Hypothesis tests for two independent samples

- **Compare two proportions**
- **Compare mean values of two populations**
- **Compare two variances**

Problem 3. *Compare two mean values*

Let $(X_1, X_2, \ldots, X_n)$ be a sample of *n* independent observations from a variable X with expectation μ_1 and variance σ_1^2

$(Y_1, Y_2, \ldots, Y_m)$ be a sample of *m* independent observations from a variable Y with expectation μ_2 and variance σ_2^2

Problem: Compare two expectations μ_1 and μ_2 .

→ Estimate and compare two mean values $\overline{X}$ and $\overline{Y}$.

The problem can be solved by using the following Theorem:

Theorem. *Let* $(X_1, X_2, \ldots, X_n)$ *and* $(Y_1, Y_2, \ldots, Y_m)$ *be two samples of n independent observations selected correspondingly from a variable X with sample mean* $\overline{X}$ *and sample variance* S_X^2 *and from a variable Y with sample mean* $\overline{Y}$ *and sample variance* S_Y^2 *(both variables are normal distributed). Then the (new) variable*

$$t = \sqrt{\frac{n.m}{n+m}} . \sqrt{\frac{n+m-2}{n.S_X^2 + m.S_Y^2}} . (\overline{X} - \overline{Y})$$

has Student distribution with (n+m-2) degrees of freedom.

Hypothesis Tests

A. Two-tail Test: Hypothesis

$$H: Mean(X) = Mean(Y)$$

Alternative Hypothesis

$$K: Mean(X) \text{ differs from } Mean(Y)$$

B. Right one-tail Test: Hypothesis

$$H: Mean(X) = Mean(Y)$$

Alternative Hypothesis $K: Mean(X) > Mean(Y)$

C. Left one-tail Test: Hypothesis

$$H: Mean(X) = Mean(Y)$$

Alternative Hypothesis $K: Mean(X) < Mean(Y)$

Steps of testing

Step 1. Estimate sample mean values $Mean(X)$, $Mean(Y)$ and sample variances $Var(X)$, $Var(Y)$

Step 2. Calculating perform the quantity

$$t = \sqrt{\frac{n.m}{n+m}}.\sqrt{\frac{n+m-2}{n.Var(X)+m.Var(Y)}}.(Mean(X)-Mean(Y))$$

Step 3 (Version A- Computer). Taking a variable ***T(n+m-2)*** of Student distribution with (n + m - 2) degrees of freedom calculate the probability

$$b = P\ \{\ |T(n+m-2)|\ >=\ |\ t\ |\ \}$$

(for 2-tails test); or

$$b = P\ \{\ T(n+m-2)\ >=\ t\ \}$$

(for right 1-tail test); or

$$b = P\ \{\ T(n+m-2)\ =<\ t\ \}$$

(for left 1-tail test, then **t < 0**)

Step 4. Compare the probability ***b*** with a given ahead significance level ***alpha*** (=5%, 1%, 0.5% or 0.1%):

+ If ***b >= alpha*** → accept Hypothesis **H** and conclude ***Mean(X) = Mean(Y)***

+ If ***b < alpha*** → reject Hypothesis **H** and confirm ***Mean(X) khác Mean(Y)*** (for 2-tails test); or

Mean(X) > Mean(Y) (for right 1-tail test); or

Mean(X) < Mean(Y) (for left 1-tail test)

Version B. Using Student distribution table

Looking in Table of Student distribution find out critical value $T(n+m-2,alpha/2)$ of Student distribution with $n+m-2$ degrees of freedom ($alpha$ is a given ahead significance level =5%, 1% or 0.5%)

Decide

- Reject Hypothesis **H:** = if

$$t > T(n+m-2,alpha/2)$$

- Accept Hypothesis **H:** = if

$$t =< T(n+m-2,alpha/2)$$

Version C. Using confidence intervals

When degree of freedom (sample size) is large, Student distribution approximates Normal distribution. Then we can use confidence intervals (with significance level of 5%) for testing:

$$\left[Mean(X) - 1.96 * \sqrt{Var(X)/n}; Mean(X) + 1.96 * \sqrt{Var(X)/n} \right]$$

$$\left[Mean(Y) - 1.96 * \sqrt{Var(Y)/m}; Mean(Y) + 1.96 * \sqrt{Var(Y)/m} \right]$$

