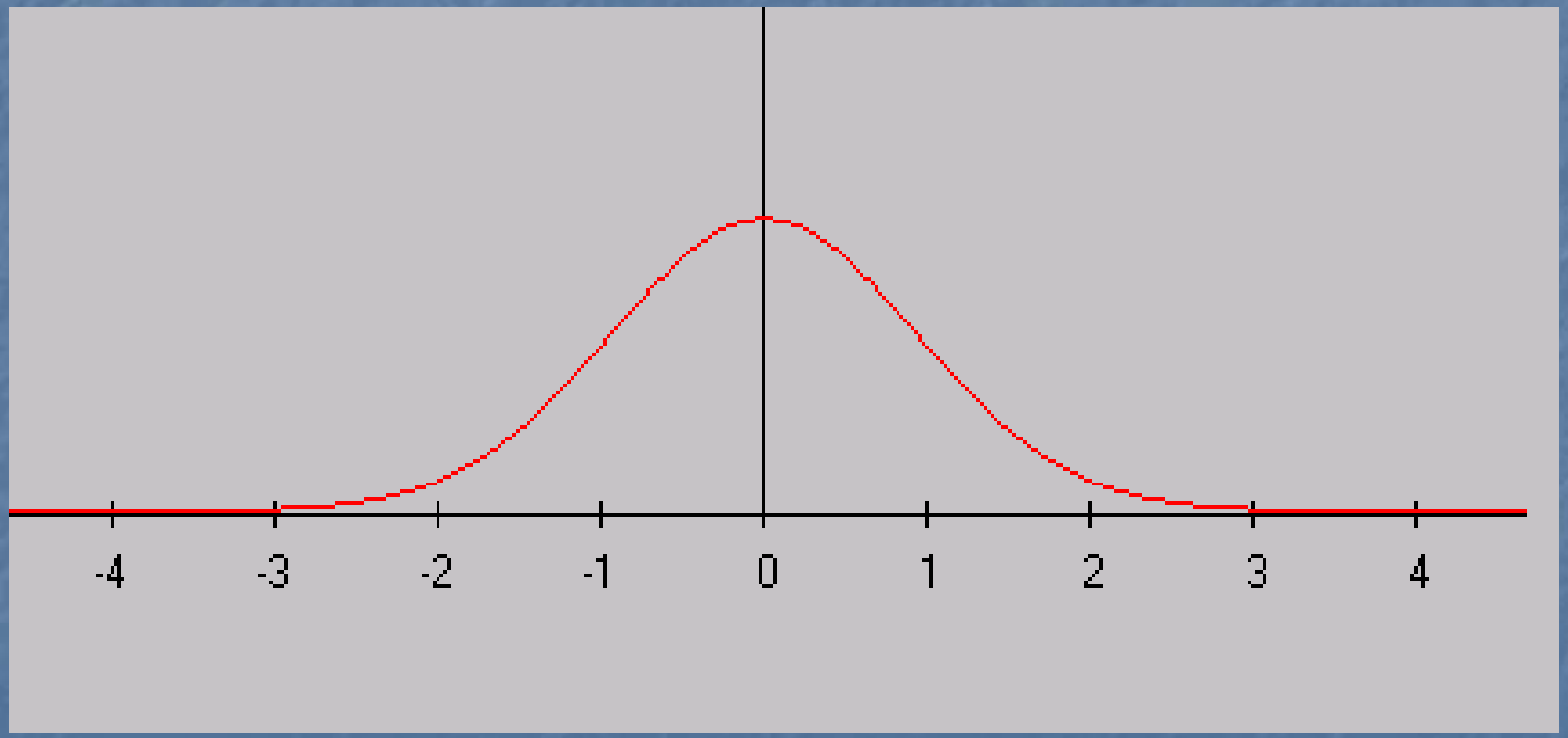
- For n large the calculation is very cumbersome → need to have another method for evaluation

# Distribution of sample rate (proportion)

**MOIVRE-LAPLACE THEOREM.** Let **X** be a binary variable taken value **1** with probability **p** and value **0** with probability **1 – p** . For the sample **x(1), x(2), … , x(n)** of **X** with **n** observation let **m(p) / n** be the proportion 1's number per sample size. Then the proportion is a quantity with distribution approximate to **Normal distribution** with mean value (expectation) **p** and variance **p . (1-p) / n** when the sample size **n** is large.

