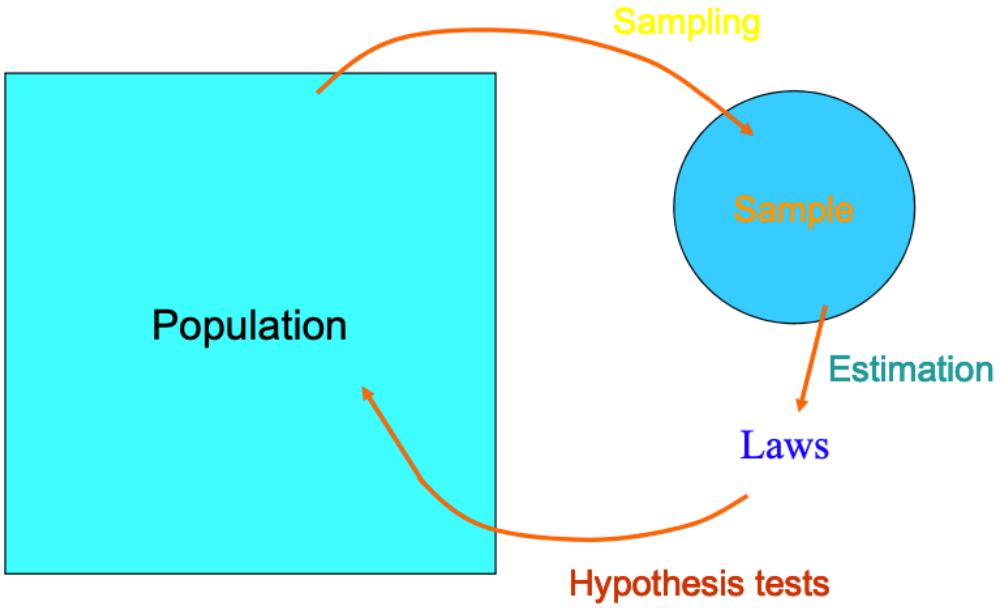
PARAMETER ESTIMATION

Road Map

- What is parameter estimation?
- Estimation of rate (proportion)
- Confidence interval of estimation (Interval estimation)
- Estimation of expectation

What is parameter estimation?

- Use sample data to estimate the measurable characteristics (parameters) of a larger population
- Parameter estimation is the process of using data to infer the values of unknown parameters within a statistical model



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 ?

Parameter estimation

- Usually it is very hard to determine exactly the real value of the 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 the given precise level, we need to know distribution of the estimation

Distribution of a 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

Distribution of a variable

- Example:
  - Tossing a coin: possibility to get “figure side” = 1/2

    → uniform distribution of 2 values “figure side” and “number side”

  - Tossing a dice: probability to get the side with 6 points = 1/6

    → uniform distribution of 6 values *, **, ***, ****, ***** and ******

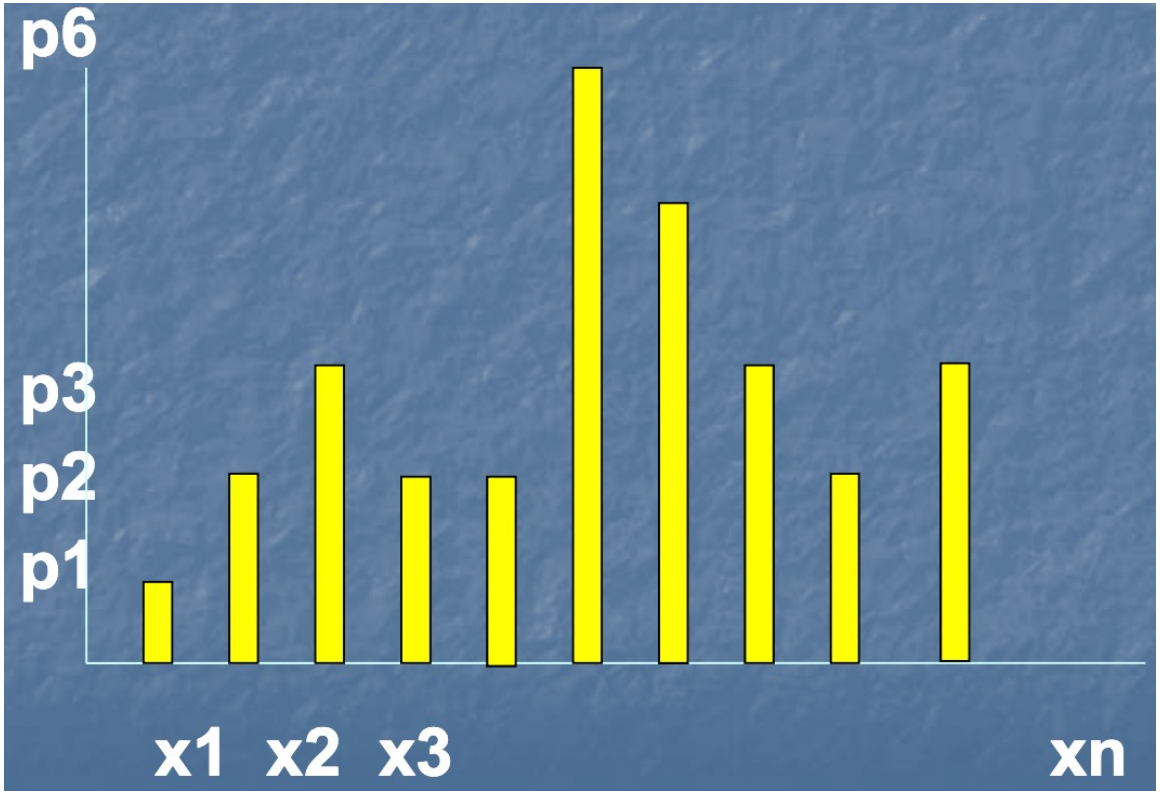
Concept of probability distribution

- (1) 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 … P{X=Xn} = pn >= 0					
p1 + p2 + … + pn = 1 (100%)					

Concept of probability distribution

- Discrete distributions:



Concept of probability distribution

- Discrete probability distributions:
  - The probability distribution is defined by a probability function, denoted by f(x), that provides the probability for each value of the random variable.

Expected value

- The expected value, or mean, of a random variable is a measure of its central location:

$$E(x) = \mu = \Sigma x f(x)$$

- The expected value is a weighted average of the values the random variable may assume. The weights are the probabilities.

Variance and standard deviation

- The variance summarizes the variability in the values of a random variable.

$$Var(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x)$$

- The variance is a weighted average of the squared deviations of a random variable from its mean. The weights are the probabilities.
- The standard deviation σ is defined as the positive square root of the variance.

Concept of probability distribution

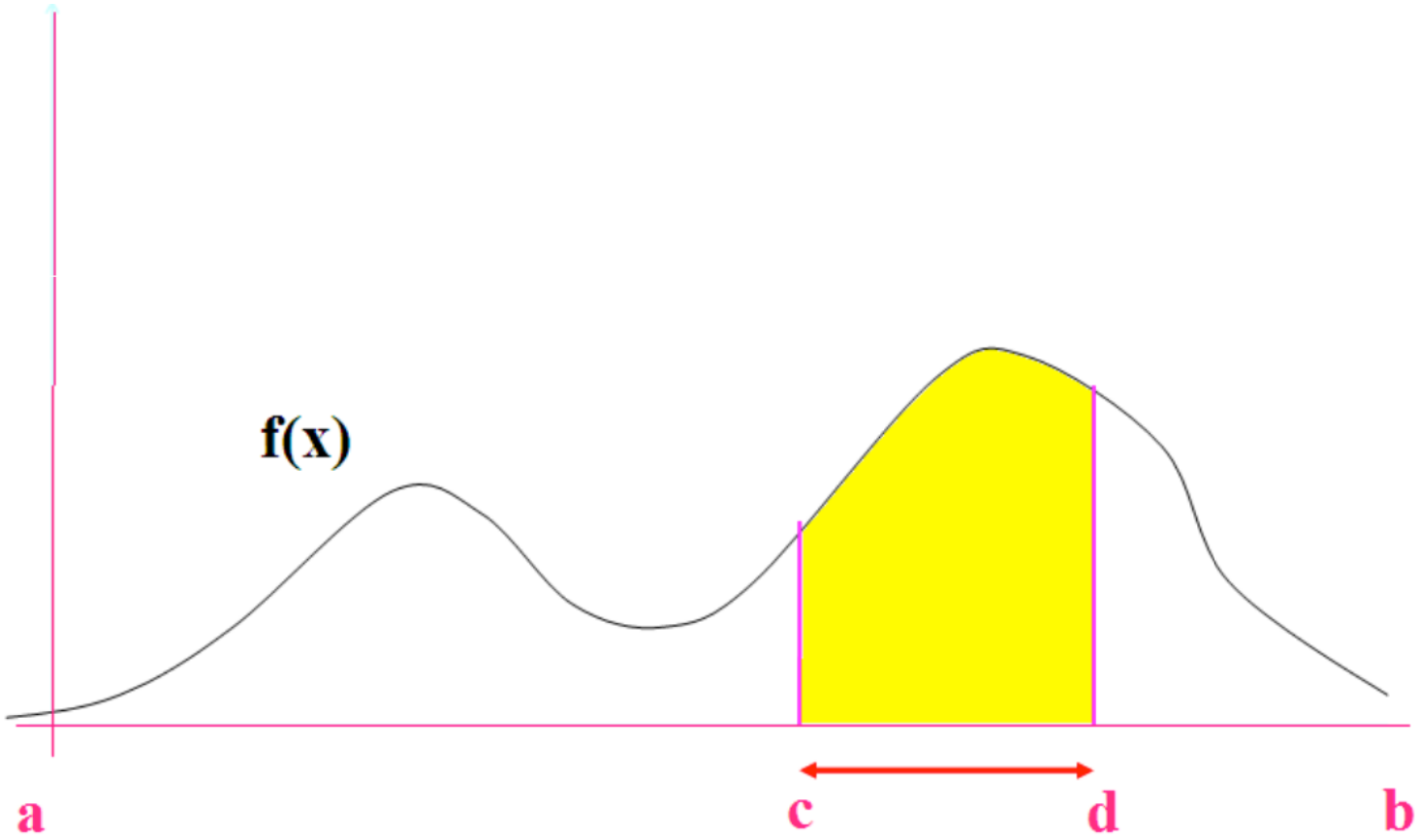
- (2) Continuous distributions: Variable X taken value x inside interval (a;b) with density function f(x) >= 0