Decide

Reject Hypothesis **H: =** **if** the two intervals disjoin

Accept Hypothesis **H: =** **if** the two intervals have nonempty intersection

SPSS

Test 4. *Compare two independent samples - Mann-Whitney non-parametric Test*

Test 3 is powerful under assumption of **Normal distribution** of variables ***X*** and ***Y*** , or sample sizes ***n*** and ***m*** are large (>40). Without the above assumption we must use **"non-parametric"** methods

Mann-Whitney Test is a non-parametric test comparing 2 independent samples with **Hypothesis**

H: two variables X and Y have common distribution
(two samples have been selected from a homogeneous population)

and **Alternative Hypothesis**

K: distributions of X and Y are different
(two sample have been selected from different populations)

Non-parametric tests are based on comparing **ranks** of values of concerned variables instead of comparing directly the values of variables.

Definition. *Given a sequence* $a_1, a_2, \ldots, a_n$ *of numbers. Let the sequence be reordered into increasing sequence* $a_{k_1} \le a_{k_2} \le \ldots \le a_{k_n}$
Then rank ***h(.)*** *of elements in the original sequence is defined as the follows:*

$$h(a_{k_p}) = p \quad \text{if} \quad a_{k_{p-1}} < a_{k_p} < a_{k_{p+1}}$$

$$h(a_{k_p}) = \frac{(2r + s)}{2}$$

$$\text{if} \quad a_{k_{r-1}} < a_{k_r} = a_{k_{r+1}} = \ldots = a_{k_p} = \ldots = a_{k_{r+s}} < a_{k_{r+s+1}}$$

Weight Stem-and-Leaf Plot

Frequency	Stem	&	Leaf	
3.00	9	.	005	3
3.00	10	.	003	6
6.00	11	.	005578	12
10.00	12	.	0000005688	22
12.00	13	.	000000000001	34
14.00	14	.	00000035557889	48
14.00	15	.	00000000004559	62
17.00	16	.	00000000445555567	79
18.00	17	.	00000000000245559	97
21.00	18	.	000000000000255556678	118
18.00	19	.	00000000055555778	136
33.00	20	.	000000000000000000002233345566799	140
21.00	21	.	000000000000055555555	107
28.00	22	.	0000000000000000002555566778	86
25.00	23	.	0000000000000000000002459	58
21.00	24	.	000000000000000002448	33
12.00	25	.	000000000000	12

Procedure of Testing

Step 1:

Put together two sample $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_m)$ into a common sequence of $(n + m)$ numbers,

Determine the rank of each element in that sequence and calculate the ranks sum of each sample:

$$R_1 = \sum_{i=1}^{n} h(X_i) \quad \text{(sum of ranks in the first sample)}$$

$$R_2 = \sum_{j=1}^{n} h(Y_j) \quad \text{(sum of ranks in the second sample)}$$

Step 1 (continued):

Determine the rank statistics:

$$U_1 = n.m + \frac{n(n+2)}{2} - R_1$$

$$U_2 = n.m + \frac{m(m+2)}{2} - R_2$$

$$U = min(U_1, U_2) \quad ; \quad \bar{U} = \frac{n.m}{2}$$

$$S_U = \sqrt{\frac{n.m.(n+m-1)}{12}}$$

LEMMA. Suppose $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_m)$ be independent samples from two continous varables X and Y. Suppose that hypothesis H is true. Then variable U has distribution converging very fast to the Normal distribution $N(\overline{U}, S_U^2)$, therefore the distribtion of the variable

$$u = \frac{U - \overline{U}}{S_U}$$

converges very fast to the standard Normal distribution $N(0,1)$.

REMARK. In the above Lemma, to conclude that distributions of U and u are close to normal distributions it is enough to have the sample sizes **greater than 8**.