# Normal distribution (Gauss distribution)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left( - \frac{x - \mu^{2}}{\sigma^{2}} \right)$$




Normal distribution defined by its “expectation” và “variance”

# Distribution of sample rate (proportion)

- Moivre-Laplace Theorem can be used to evaluate errors in estimation of proportion:

  allows to determine Confidence Interval of the estimation

# Confidence interval of estimation (interval estimation)

- **Confidence Interval** of estimation is an interval containing the estimated value of parameter, informing the **true value** of parameter can be some point inside the interval with given probability **a** .

- For a variable with normal distribution with expectation $p$ and variance $p\,.\,(1\text{-}p)\,/\,n$

  **95% Confidence Interval** of estimation of $p$ is the interval

$$\left(p - 1.96 * \sqrt{p.(1-p)/n}\,;\; p + 1.96 * \sqrt{p.(1-p)/n}\right)$$

# Confidence interval of proportion

**Because estimation of proportion (by Moivre – Laplace Theorem) is a quantity with distribution approximate to Normal Distribution, 95% Confidence Interval of proportion estimation is**

$$\left[ \hat{p} - 1.96 * \sqrt{\hat{p}.(1 - \hat{p}) / n} ; \hat{p} + 1.96 * \sqrt{\hat{p}.(1 - \hat{p}) / n} \right]$$

**where**

$$\hat{p} = m(p) / n$$

# Application

**Problem: How to estimate the amount of fishes in a lake?**

**Step 1. The amount of fishes in a lake is $N$ =?**

- **Nesting 1st time to capture certain amount *m1* of fishes**
- **Mark each fish of that amount. Then release those fishes back into the lake. Hence the true proportion of marked fishes in the lake equals**

$$p = m1 / N$$

**Step 2.** Nesting 2nd time to capture another amount ***n*** of fishes

- Count the amount ***m2*** of marked fishes among ***n*** fishes captured in the 2nd time
- Estimate the proportion ***p*** of marked fishes by ***p' = m2 / n*** with 95% confidence interval

$$\left[ p' - 1.96 * \sqrt{p'.(1 - p') / n} ; p' + 1.96 * \sqrt{p'.(1 - p') / n} \right]$$

**Step 3.** We are sure (with 95% possibility) that the true proportion $p$ of marked fishes in the lake should be a certain number inside the confidence interval, that means

$$p = m1/N \geq p' - 1.96 * \sqrt{p'.(1 - p')/n};$$

$$p = m1/N \leq p' + 1.96 * \sqrt{p'.(1 - p')/n}$$

We can be sure (with 95% certainty) that the amount of fishes in the lake should be a number between

$$m1/(p' + 1.96 * \sqrt{p'.(1 - p')/n}) \leq N$$

$$N \leq m1/(p' - 1.96 * \sqrt{p'.(1 - p')/n})$$

# Estimation of Expectation

**Expectation of variable =**

**Mean value of variable in whole population**

**For estimation of expectation of a quantitative variable X , a sample x(1), x(2), … , x(n) can be chosen and sample mean value (sample average)**

$$\bar{X} = \frac{1}{n}(x(1) + x(2) + \ldots + x(n))$$

**Can be taken as an estimated value of expectation parameter E(X) of X**

**→ That manner (of estimation) is correct or not?**

**THEOREM (Law of Large Numbers).** When the sample size $n$ tends to infinity (is very large), the sample mean value

$$Mean(X) = \bar{X} = \frac{1}{n}(x(1) + x(2) + \ldots + x(n))$$

will convergent to the true value of expectation (theoretical mean value) of $X$ .

**CONCLUSION :** Sample mean value is a "good" estimation of Expectation:

$$Mean_n(X) \xrightarrow{n \to \infty} E(X)$$

The estimation is very close to true value of expectation if sample size ***n*** is very large.

**Problem:** Although Sample mean value is a "good" estimation of Expectation, there exists always some error of that estimation

→ How to evaluate the error in that estimation?