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

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

Concept of probability distribution

- (2) 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), …, 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:

- The proportion: m(p) / n

  tends to p when n tends to infinity (is very large)
- The theorem proved mathematically shows that 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 the values:

  0/n, 1/n, 2/n, … , (n-1)/n, n/n,

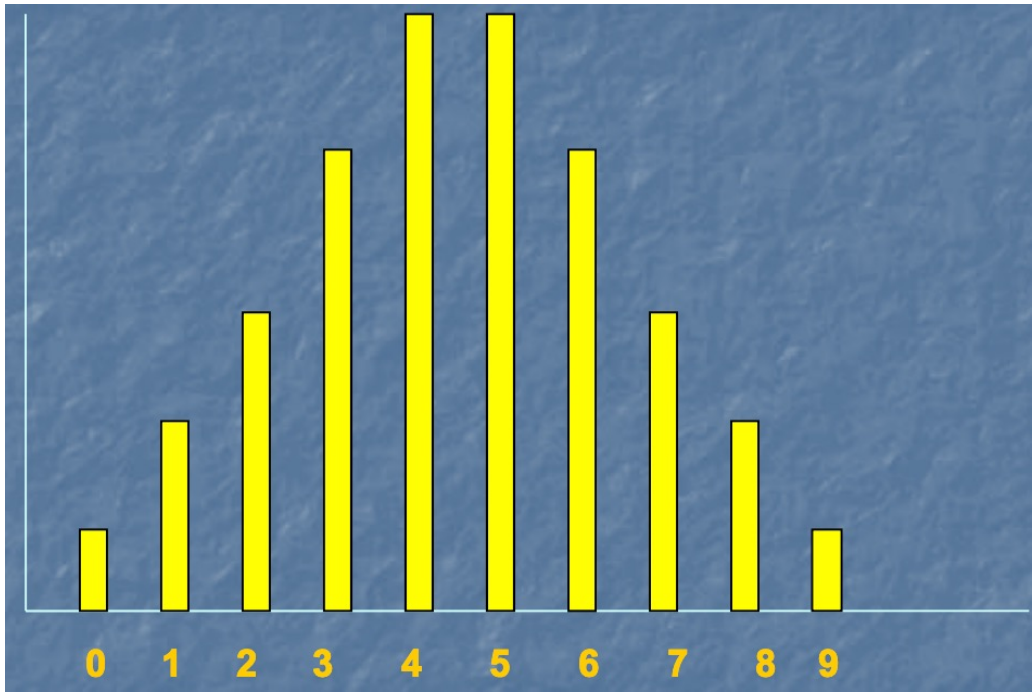
  each with certain "possibility" (probability)

Distribution of sample rate (proportion)

- Quantity m(p)/n is a random variable with binomial distribution with parameters p and n noted by B(p;n)

Binomial Distribution B(p;n)

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



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

Binomial Probability Distribution

- Expected value:

$$E(x) = \mu = np$$

- Variance:

$$Var(x) = \sigma^2 = np(1 - p)$$

- Standard deviation:

$$SD(x) = \sigma = (np(1 - p))^{1/2}$$

Binomial Probability Distribution

- Use excel to compute binomial probabilities
  - BINOMDIST is used to compute the probability and cumulative probability.

Distribution of a sample rate (proportion)

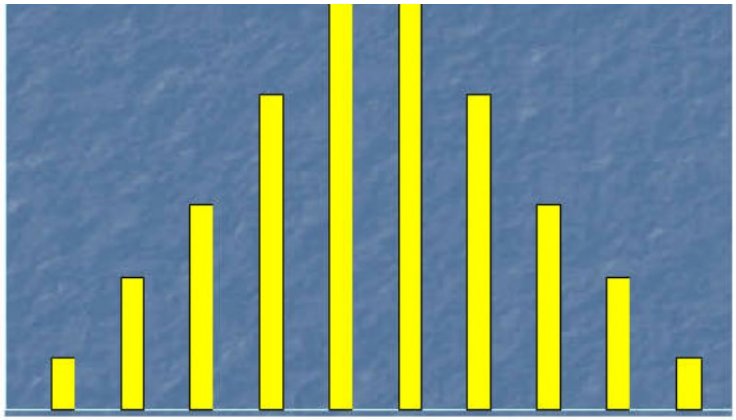
- Binomial distribution can be used to evaluate error in estimating p by m(p)/n
- For small n, the calculation with binomial distribution is practicable
- For large n, the calculation is very cumbersome → approximate Binomial distribution by Normal distribution:

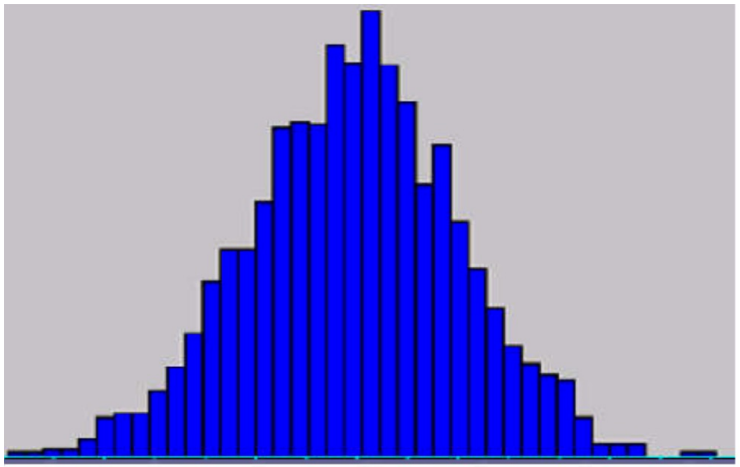
$$B(\mathrm{p};\mathrm{n}) \sim \mathrm{N}(\mathrm{p};\mathrm{p}(1-\mathrm{p})/\mathrm{n})$$

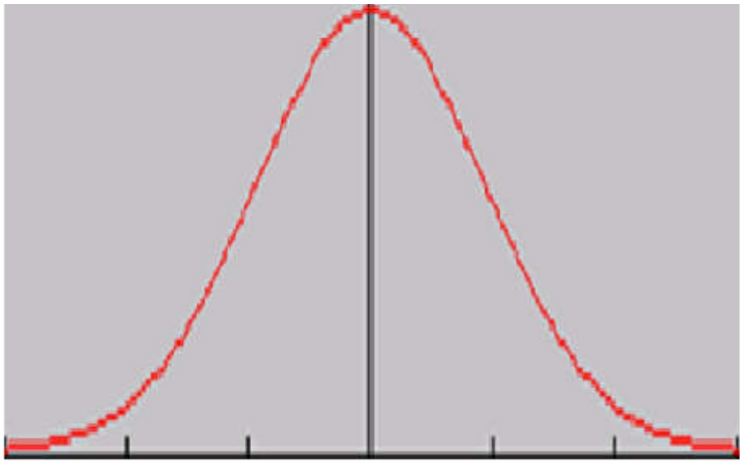
Distribution of estimation of proportion

- The sampling distribution of $\bar{p}$ plays a key role in computing the margin of error for this interval estimate
- The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever $np \geq 5$ and $n(1-p) \geq 5$

Approximation of B(p;n) by N(p;p(1-p)/n)

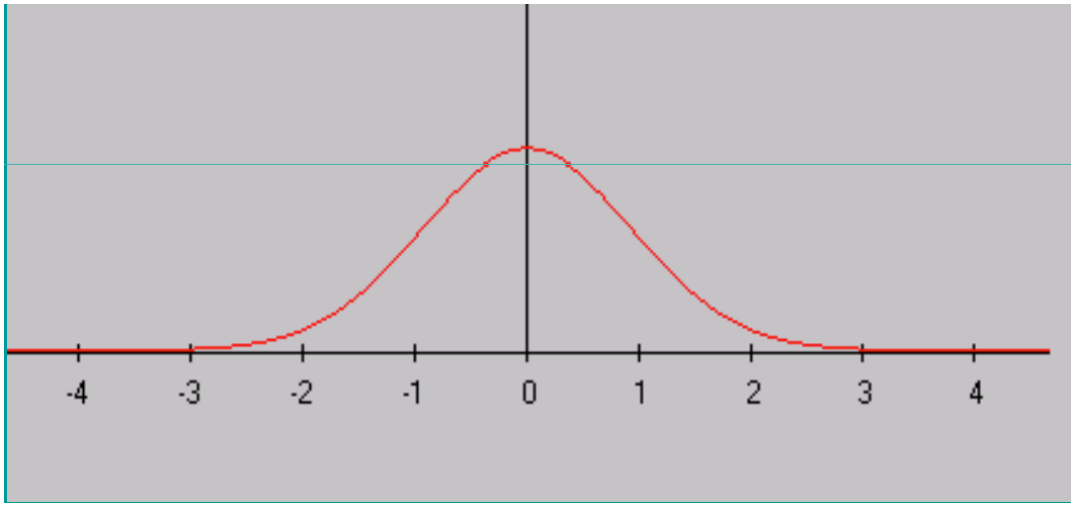






Normal distribution (Gaussian distribution) $N(\mu; \sigma^2)$

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



Normal distribution is defined by its expectation μ and variance σ^2

Normal distribution $N(\mu; \sigma^2)$

- Expected value:

$$E(X) = \mu$$

- Variance:

$$Var(X) = \sigma^2$$

- Standard deviation:

$$SD(X) = \sigma$$

Standard normal distribution N(0;1)

- Expected value:

$$E(Z) = 0$$

- Variance:

$$Var(Z) = 1$$

- Standard deviation:

$$SD(Z) = 1$$

Distribution transformation

- Linear transformation: Y = aX + b

  E(Y) = a.E(X) + b

  Var(Y) = a^2Var(X); SD(Y) = $|a|$.SD(X)

- Standardization transformation:

  Y = (X – E(X)) / SD(X)