Steps of Hypothesis testing

Step1. Determine rank of each element in both samples and the quantity u as presented above;

Step2. Taking a variable $N(0,1)$ with standard normal distribution (normal distribution with expectancy 0 and variance 1) canculate the probability

$$b = P\{|N(0,1)| > |u|\}$$

Step 3. Compare the probability *b* with a given ahead significance level *alpha* :

* If *b* > *alpha* → accept hypothesis *H* and consider two variables *X* , *Y* as those have the same distribution, i.e. both samples were selected from a common homogeneous population

* If *b* <= *alpha* → reject hypothesis *H* and conclude *X* , *Y* are truly different, i.e. the two samples were taken from two different sources

Remark

In the above, T – tests are used for comparing **mean values** and are valid if sample size are large (> 40) or the condition of Normal distribution are fulfilled

The non-parametric Mann-Whitney test is used to compare **two medians**, is applicable even when there is no assumption of Normal distribtion and sample sizes are not very large. When the sample size are large the non-parametric and T tests are equivalent

Test 7. *Compare two variances*

Variance represents precision of a measure or of an estimation. The **smaller variance** corresponds the **more accurate** measure. Therefore the evaluation of measure's accuracy can be done by comparing variances. The comparison can be processed by assess ratio of two variances.

Testing problem

Let $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_m)$ be samples taken from two Normal variables $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$.

Hypothesis H: $\sigma_1^2 = \sigma_2^2$

Alternative hypothesis K: $\sigma_1^2 \neq \sigma_2^2$

Steps of testing process

Step 1. Estimate sample variances S_X^2, S_Y^2 and perfom the ratio

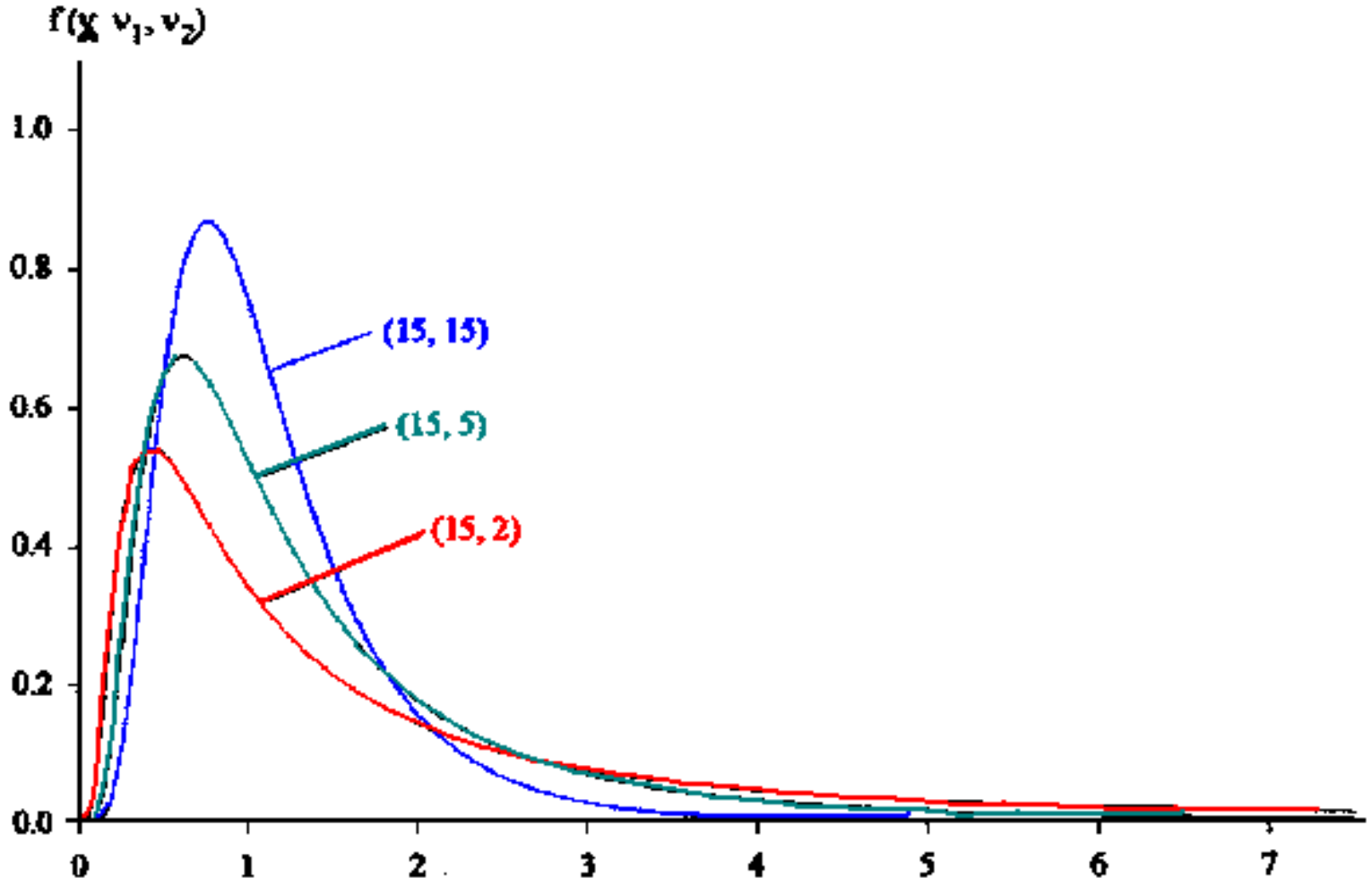
$$F = \frac{S_X^2 (n-1)}{S_Y^2 (m-1)} \quad \text{if } S_X^2 > S_Y^2$$