→ Need to know about **distribution** of sample mean value

# Distribution of sample mean value

**THEOREM. Let variable X have normal distribution with expectation $\mu$ and variance $\sigma^2$ and select a sample x(1), x(2), … , x(n) of that variable. Then the sample mean value**

$$\bar{X} = \frac{1}{n}(x(1) + x(2) + \ldots + x(n))$$

**is a quantity with normal distribution with mean value equal $\mu$ and variance equal $\sigma^2 / n$ .**

**→ The Theorem gives a base for determining Confidence Interval of estimation to evaluate the error of estimation**

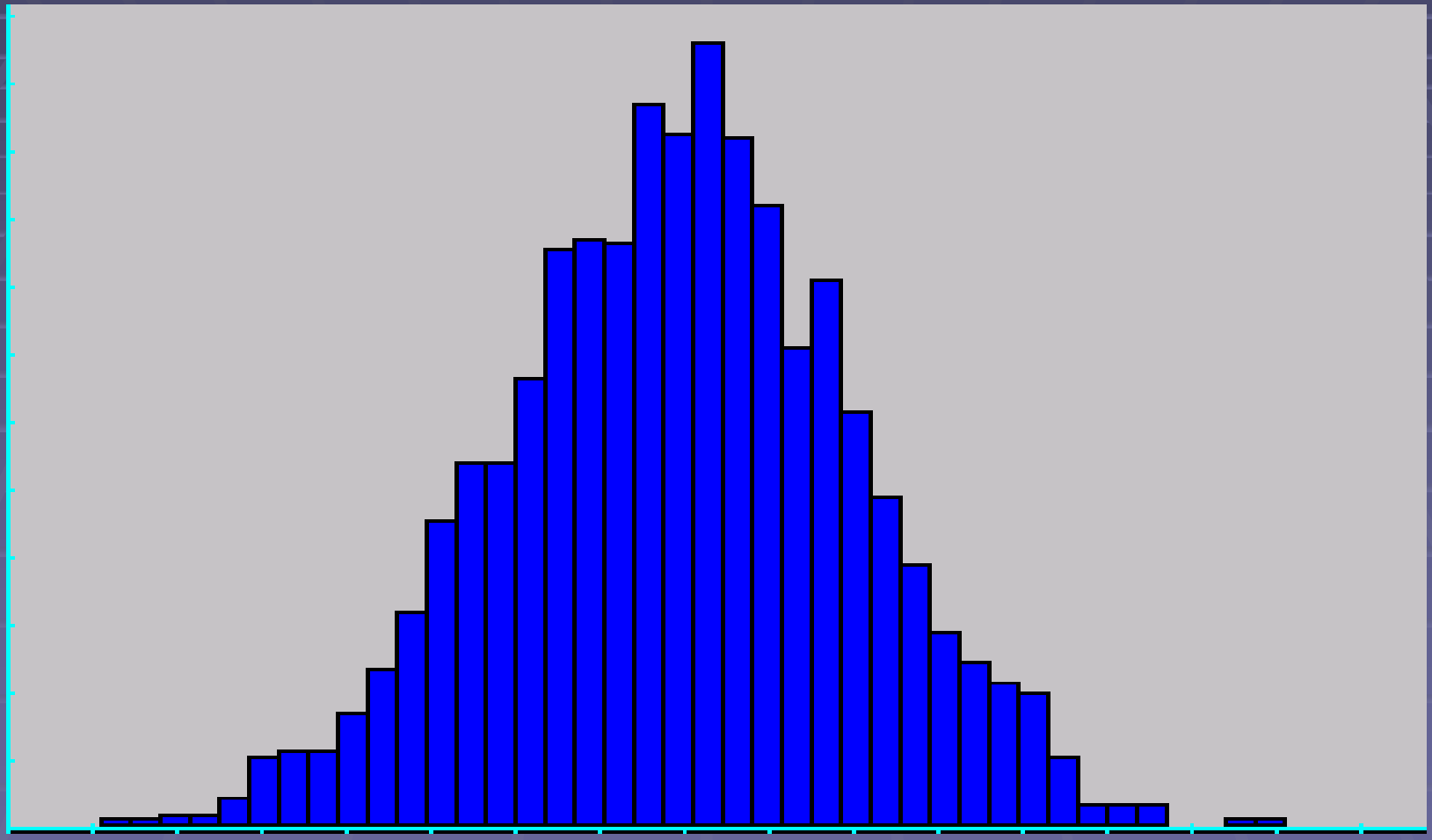
# Confidence Interval of sample mean value

- **Confidence Interval** of estimation is an interval containing the estimated value, confirming the true value of estimated parameter should be a point of that interval with a given probability **a**