  Then: E(Y) = 0; Var(Y) = 1; SD(Y) = 1

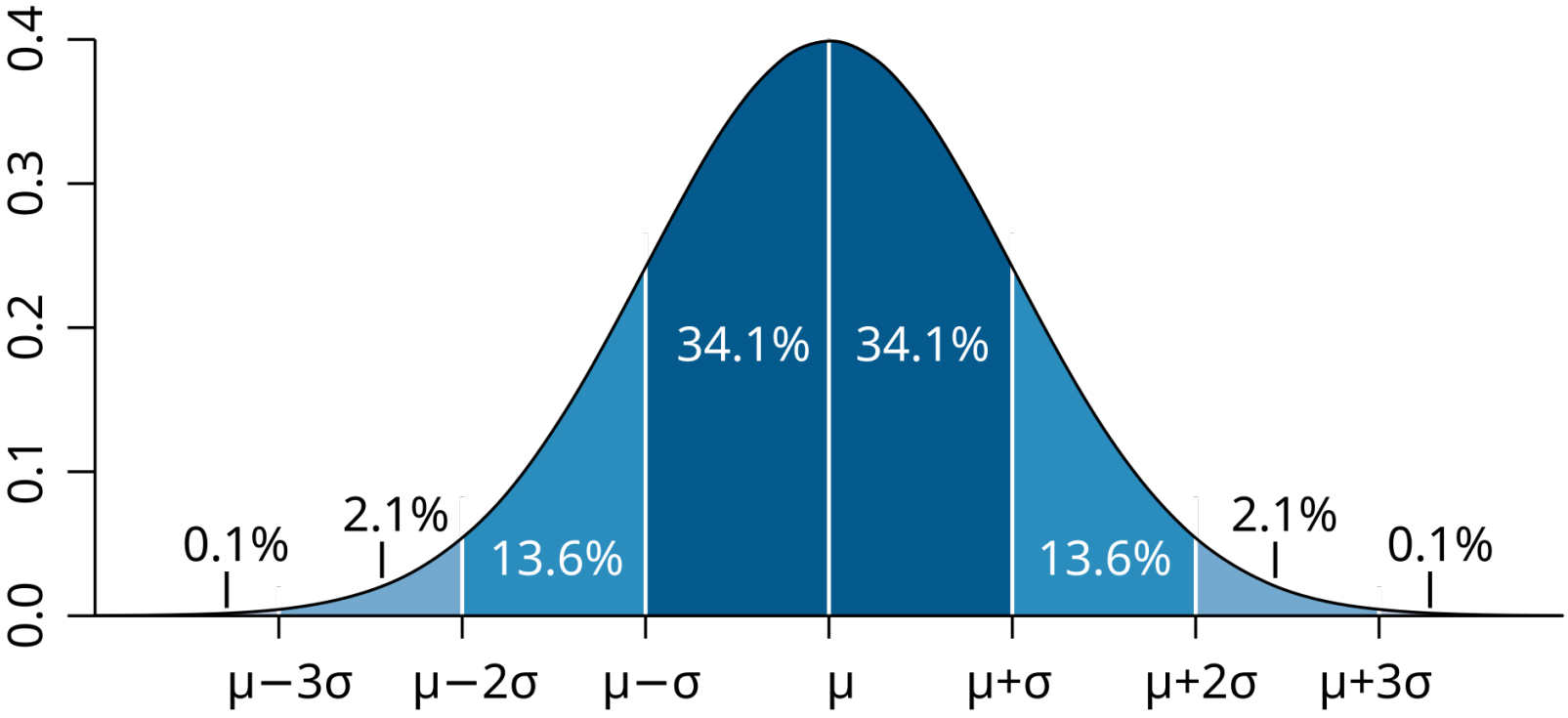
- Standardization of normal distribution $X \sim N(\mu; \sigma^2)$

  $Z = (X - \mu) / \sigma$

  Then E(Z) = 0; Var(Z) = 1; SD(Z) = 1

  Z is called standard normal distribution, $Z \sim N(0;1)$

Normal probability distribution



Using Excel to Compute Normal Probabilities

- Excel has two functions for computing cumulative probabilities and x values for any normal distribution:
- NOMDIST is used to compute the cumulative probability given an x value.
- NORMINV is used to compute the x value given a cumulative probability.

Confidence interval of estimation

- Confidence interval of an estimation is a segment on the number line that overlaps the estimated value of a parameter, indicating the true value of the parameter is likely to lie in that line with a given probability a (usually take a = 95% → 95% CI)

Confidence interval of estimation (interval estimation)

- Confidence interval of an 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-p)/n, then

  95% CI 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

- Proportion is a quantity with distribution approximate to Normal Distribution, we have:

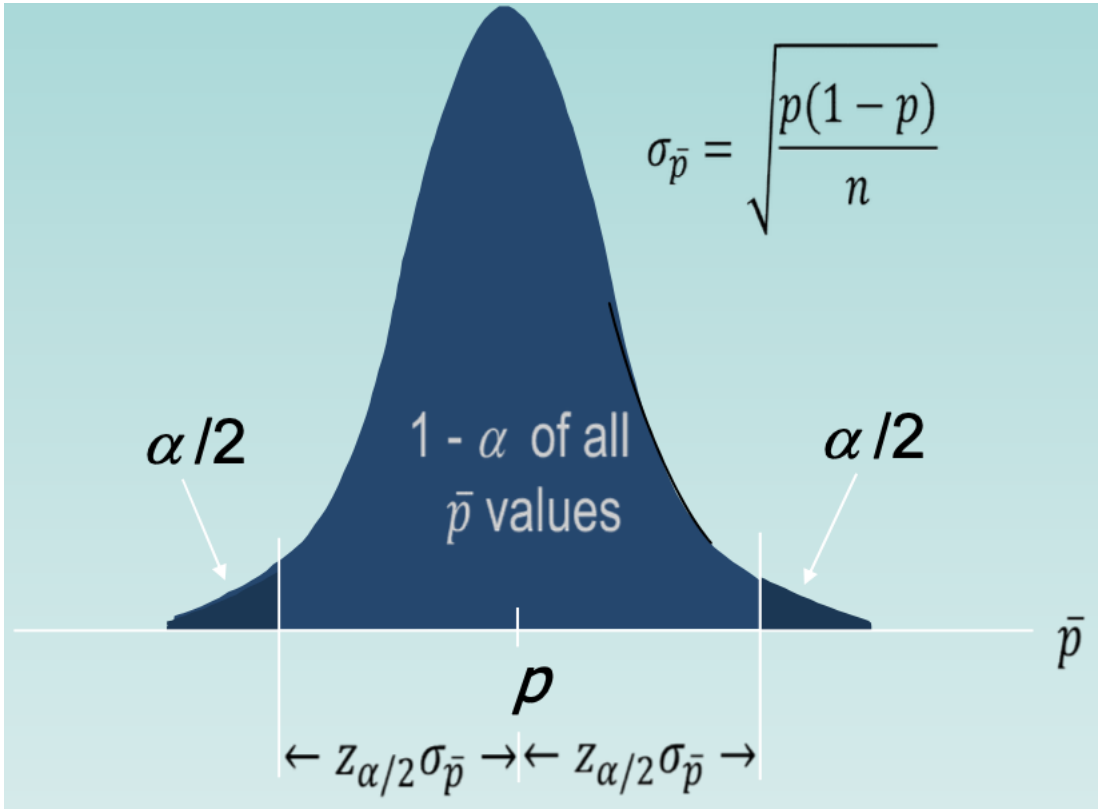
95% Confidence Interval of proportion estimation is

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

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

Interval Estimate of a Population Proportion

- Normal approximation of sampling distribution of $\bar{p}$



Interval Estimate of a Population Proportion

- Interval estimate

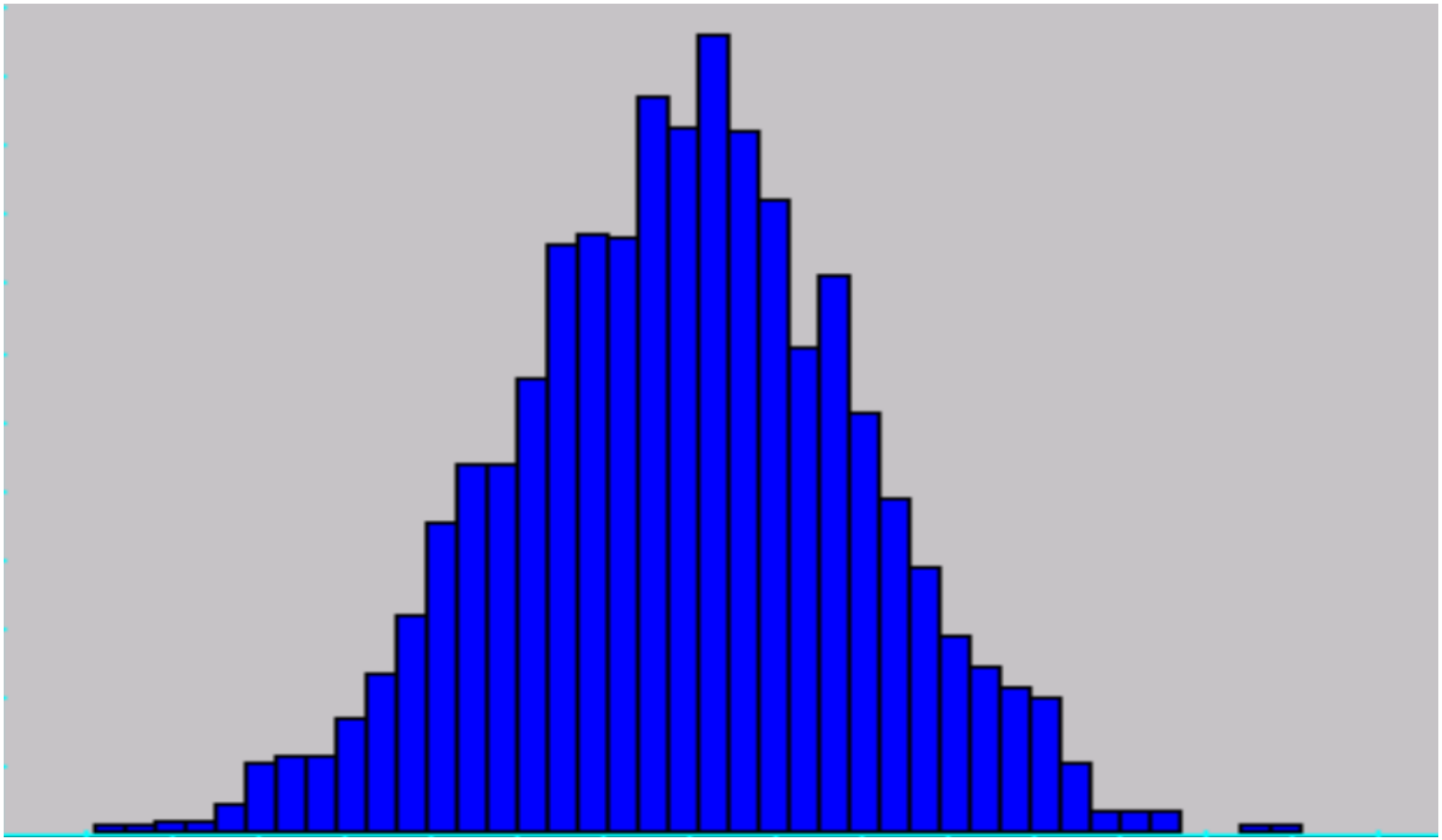
$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where: $1 - \alpha$ is the confidence coefficient,

