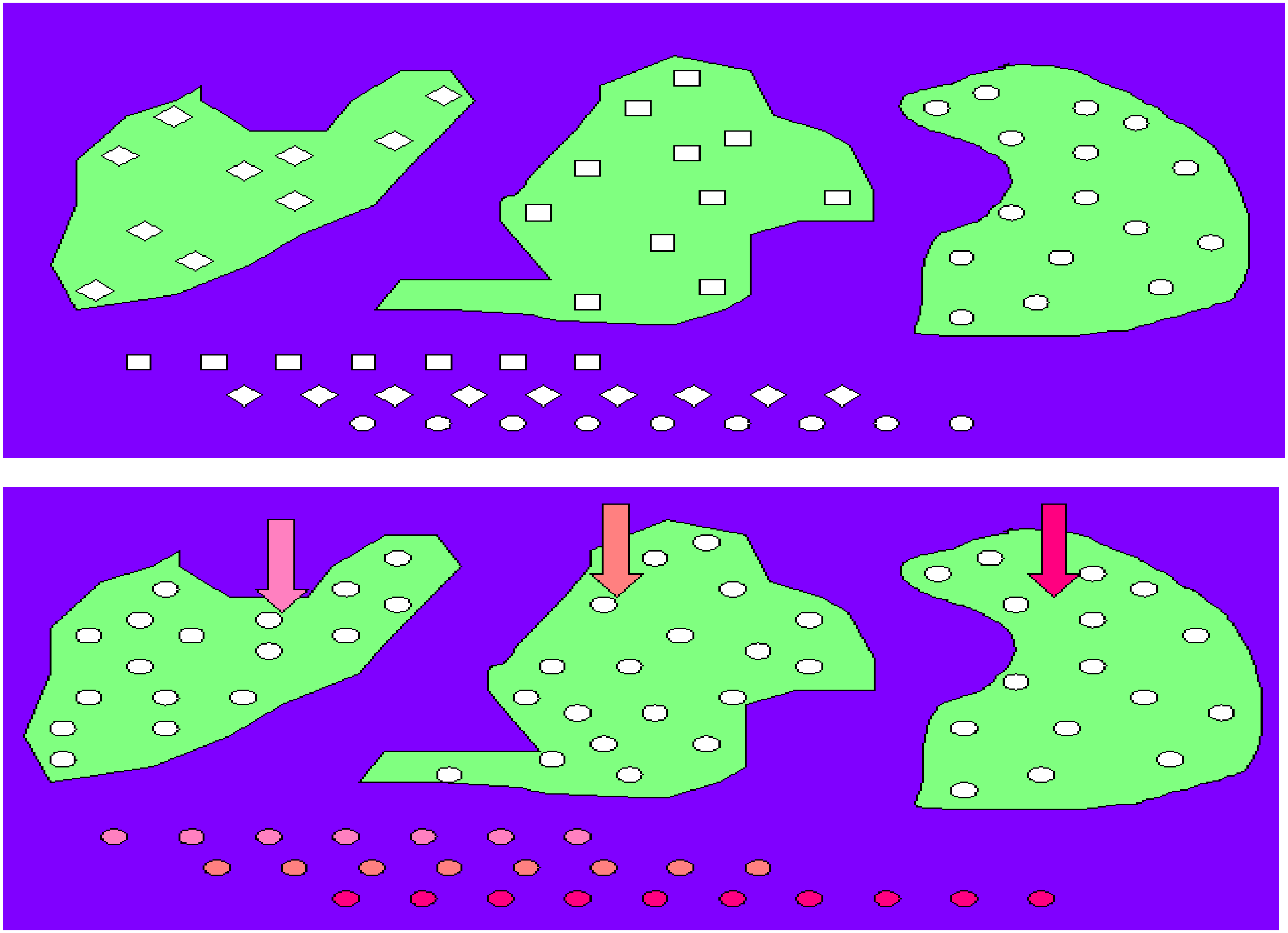
Model of multi-independent samples



Hypothesis tests for several independent samples

- **Compare several proportions**

- **Compare mean values of several populations**

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

$$p_1, p_2, \dots, p_k$$

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

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

Alternative Hypothesis

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

Data: Perform a *2xk* 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:

Table 1. Observed frequency

	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

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

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

Compare several proportions

- If the hypothesis is correct, the proportion of occurrence of 1 estimated commonly to all columns (conditions) is equal to

$$n_1 / n$$

- The proportion of occurrence of 0 estimated commonly to all columns is equal to

$$n_0 / n$$

Perform the table of expected (theoretical) frequencies of the hypothesis:

Table 2. predicted (expected) frequency

	Group 1	Group2		Group k	
X = 1	$n^{(1)} * \frac{n_1}{n}$	$n^{(2)} * \frac{n_1}{n}$	. . .	$n^{(k)} * \frac{n_1}{n}$	n_1
X = 0	$n^{(1)} * \frac{n_0}{n}$	$n^{(2)} * \frac{n_0}{n}$	. . .	$n^{(k)} * \frac{n_0}{n}$	n_0
	$n^{(1)}$	$n^{(2)}$		$n^{(k)}$	n

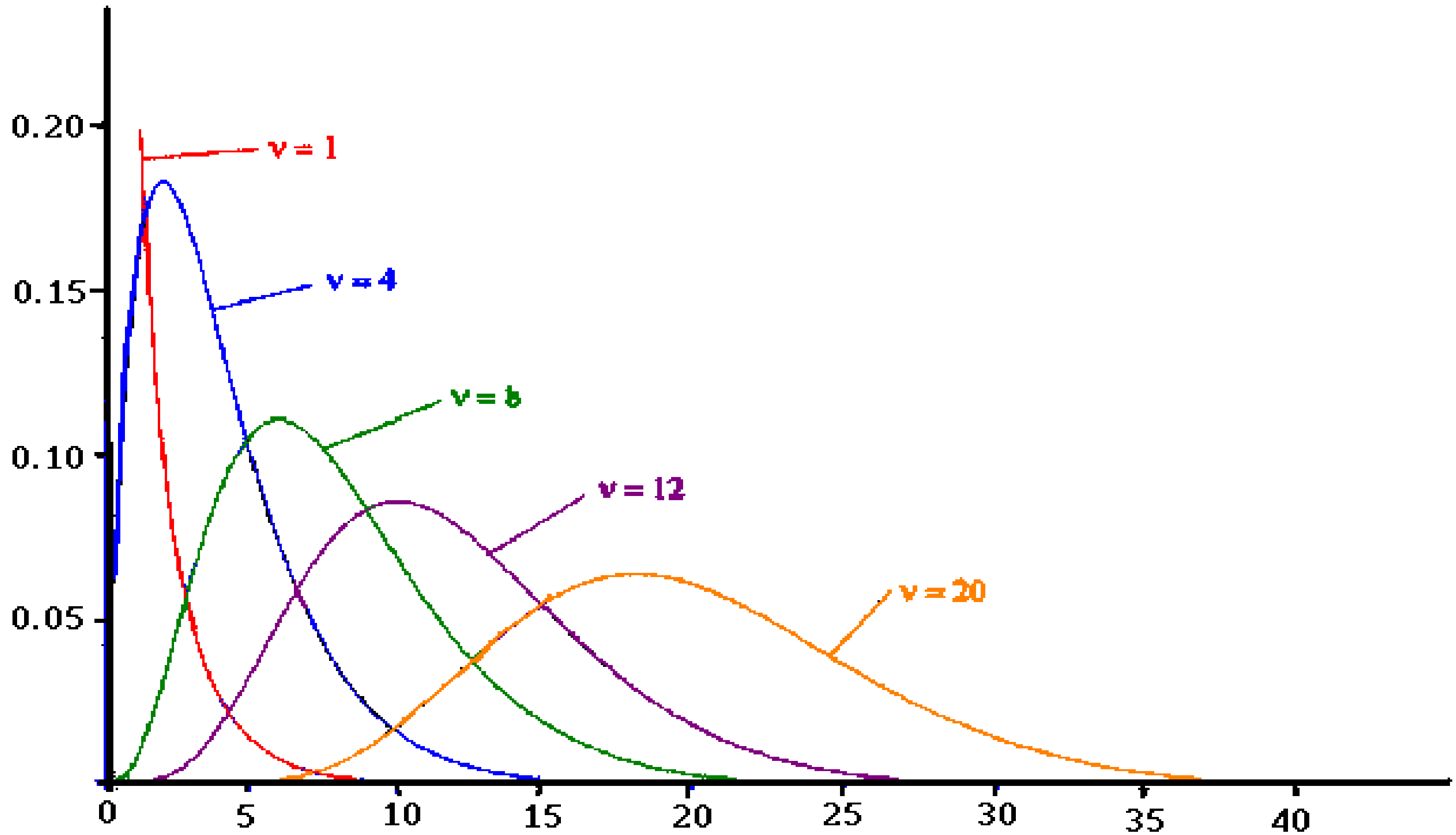
Perform the table of the test statistic:

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

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

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

Density function of Chi – squared distribution



Using Excel to Compute *Chi – squared* Distribution

- Excel has two functions for computing cumulative probabilities and *x* values for any *Chi - squared* distribution:
    - CHI.DIST is used to compute the cumulative probability given an *x* value, p-value.
    - CHI.INV is used to compute the *x* value given a cumulative probability, critical value.