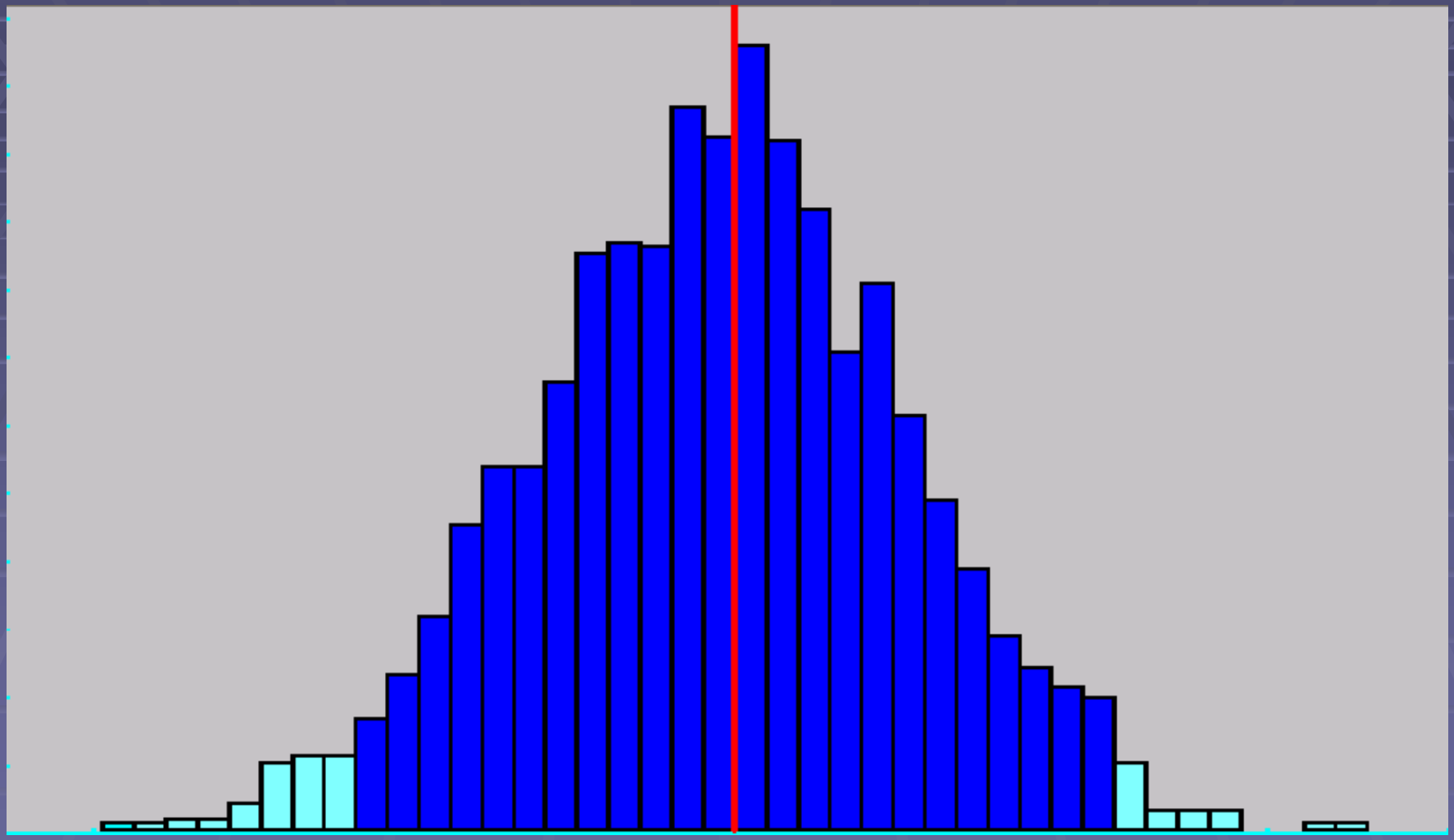
- For a normal distributed estimation quantity with expectation $\bar{X}$ and variance $\sigma^2 / n$ , the **95% Confidence Interval** (a = 95%) is defined by