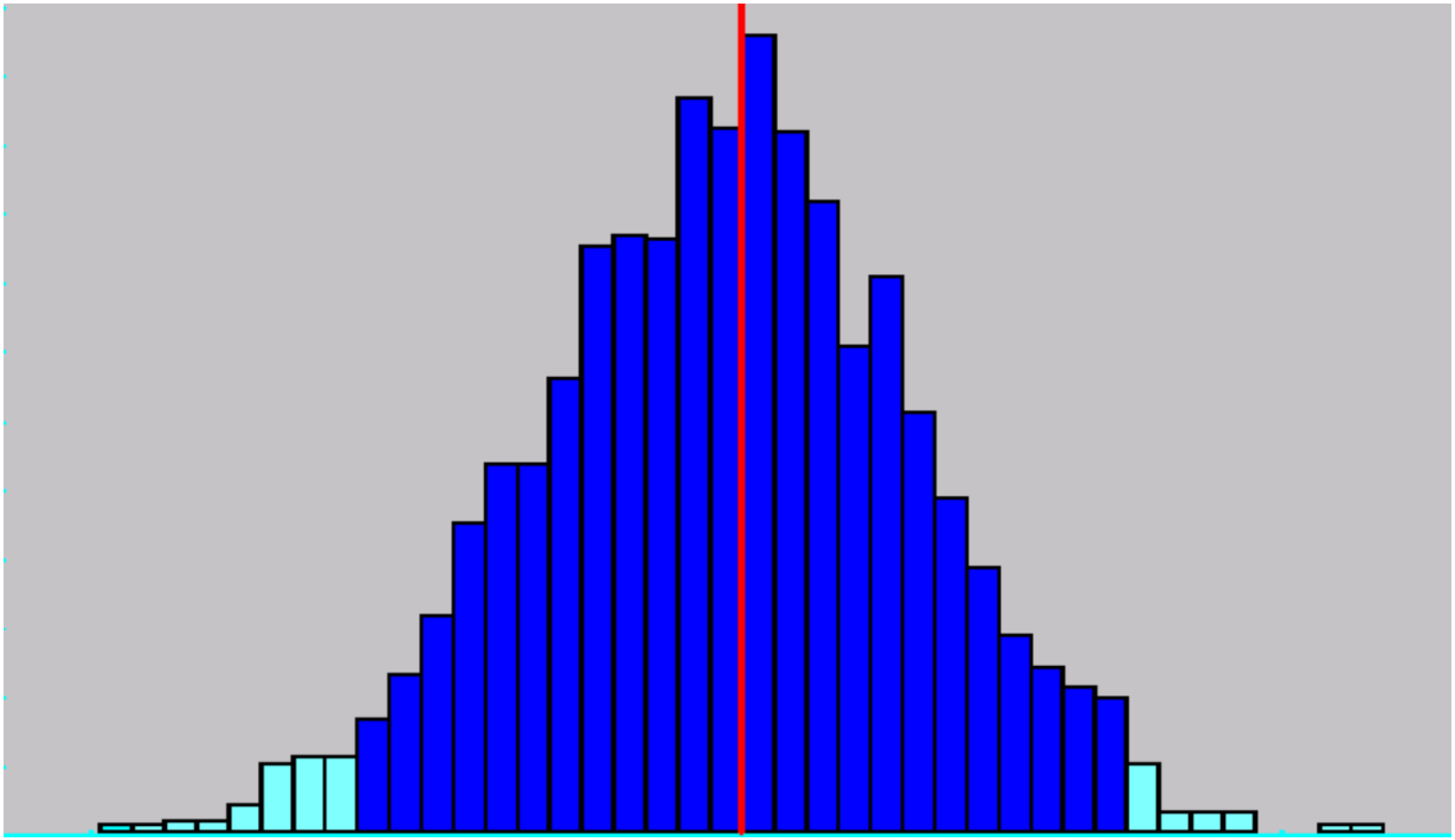
$z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution, and