or

$$F = \frac{S_Y^2 (m-1)}{S_X^2 (n-1)} \quad \text{if } S_Y^2 > S_X^2$$

LEMMA. Suppose $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_m)$ be independent samples from two Normal distributed varables X and Y. Suppose that hypothesis H is true. Then the ratio F is a variable with Fisher-Snedecordistribution of n and m degree of freedom (for the first case) or m and n degree of freedom (for the second case).

Fisher (F) distribution



Parameter of Fisher distribution is “degree of freedom” (ν_1, ν_2)

By virtue of the above Lemma we can continue testing process:

Step 2. For the first case taking a variable $FS(n-1, m-1)$ which have Fisher wíth $(n-1, m-1)$ degree of freedom (for the second case the procedure is similar with degree of freedom in the invers order) calculate the probability

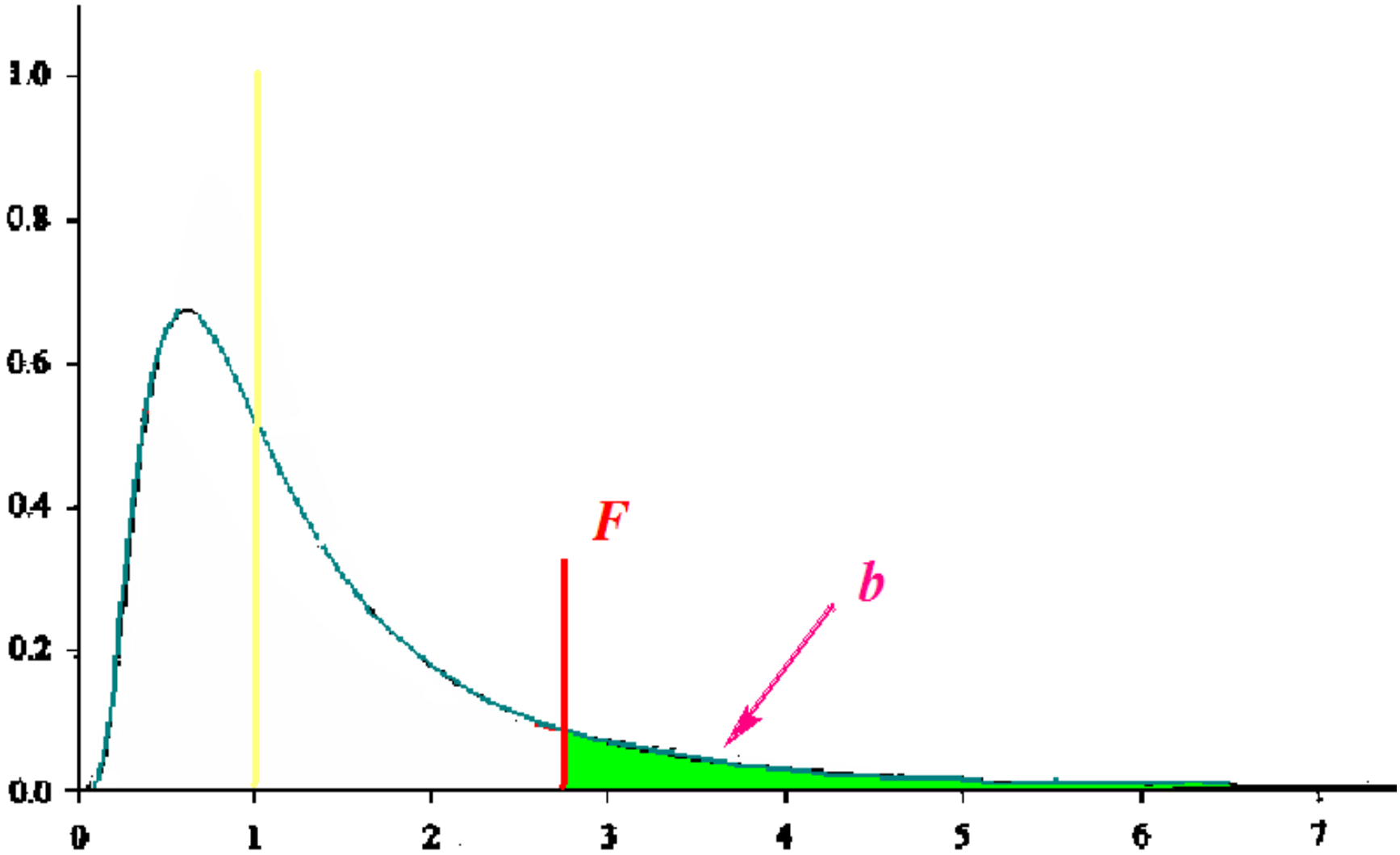
$$b = P\{ FS(n-1, m-1) > F \}$$

Step 3. Compare the probability b with a given ahead significance level $alpha$

* If $b > alpha$ → accept Hypothesis H , conclude the equality of two variances

* If $b <= alpha$ → reject Hypothesis H , confirm the difference of two variances

Fisher distribution



SPSS

Compare two proportions – the case of large sample sizes (using Normal distribution)