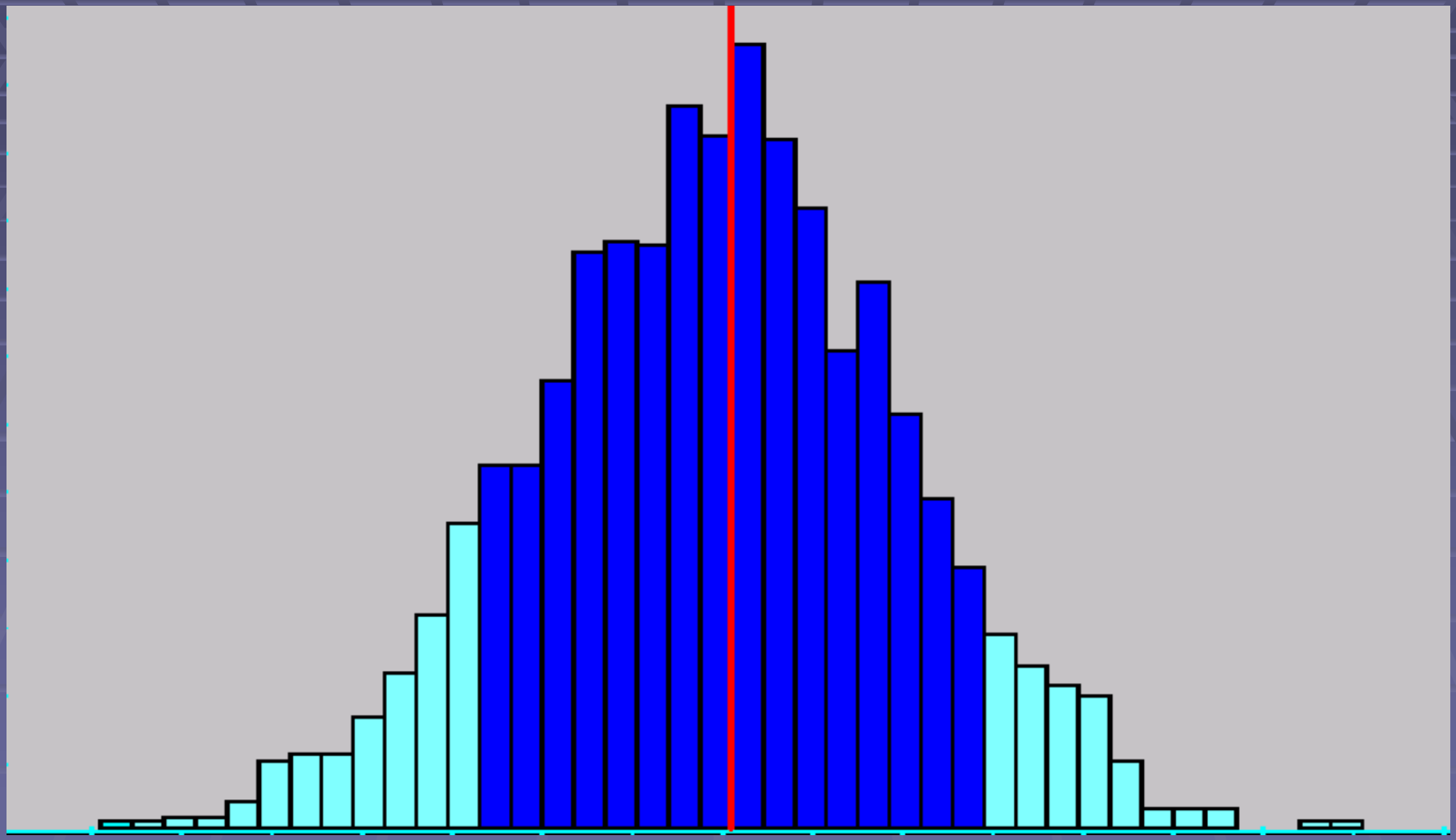
$$\left[ \bar{X} - 1.96 * \sigma / \sqrt{n} ; \bar{X} + 1.96 * \sigma / \sqrt{n} \right]$$