$\bar{p}$ is the sample proportion

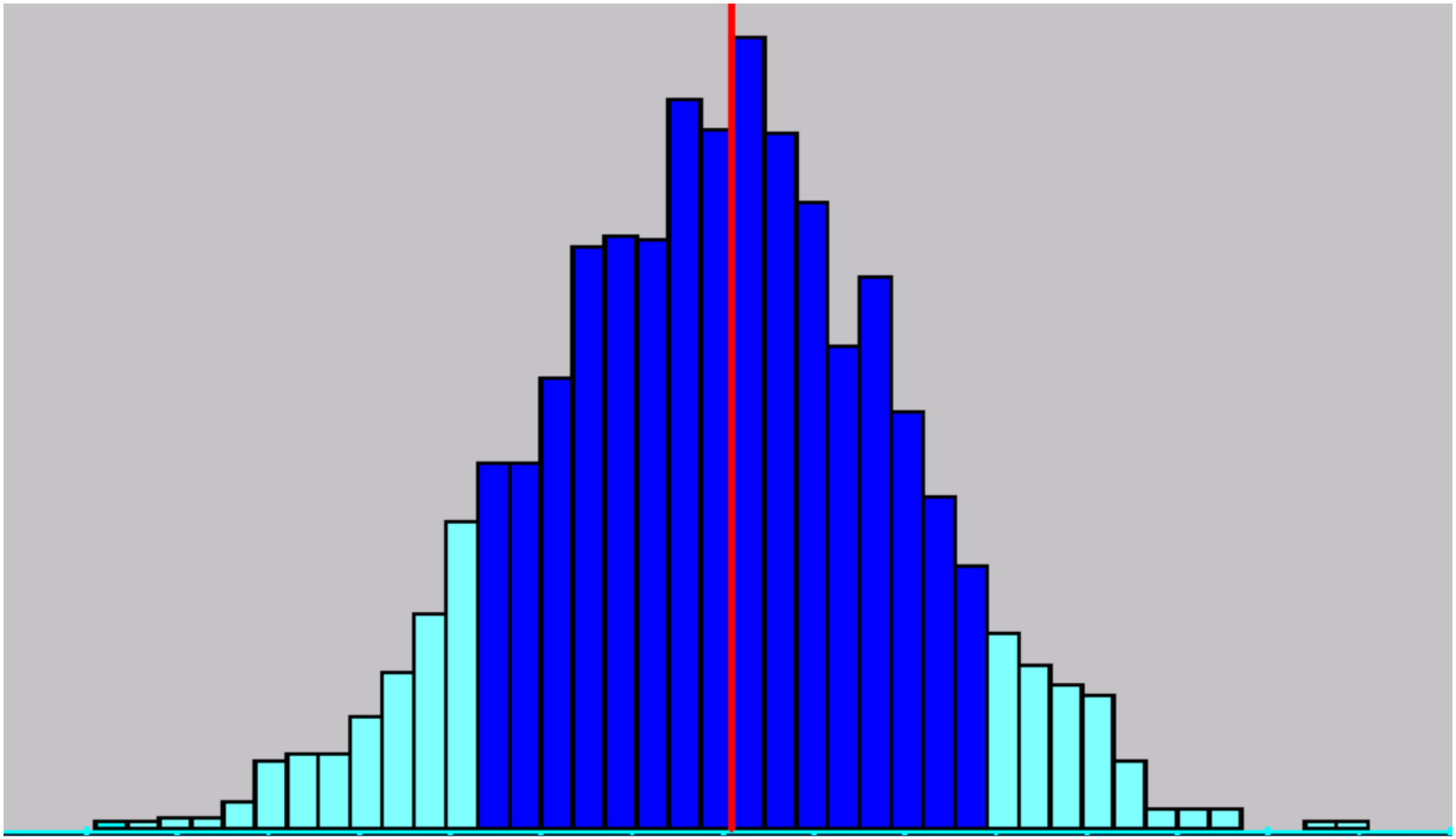
Normal distribution



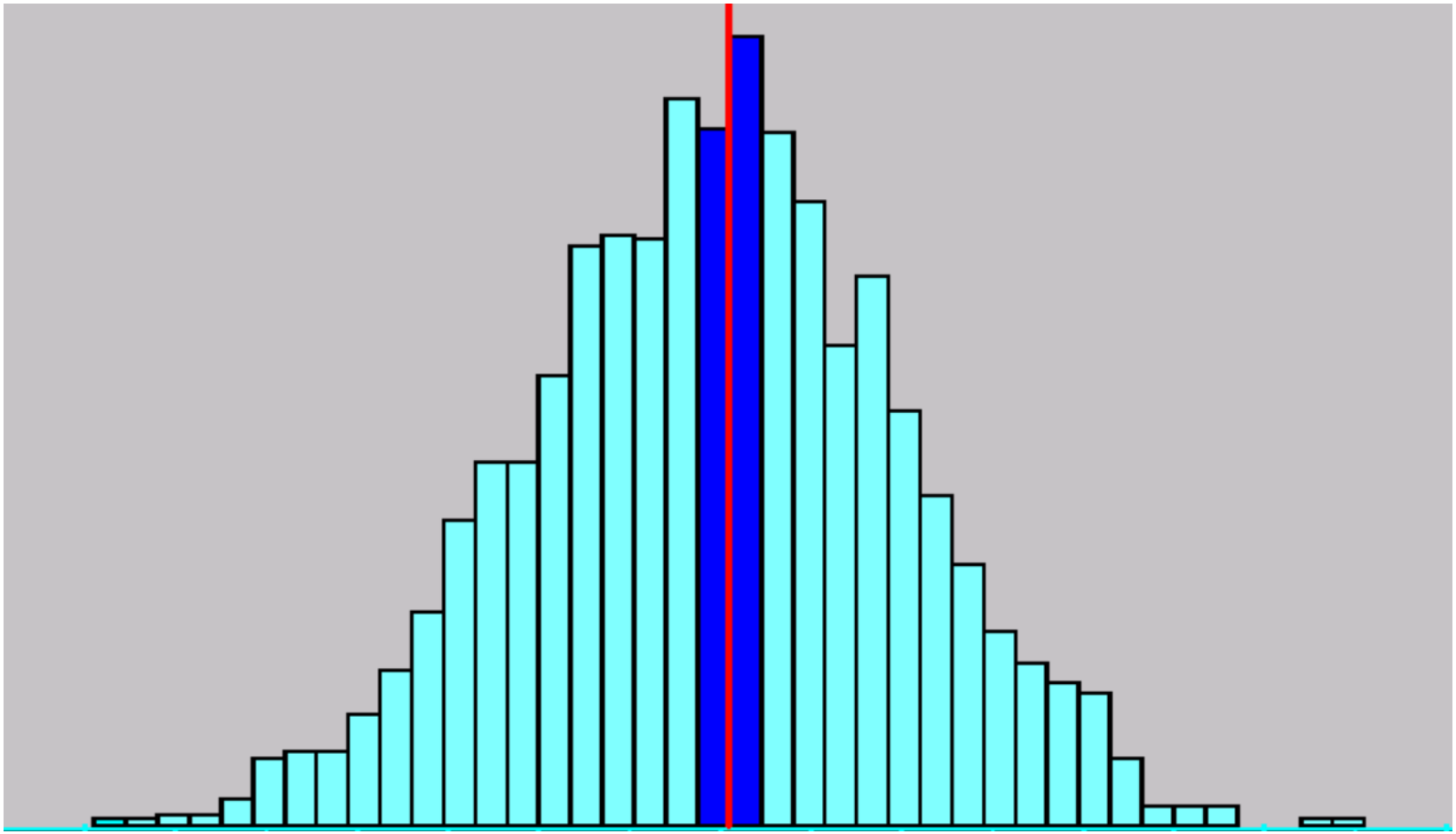
95% Confidence Interval



70% Confidence Interval



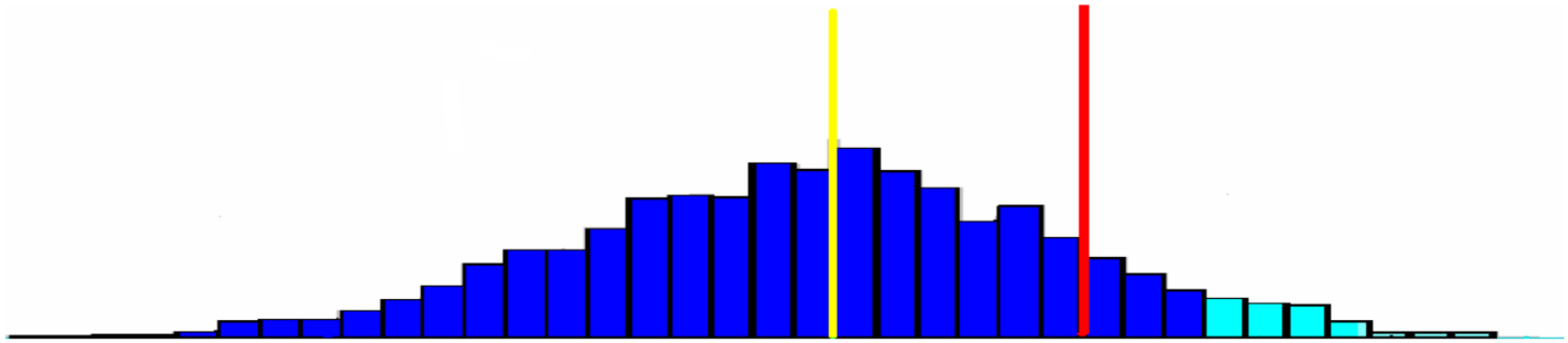
10% Confidence Interval



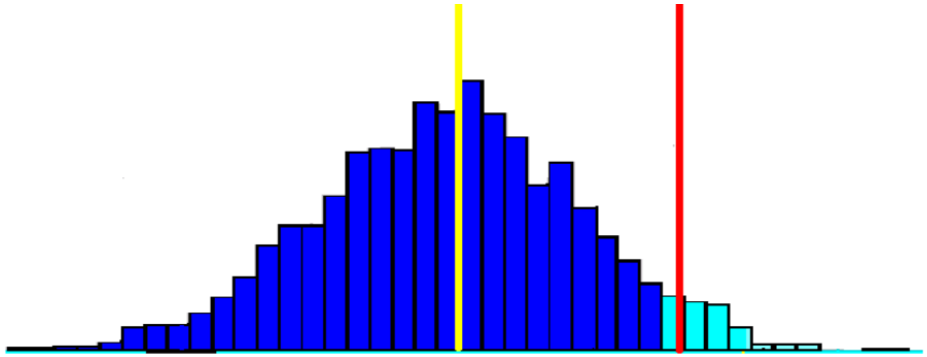
Sample size determining

- 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%

Sample size determining



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



Sample size for an interval estimate of a population proportion

- Margin of error

$$E = z_{\alpha/2}\sqrt{\frac{\bar{p}\,(1-\bar{p})}{n}}$$

- Solving for the necessary example size n, we get:

$$\mathrm{n} = \frac{(z_{\alpha/2})^2\ \bar{p}\,(1-\bar{p})}{E^2}$$

- However, $\bar{p}$ will not be known until after we have selected the sample. We will use the planning value p^* for $\bar{p}$

Estimation of expectation

- Expectation of variable = mean value of variable in the whole population
- To estimate the expectation of a quantitative variable X, a sample x(1), x(2), …, x(n) is selected 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

Estimation of expectation:

- Sample mean value is a "good" estimation of expectation if:

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

- The estimation is very close to the true value of expectation parameter if sample size n is very large

Estimation of expectation:

- 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 μ and variance σ^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 μ and variance equal σ^2 / n

Interval estimate of a population mean

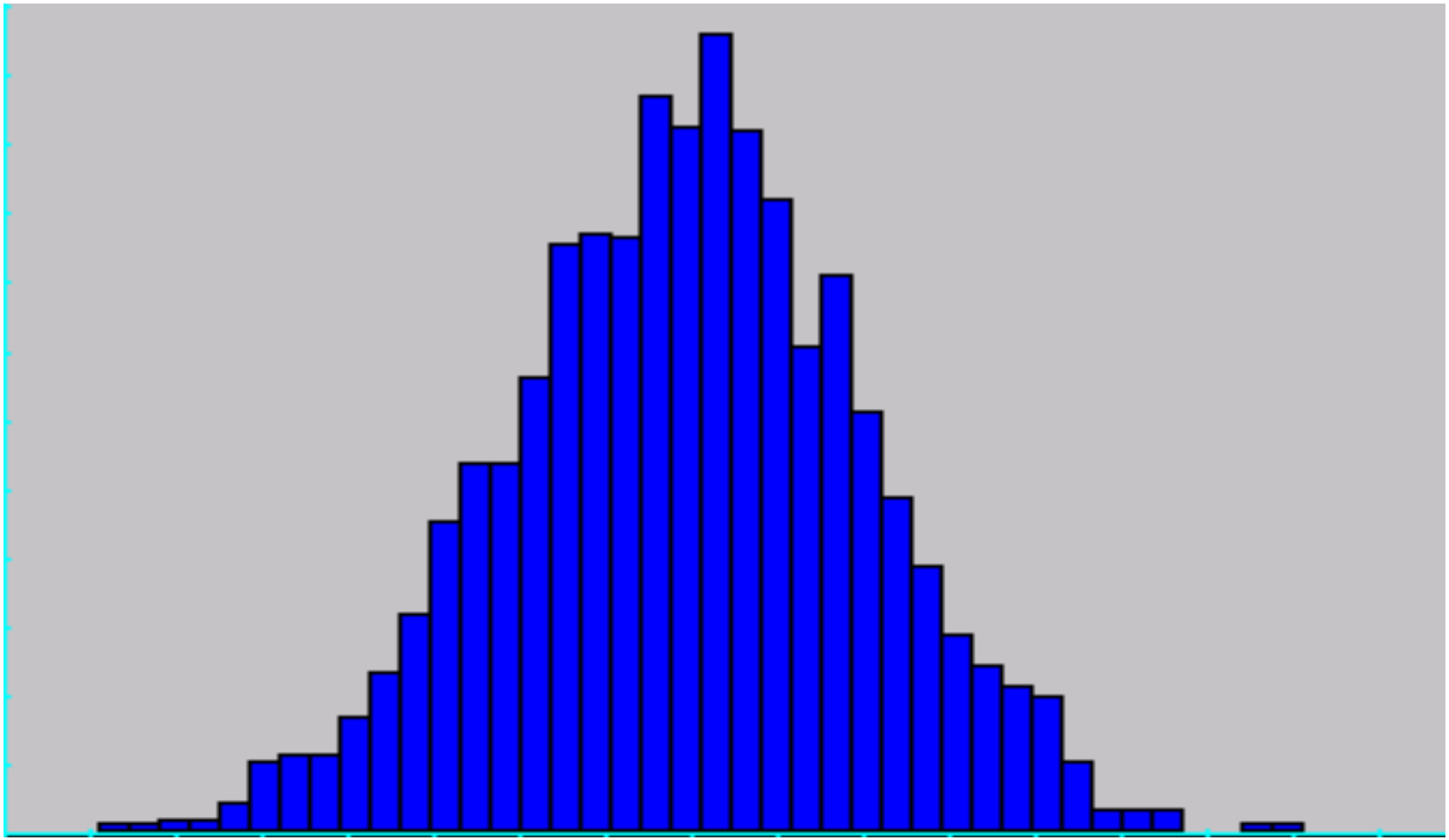
- In order to develop an interval estimate of a population mean, the margin of error must be computed using either:
  - The population standard deviation σ
  - Or the sample standard deviation s
- σ is rarely known exactly, but often a good estimate can be obtained based on historical data or other information
- We refer to such cases as the σ known case