Let $(X_1, X_2, \ldots, X_n)$ be a sample of a binary variable X taking value 1 with probability p_1 and value 0 with probability $(1 - p_1)$, $(Y_1, Y_2, \ldots, Y_m)$ be a sample of a binary variable Y taking value 1 with probability p_2 and value 0 with probability $(1 - p_2)$;

$p_1, p_2 \in (0,1)$.

Consider the Hypothesis H: $p_1 = p_2$

and Alternative Hypothesis K: $p_1 \neq p_2$

Note. Variable X has expectation p_1 and variance $p_1(1 - p_1)$.

Variable Y has expectation p_2 and variance $p_2(1 - p_2)$.

Therefore we can treat the testing problem as a special problem of comparing two mean values (expectations) p_1 and p_2.

If the Hypothesis H is true then use the two samples $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_m)$ as samples collected from one variable and estimate the common variance of X and Y by

$$\frac{m_1 + m_2}{n_1 + n_2}.\left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right) = \frac{m_1 + m_2}{n_1 + n_2}.\frac{n_1 + n_2 - m_1 - m_2}{n_1 + n_2}$$

then perform a statistic

$$u = \left(\frac{m_1}{n_1} - \frac{m_2}{n_2}\right) \Big/ \left[\sqrt{\frac{m_1 + m_2}{n_1 + n_2}.\frac{n_1 + n_2 - m_1 - m_2}{n_1 + n_2}}.\sqrt{\frac{n_1 + n_2}{n_1.n_2}}\right]$$

for testing, where m_1 and m_2 respectively are the numbers of values 1 appeared in the above two samples.