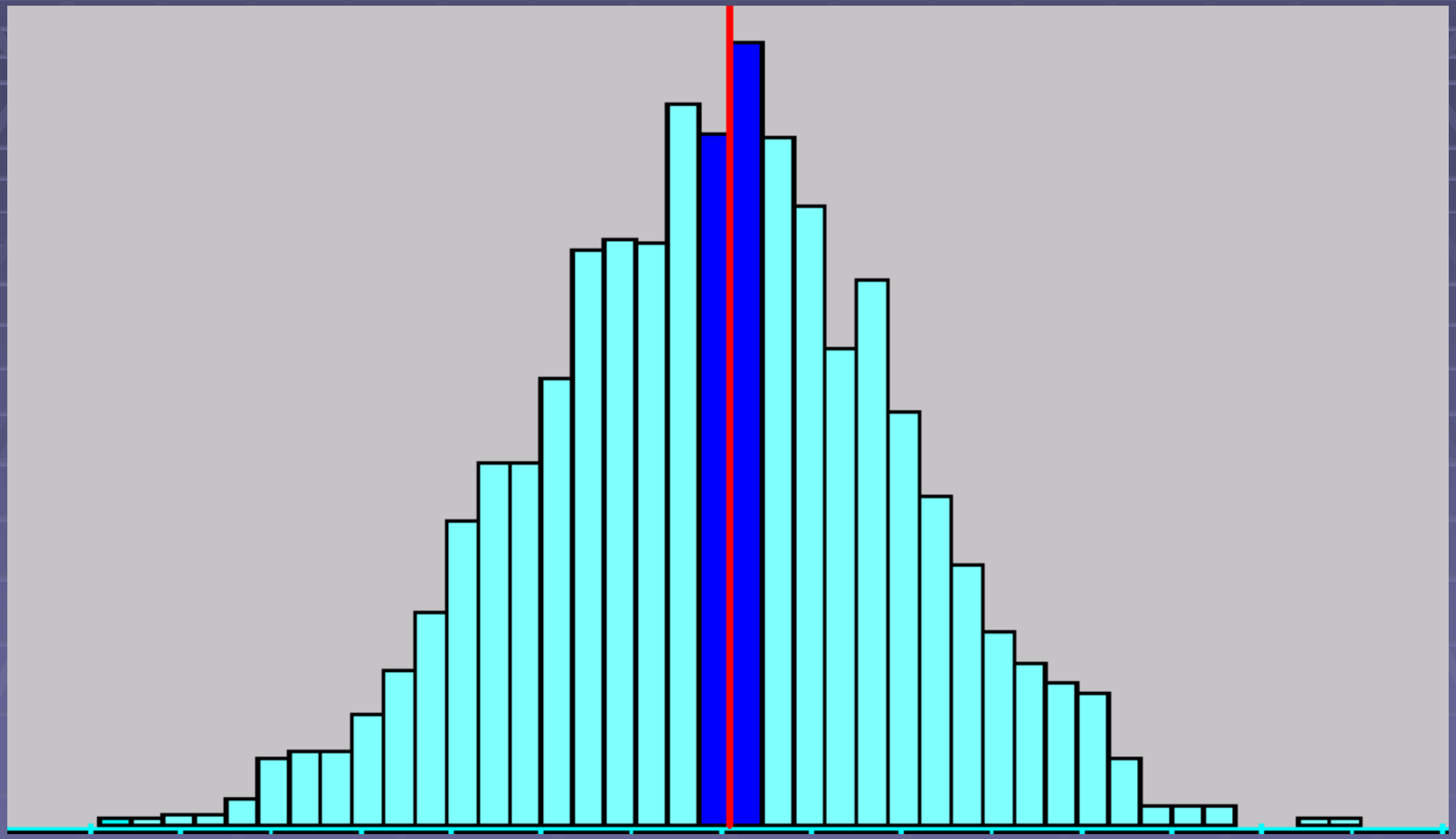
# Normal distribution



# Confidence Interval







# Confidence Interval for Non-normal distributed variable

**CENTRAL LIMIT THEOREM.** Suppose that $X$ be a variable with expectation $\mu$ and variance $\sigma^2$. Let $x(1), x(2), \ldots, x(n)$ be a sample of $X$ with $n$ observations and

$$\bar{X} = \frac{1}{n}(x(1) + x(2) + \ldots + x(n))$$

be a sample mean value. Then the mean value has distribution approximate to a normal distribution with expectation $\bar{X}$ and variance $\sigma^2 / n$ when sample size n is large

# Confidence Interval of sample mean value for non-normal variable

The above theorem provides a base to give **Confidence Interval** of mean value for non-normal distributed variable:

- If sample size ***n*** is **very large** then mean value of a variable with finite variance is an estimation of expectation with **95% Confidence Interval** (a = 95%) given by

$$\left[ \bar{X} - 1.96 * \sqrt{S^2 / n} ; \bar{X} + 1.96 * \sqrt{S^2 / n} \right]$$

where

$$S^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (x(i) - \bar{X})^2$$

# Application 2

- **Example.** In aquaculture, to determine the right moment for shrimp catching, the owner time by time captures small amount of shrimps to weight them. How many shrimps must be caught to see whether the average weight of all shrimps in lake is not different from standard weight more than 1 gram, knowing the shrimps weight is a quantity normally distributed with standard deviation equal 10 grams?