Confidence interval of sample mean value

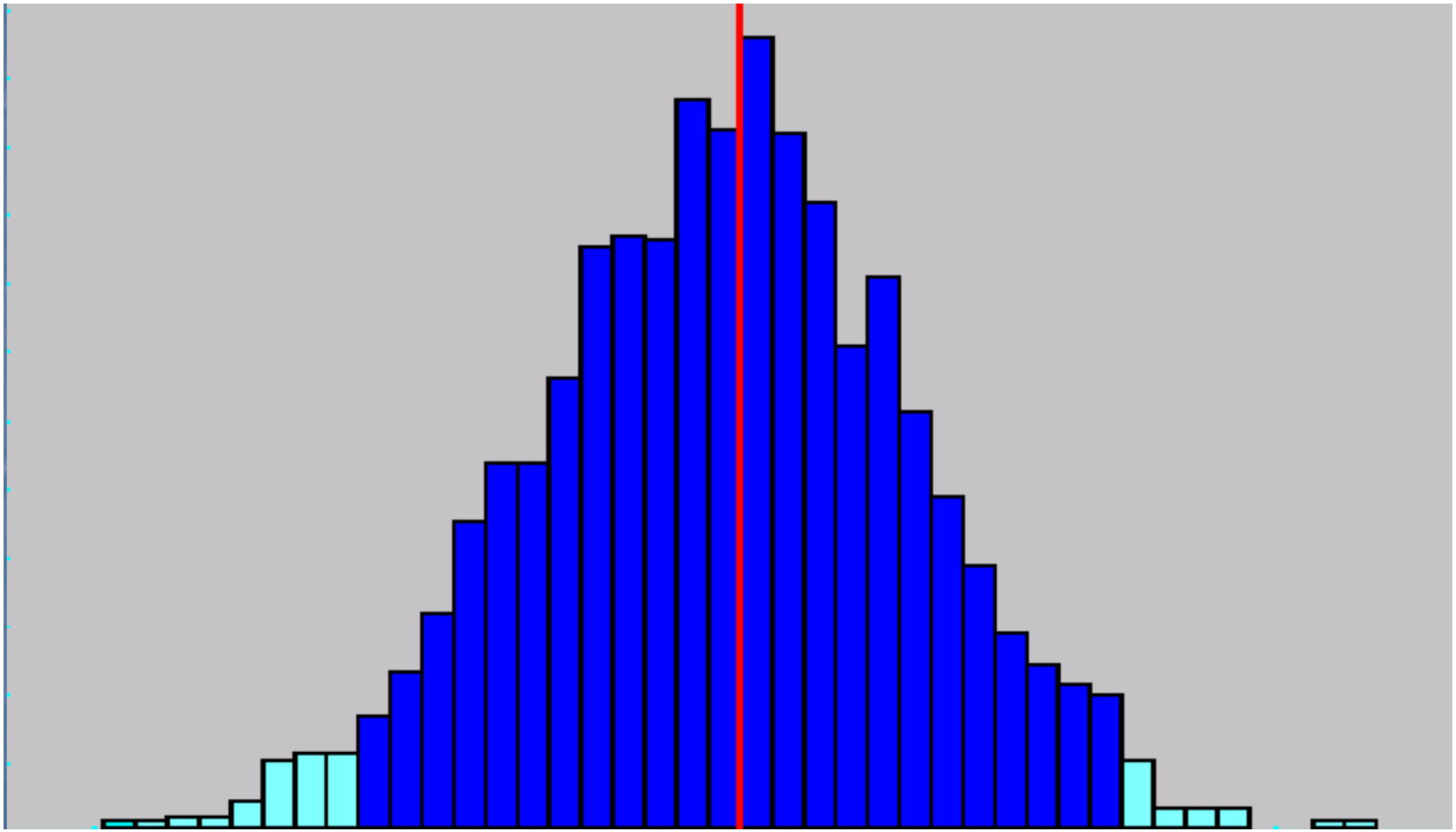
- Confidence interval of an estimation is an interval containing the estimated value, confirming the true value of the estimated parameter should be a point of that interval with a given probability a
- For a normal distributed estimation quantity with expectation $\bar{X}$ and variance σ^2/n, the 95% confidence interval (a = 95%) is defined by