By Central Limit Theorem, when sample sizes are large, the difference Mean(X) - Mean(Y) has a distribution very close to Normal distribution. Then the testing procedure can be as follows:

Step 1. Calculate value of statistic

$$u = \left(\frac{m_1}{n_1} - \frac{m_2}{n_2} \right) \Big/ \left[\sqrt{\frac{m_1 + m_2}{n_1 + n_2} \cdot \frac{n_1 + n_2 - m_1 - m_2}{n_1 + n_2}} \cdot \sqrt{\frac{n_1 + n_2}{n_1 . n_2}} \right]$$

Step 2. Taking Normal distribution N(0,1) find the probability

$$b = P\{ |N(0,1)| > |u| \}$$

Step 3. Compare the probability **b** to a given ahead significance alpha

* If **b** > **alpha** → Accept Hypothesis **H** , confirm the equality of two proportions

* If **b** <= **alpha** → Reject Hypothesis **H** and conclude two proportions to be different

(One-tail tests can be done similarly)

Version B. Using Normal distribution table

Looking in Table of Normal distribution find out critical value *u(alpha/2)* of Normal distribution (for *alpha* = 5% the critical value equals 1.96)

Decide

- Reject Hypothesis **H:** = if

u >= u(alpha/2)

- Accept Hypothesis **H:** = if

u < u(alpha/2)

Version C. Using confidence intervals

Use confidence intervals (with significance level of 5%) of estimated proportions for testing:

$$\left[\frac{m_1}{n_1} - 1.96 * \sqrt{\frac{m_1}{n_1}\left(1 - \frac{m_1}{n_1}\right) / n_1}\,;\, \frac{m_1}{n_1} + 1.96 * \sqrt{\frac{m_1}{n_1}\left(1 - \frac{m_1}{n_1}\right) / n_1}\right]$$

$$\left[\frac{m_2}{n_2} - 1.96 * \sqrt{\frac{m_2}{n_2}\left(1 - \frac{m_2}{n_2}\right) / n_2}\,;\, \frac{m_2}{n_2} + 1.96 * \sqrt{\frac{m_2}{n_2}\left(1 - \frac{m_2}{n_2}\right) / n_2}\right]$$

Decide

Reject Hypothesis **H: = if** the two intervals disjoin

Accept Hypothesis **H: = if** the two intervals have nonempty intersection

SPSS

Compare several proportions

Let X be a binary variable taking two values 0 and 1 . Collecting data from that variable under k different conditions we have a sample containing k groups of observations related with the conditions

Let $p_1, p_2, \ldots, p_k$

be probabilities of appearance of value 1 of variable X under each of the above k conditions.

Hypothesis H : $p_1 = p_2 = \ldots = p_k$

Alternative Hypothesis

K: *there is certain difference between* $p_1, p_2, \ldots, p_k$

Data: Perform a *2*x*k* table of *2* rows and *k* columns: each column for one group, the 1rst row for value 1, the 2nd row for value 0 of the variable at observations:

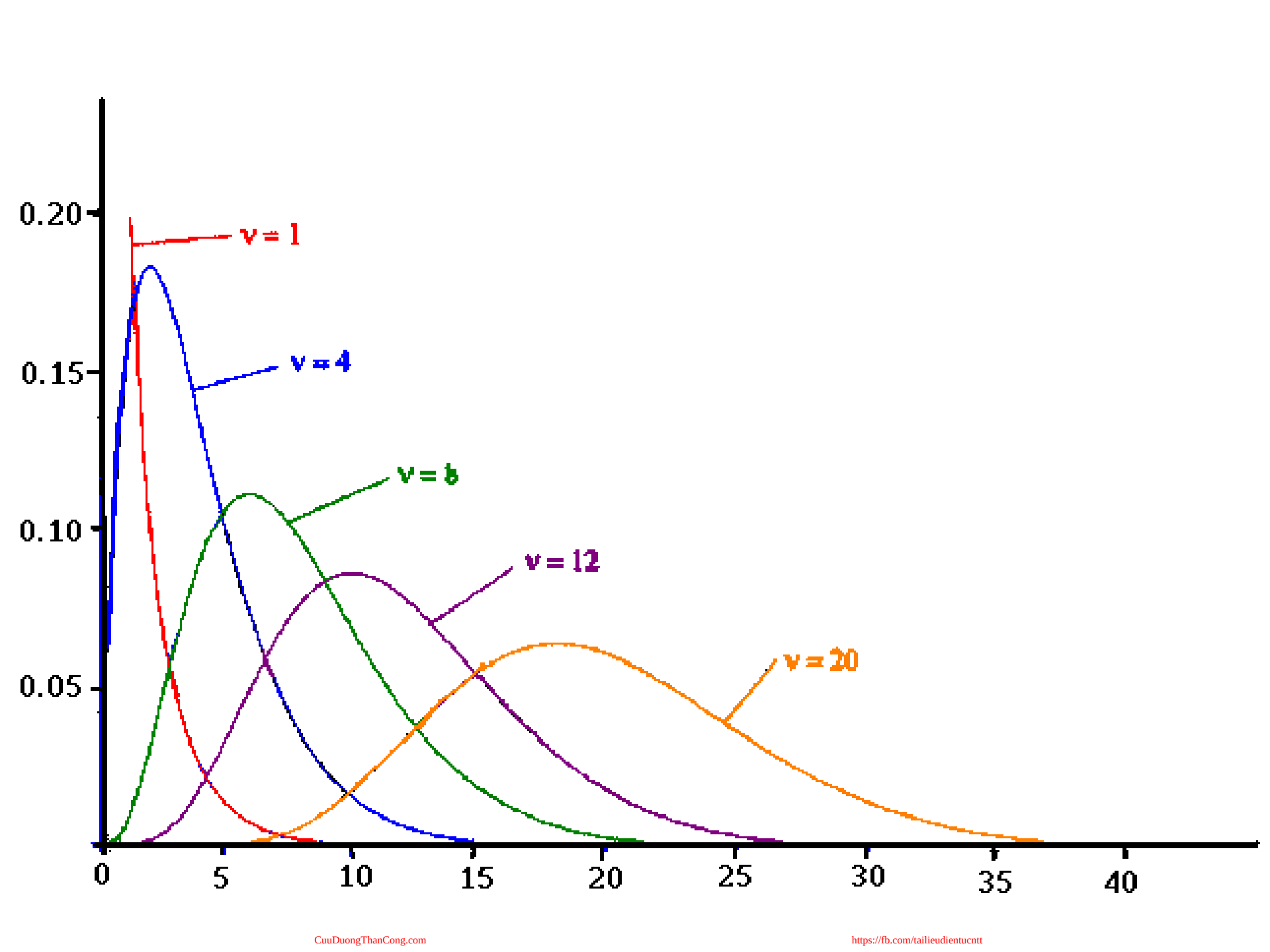
	Group 1	Group2		Group k	
X = 1	n_{11}	n_{12}	. . .	n_{1k}	n_1
X = 0	n_{01}	n_{02}	. . .	n_{0k}	n_0
	$n^{(1)}$	$n^{(2)}$		$n^{(k)}$	n

$$n_1 = n_{11} + n_{12} + \ldots + n_{1k} \quad ; \quad n_0 = n_{01} + n_{02} + \ldots + n_{0k}$$

$$n^{(j)} = n_{j1} + n_{j0} \quad ; \quad j = 1, 2, \ldots, k \quad ; \quad n = n_0 + n_1$$

$$\chi^2 = \sum_{j=1}^{k} \sum_{i=0}^{1} (n_{ij} - \frac{n_i . n^{(j)}}{n})^2 / (\frac{n_i . n^{(j)}}{n})$$

LEMMA. Suppose that hypothesis H is true. Then variable χ^2 has distribution approximate to the Chi-square distribution with $(k-1)$ degrees of freedom $\chi^2_{(k-1)}$.



Note.

When degree of freedom tends to infinity, the **Chi-square distribution** converge to **Normal distribution!**

Version A (computer):