- Assume that the real average weight of shrimps in the lake is **c**, and the standard weight for fishing is **b** . Then if a sample with **n** shrimps is performed, the estimated sample mean value is a normal distributed with mean **c** and variance **100/n**
- **95% confidence interval** of that estimation is

$$\left[ c - 1.96 * \sqrt{100/n} ; c + 1.96 * \sqrt{100/n} \right]$$

the **real average weight** of all shrimps does not differ from **b** more than **1**gr if the confidence interval contains the value **b** , therefore

$$1.96 * \sqrt{100/n} < 1$$

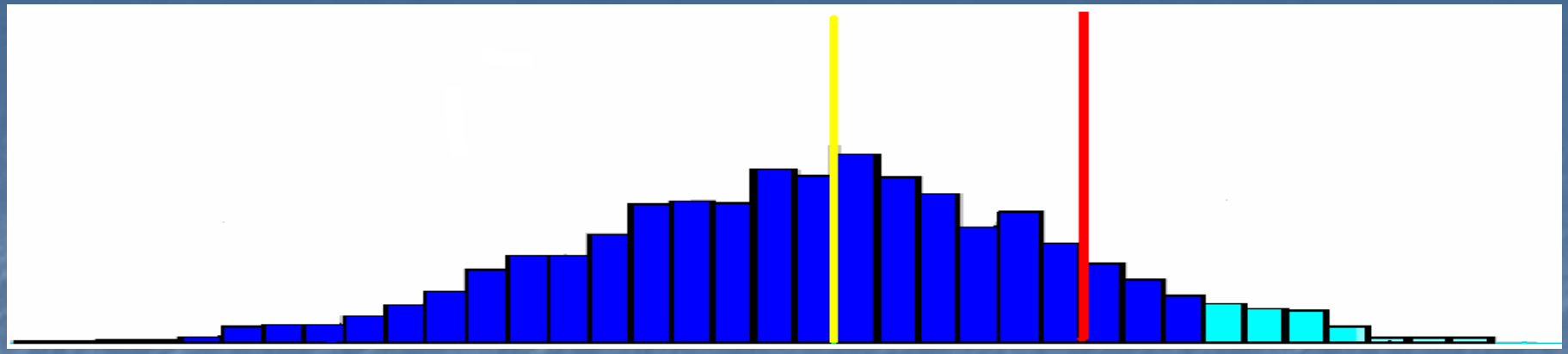
Then **n > 384.**

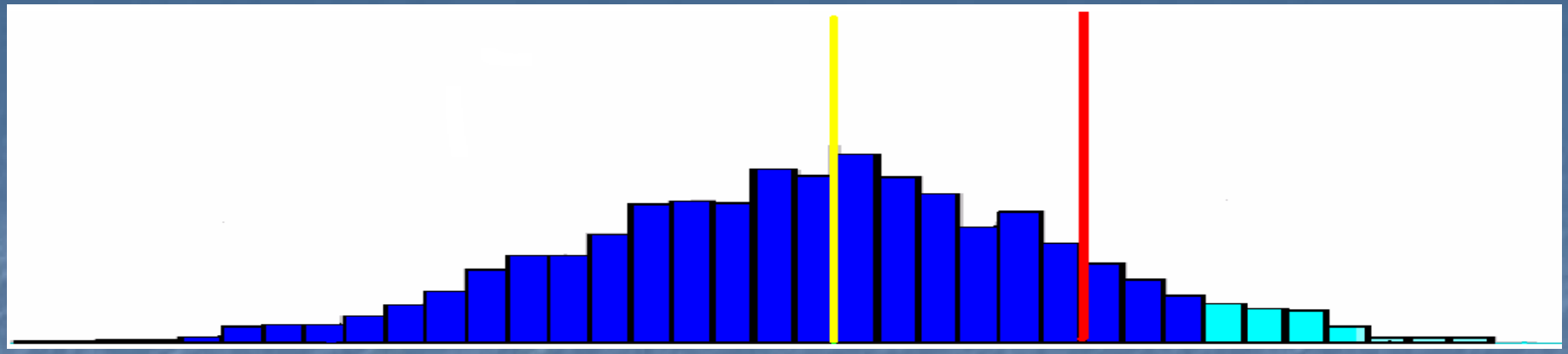
# Application 3

- **Example.** Malnutrition rate of under 8 children counted 35% for the period 2000-2005. There is an opinion saying that children nutrition is improved after 2005 and now malnutrition rate has been decreased to 30%. To check if the opinion is correct or not, we must collect data from a sample of certain amount of children.
- PROBLEM: How many children must be taken in the sample to have correct conclusion with confidence level of 95% (or 90%, 99%)?