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

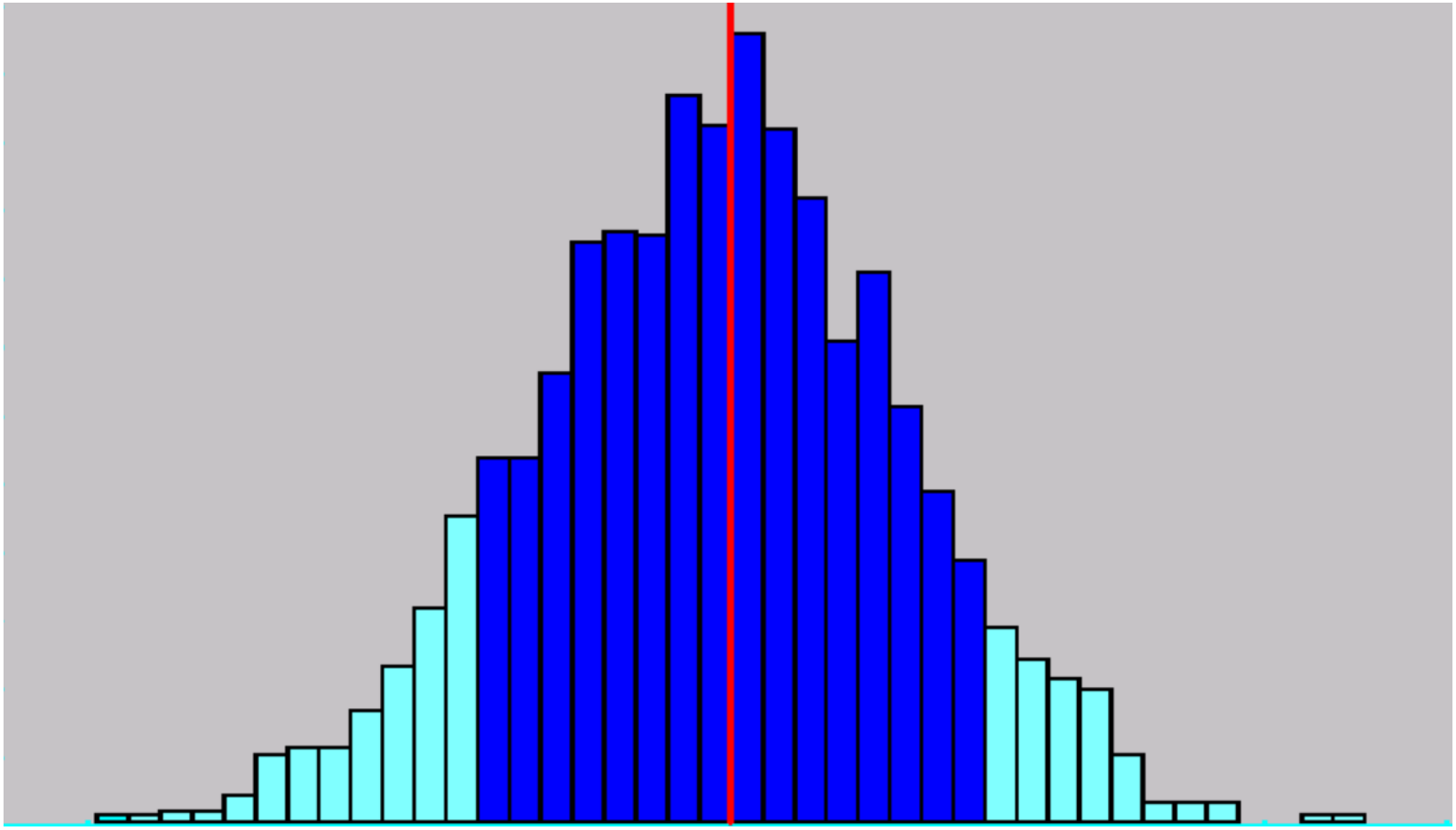
Normal distribution



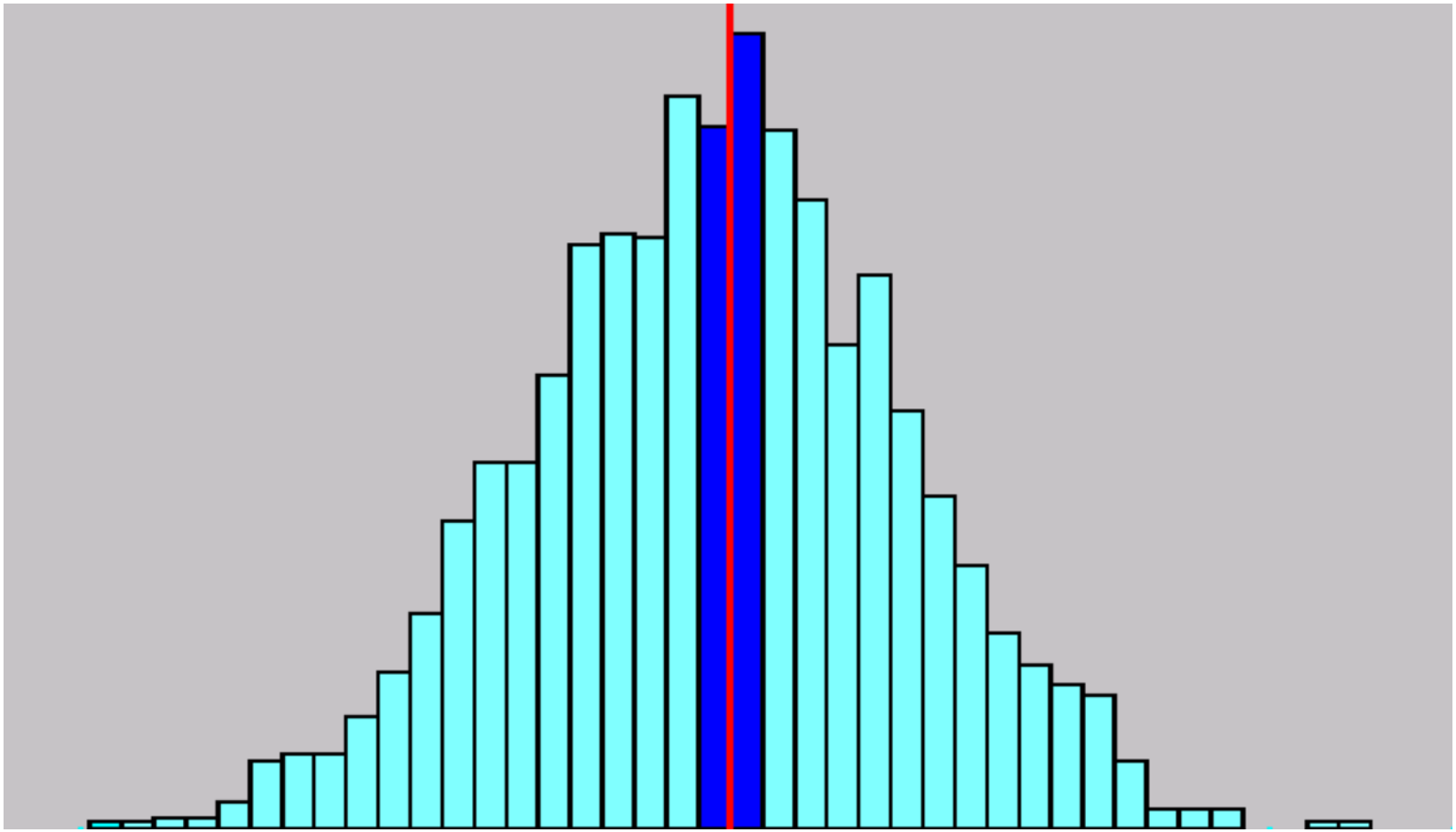
95% Confidence Interval



70% Confidence Interval

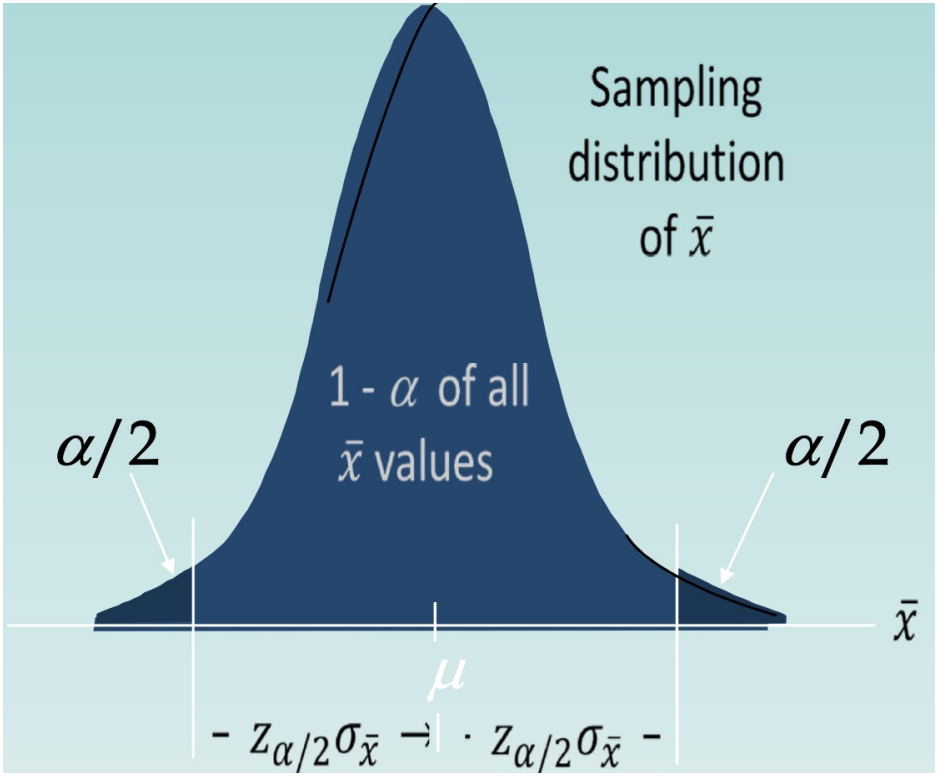


10% Confidence Interval

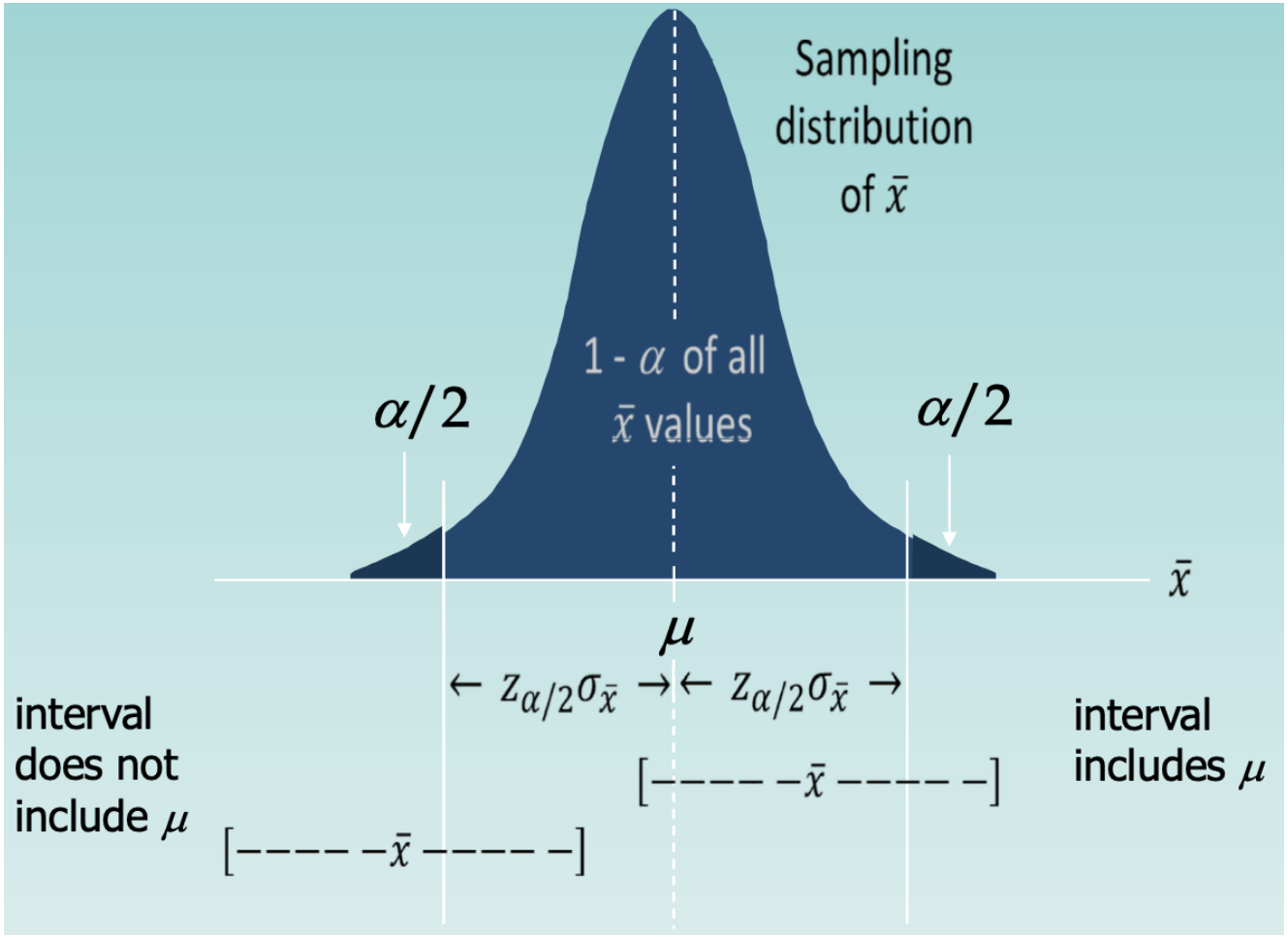


Interval Estimate of a Population Mean: σ Known

- There is a 1-α probability that the value of a sample mean will provide a margin of error of $z_{\alpha/2}\sigma_{\bar{x}}$ or less



Interval Estimate of a Population Mean: σ Known



Interval Estimate of a Population Mean: σ Known

- Interval estimate of μ:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where: $\bar{x}$ is the sample mean

$1 - \alpha$ is the confidence coefficient

$z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution

σ is the population standard deviation

n is the sample size

Interval Estimate of a Population Mean: σ Known

- Values of $z_{\alpha/2}$ for the most commonly used confidence levels:

Confidence Level	α	$\alpha/2$	Table Look-up Area	$z_{\alpha/2}$
90%	0.10	0.05	0.9500	1.645
95%	0.05	0.025	0.9750	1.960
99%	0.01	0.005	0.9950	2.576

Meaning of confidence

- Because 90% of all the intervals constructed using $\bar{x} \pm 1.645\sigma_{\bar{x}}$ will contain the population mean. We say that we are 90% confident that the interval $\bar{x} \pm 1.645\sigma_{\bar{x}}$ includes the population mean μ
- We say that this interval has been established at the 90% confidence level
- The value 0.90 is referred to as the confidence coefficient

Sample size for an interval estimate of a population mean

- Let E = the desired margin of error
- E is the amount added to and subtracted from the point estimate to obtain an interval estimate
- If a desired margin of error is selected prior to sampling, the sample size necessary to satisfy the margin of error can be determined

Sample size for an interval estimate of a population mean

- Margin of error:

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Necessary sample size:

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

Sample size for an interval estimate of a population mean

- The necessary sample size equation requires a value for the population standard deviation σ
- If σ is unknown, a preliminary or planning value for σ can be used in the equation.
  - Use the estimate of the population standard deviation computed in a previous study.
  - Use a pilot study to select a preliminary study and use the sample standard deviation from the study
  - Use judgment or a “best guess” for the value of s

Interval estimate of a population mean

- With a very large sample size n and finite population variance, the sample mean becomes a reliable estimator of the population mean (expectation), due to the Law of Large Numbers and the Central Limit Theorem.
- At 95% CI, the interval estimate for the population mean is given by:

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

where:

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

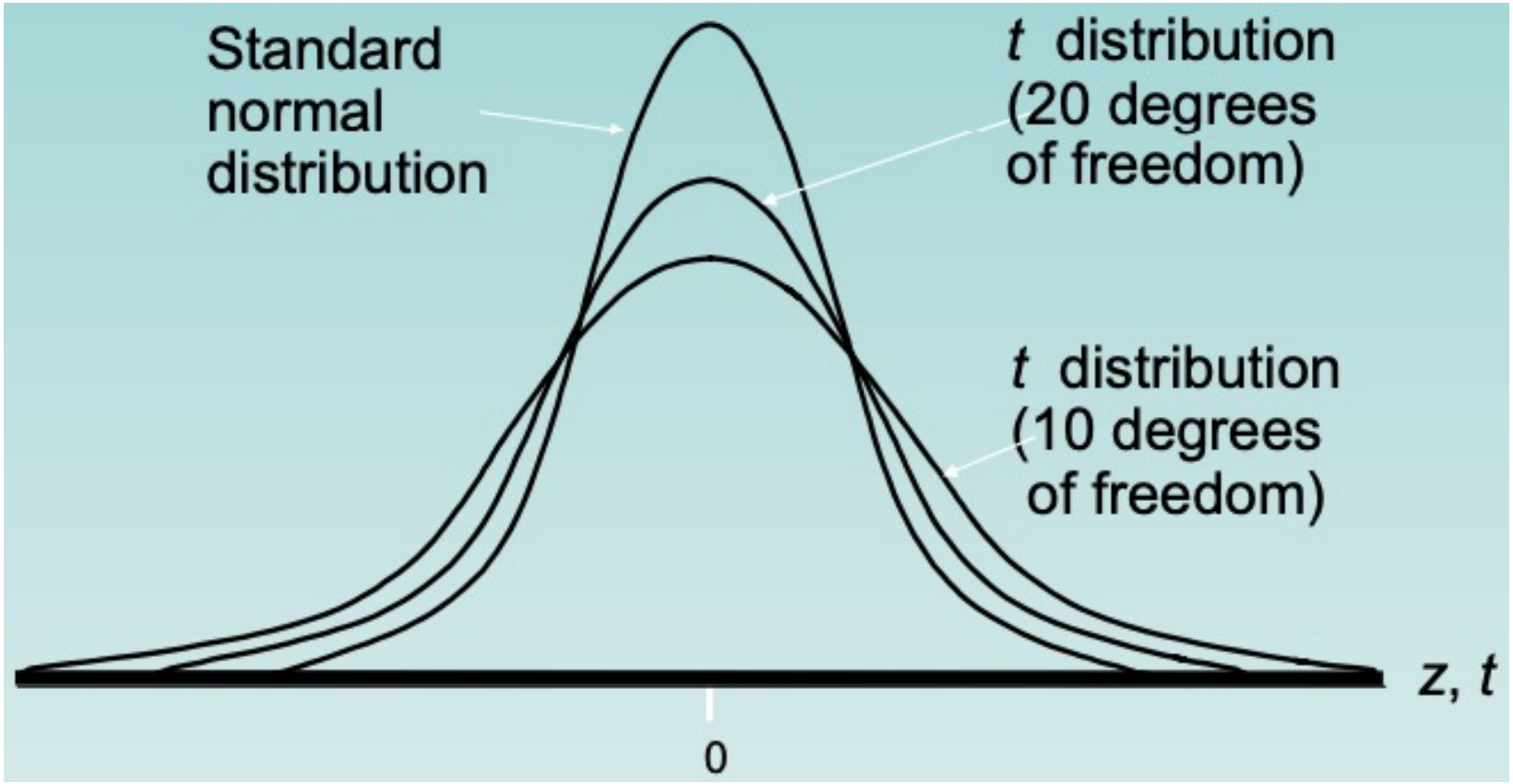
Interval Estimate of a Population Mean: σ unknown

- If an estimate of the population standard deviation σ cannot be developed prior to sampling, we use the sample standard deviation s to estimate σ
- This is the σ unknown case
- In this case, the internal estimate for μ is based on the t distribution

t-distribution

- The t-distribution is a family of similar probability distributions. A specific t-distribution depends on a parameter known as the degrees of freedom
- A t-distribution with more degrees of freedom has less dispersion
- As the degrees of freedom increases, the difference between the t-distribution and the standard normal probability distribution becomes smaller and smaller
- For more than 100 degrees of freedom, the standard normal z value provides a good approximation to the t value

t-distribution



Interval Estimate of a Population Mean: σ unknown

- At 95% confidence, $\alpha = 0.05$, and $\alpha/2 = 0.025$
- $t_{0.025}$ is based on n - 1 = 16 - 1 = 15 degrees of freedom

Degrees of Freedom	Area in Upper Tail					
	.20	.10	.05	.025	.01	.005
15	.866	1.341	1.753	2.131	2.602	2.947
16	.865	1.337	1.746	2.120	2.583	2.921
17	.863	1.333	1.740	2.110	2.567	2.898
18	.862	1.330	1.734	2.101	2.520	2.878
19	.861	1.328	1.729	2.093	2.539	2.861
.	.	.	.	.	.	.

Interval Estimate of a Population Mean: σ unknown

Degrees of Freedom	Area in Upper Tail					
	.20	.10	.05	.025	.01	.005
.	.	.	.	.	.	.
50	.849	1.299	1.676	2.009	2.403	2.678
60	.848	1.296	1.671	2.000	2.390	2.660
80	.846	1.292	1.664	1.990	2.374	2.639
100	.845	1.290	1.660	1.984	2.364	2.626
∞	.842	1.282	1.645	1.960	2.326	2.576

- Note: bottom row is standard normal z values

Using excel to compute t-distribution

- Excel has two functions for computing cumulative probabilities and x values for any t-distribution
  - T-DIST is used to compute the cumulative probability given an x value
  - T-INV is used to compute the x value given a cumulative probability

Interval Estimate of a Population Mean: σ unknown

- Interval estimate:

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where:

$\bar{x}$ = the sample mean

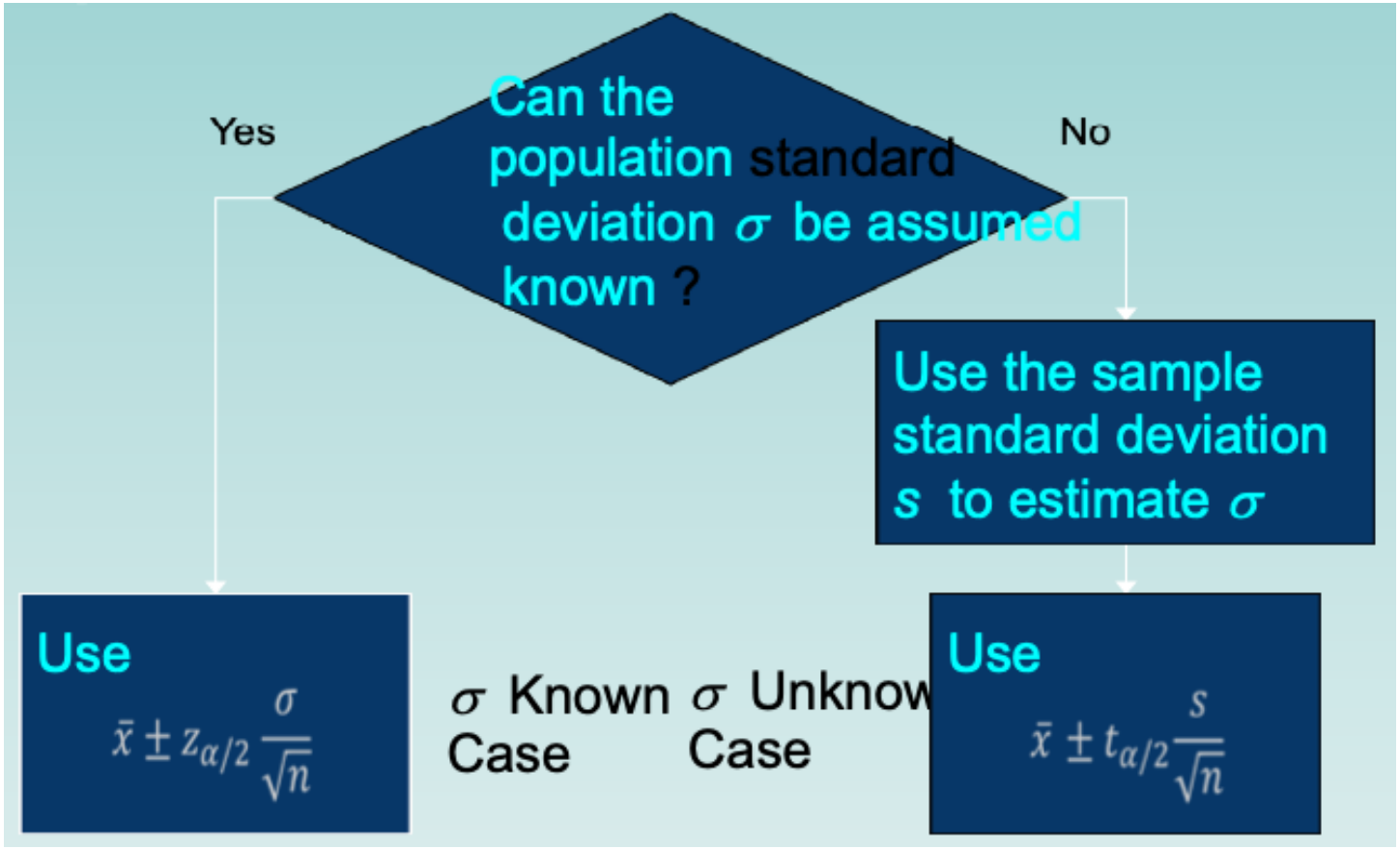
$1 - \alpha$ = the confidence coefficient

$t_{\alpha/2}$ = the t value providing an area of $\alpha/2$ in the upper tail of a t distribution with $n - 1$ degrees of freedom

s = the sample standard deviation

n = the sample size

Summary of Interval Estimation Procedures for a Population Mean



Exercise 1:

- 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 m_1 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

$$p = m_1 / N$$

Exercise 1:

- Problem: How to estimate the amount of fishes in a lake?
- Step 2: Nesting 2nd time to capture another amount n of fishes
  - Count the amount m_2 of marked fishes among n fishes captured in the 2nd time
  - Estimate the true proportion p of marked fishes by $p' = m_2 / n$ with 95% confidence interval

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

Exercise 1:

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

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

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

Exercise 1:

- Problem: How to estimate the amount of fishes in a lake?
- Step 3: We can be sure (with 95% certainty) that the amount of fishes in the lake should be a number between:

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

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

Exercise 2:

- Malnutrition rate of under 8 children counted 35% for the period 2010-2020.
- 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%)?

Exercise 2:

- 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:

$$30\% + 1.65 * (0.3 * 0.7)^{1/2} / n^{1/2} < 35\%$$

Then:

$$(0.3 * 0.7) / n < ((0.35-0.3)/1.65)^2$$

and n must be at least:

$$0.21 * 1.65 * 1.65 / 0.0025 \sim 235$$

→ Need at least 235 children in the sample

Exercise 3:

- In aquaculture, to determine the right moment for shrimp catching, the owner 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, if the shrimps weight is a quantity normally distributed with standard deviation equal 10 grams?

Exercise 3:

- Assume that the real average weight of shrimps in the lake is c, and the standard weight for catching 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 gram if the confidence interval contains the value b, therefore:

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

- Then n > 384