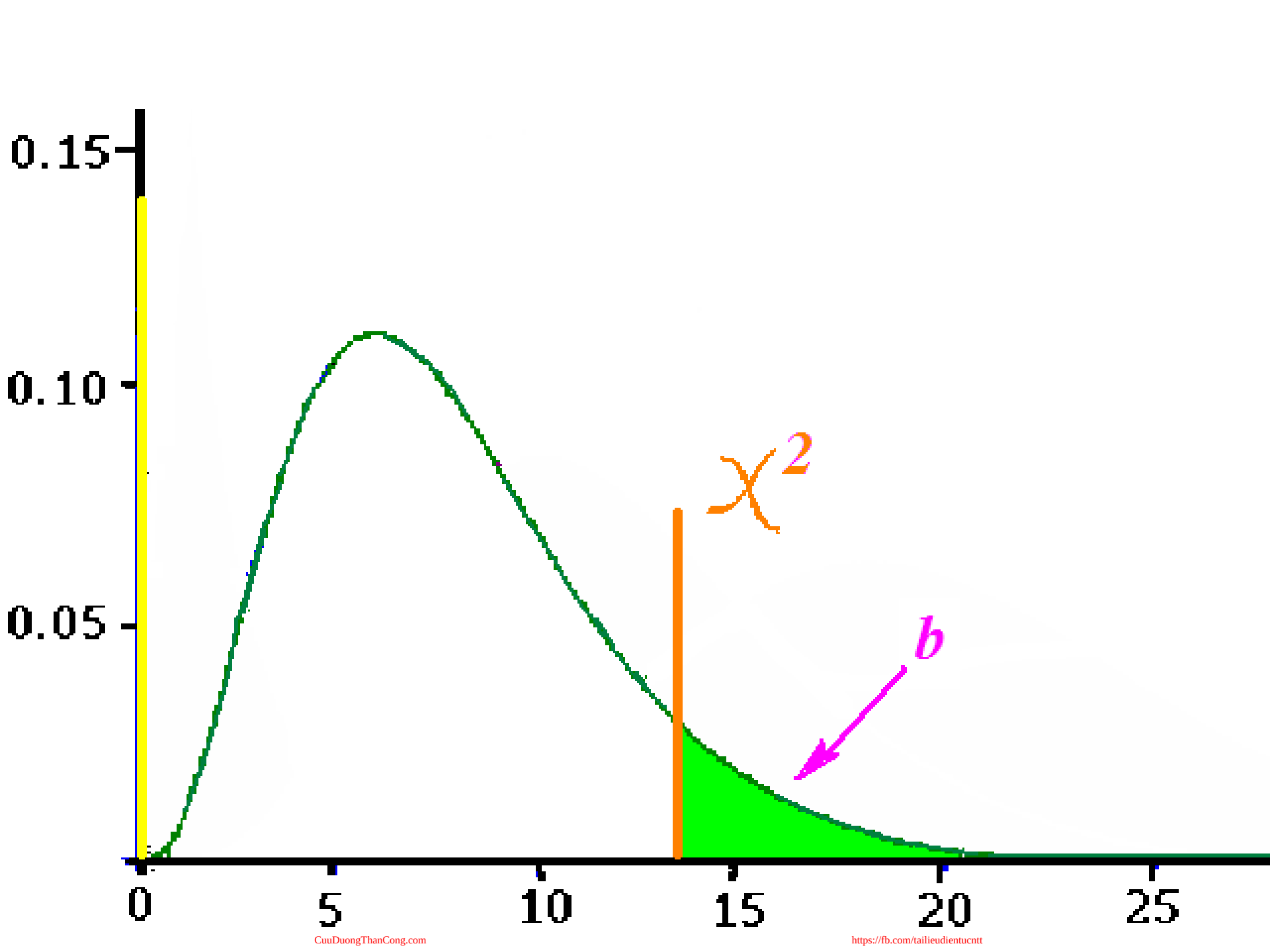
Step 1. Taking a variable **CS(k-1)** of Chi-square distribution with (k-1) degrees of freedom calculate the probability

$$b = P\{ CS(k-1) > \chi^2 \}.$$

Step 2. Compare the probability ***b*** to the given ahead significance level ***alpha*** :

* If ***b > alpha*** → accept hypothesis ***H*** , conclude the all proportions are equal

* If ***b <= alpha*** → reject hypothesis ***H*** , confirm the appearance of some difference between proportions.



Version B. Using distribution table

Looking in Table of Chi-square distribution to find out critical value $\chi^2_{(k-1)}(alpha)$ of Chi-square distribution with *k-1* degrees of freedom (*alpha* is a given ahead significance level =5%, 1% or 0.5%)

Decide

- Reject Hypothesis **H:** = if

$$\chi^2 \geq \chi^2_{(k-1)}(alpha)$$

- Accept Hypothesis **H:** = if

$$\chi^2 < \chi^2_{(k-1)}(alpha)$$

SPSS