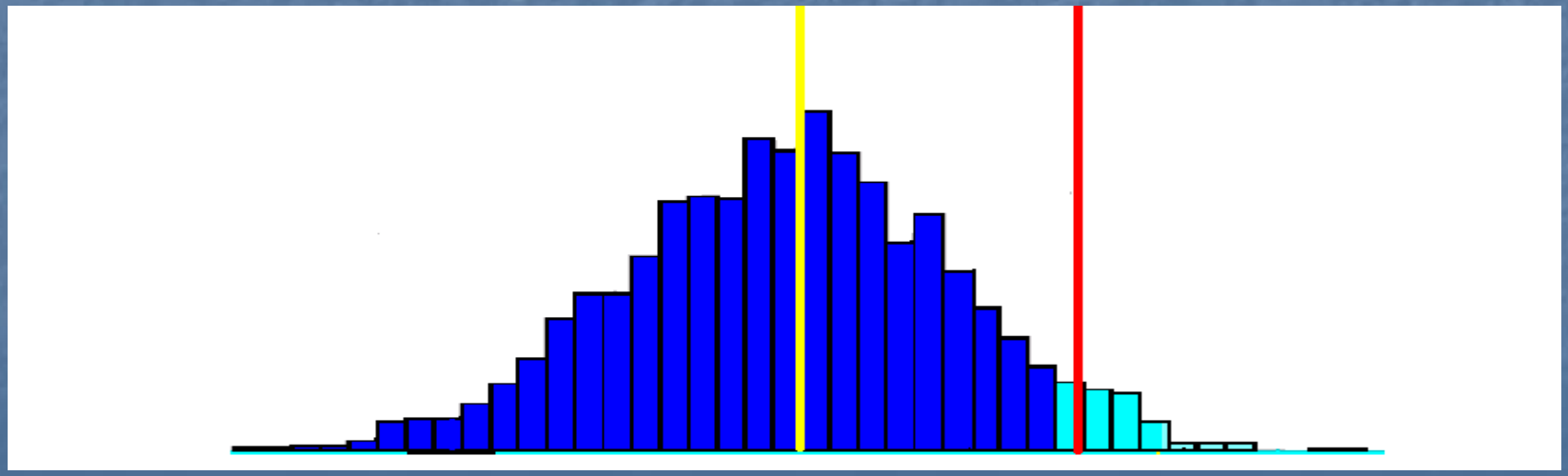
# Sample size determining

- If the opinion is right, the malnutrition rate of children must be 30%. For the sample size equal n , variance of estimated rate should be equal (0.3 * 0.7) / n . When n is small, the variance is large, the variation of estimation is large and then may be by chance the estimated rate should be more than 35% while the true rate counts only 30%.





- For larger n, variance (0.3 * 0.7) / n is smaller, the variation of the rate decreases and estimated value of the rate should not reach by chance to 35% (with confidence level 90%, 95% or 99%).



- In order that the estimate rate should not reached 35% by chance, n must be such large that variance (0.3 * 0.7) / n to be small enough so that

$$30\% + 1.65 * \sqrt{(0.3 * 0.7) / n} < 35\%$$

Then

$$(0.3 * 0.7) / n < ((35\% - 30\%)/1.65)^2$$

and n must be at least 0.21*1.65*1.65 / 0.0025 ~ 235 → need to have at least 235 children in the sample.

# Using SPSS and STATA in estimation

## *ESTIMATION OF EXPECTATION AND VARIANCE*

EXCEL :

SPSS : command

**Analyze**

**Descriptive Statistics**

**Explore…**

## *CONFIDENCE INTERVAL PLOT*

SPSS : command

**Graph**

**Error Bar ...**

**Simple + Sumary for groups os cases**

**…**

# Using EXCEL and STATA in estimation

***RATE ESTIMATION***

EXCEL :

SPSS : command

**Analyze**

**Descriptive Statistics**

**Frequency …**