Method A (p-value):

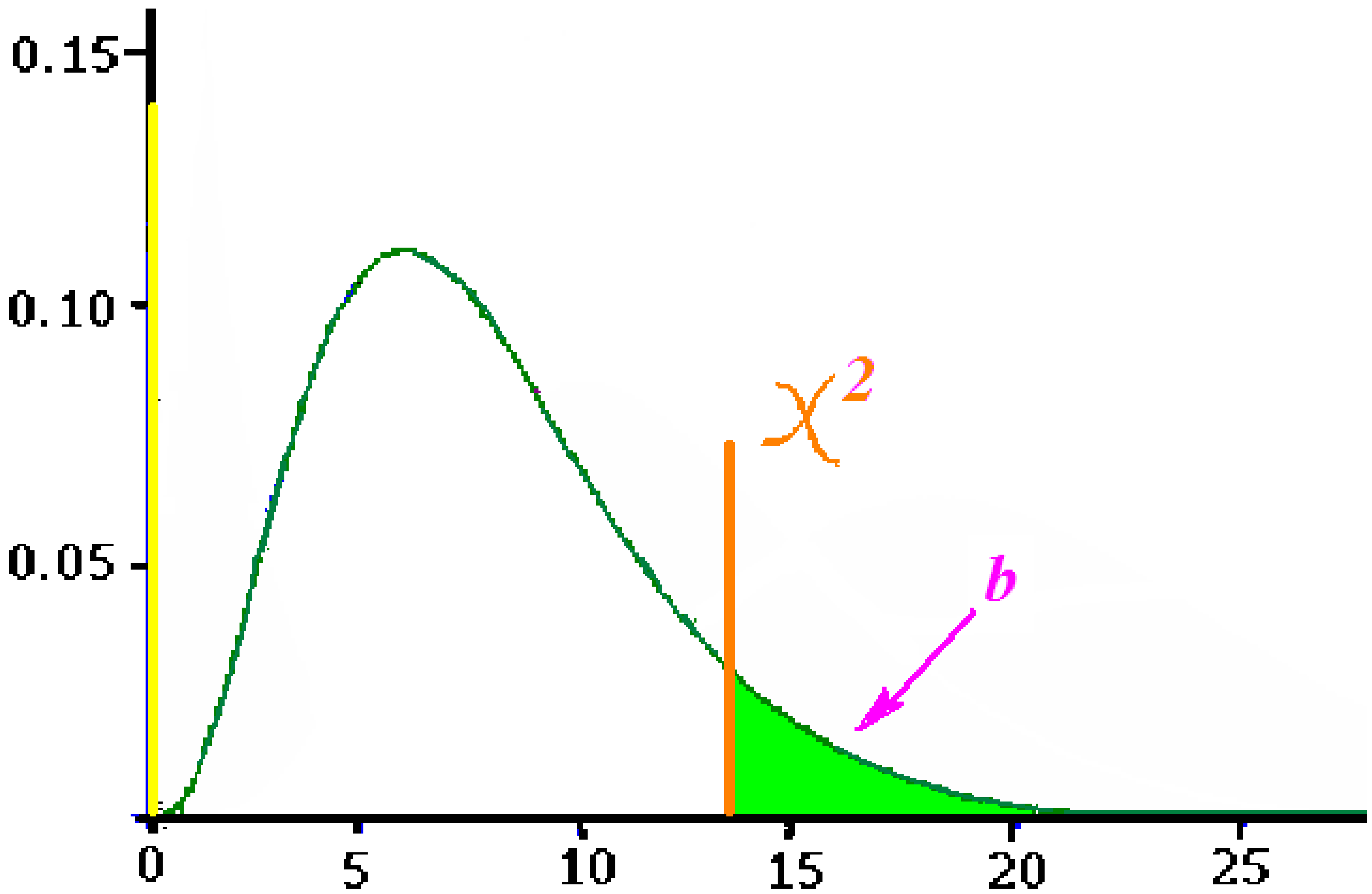
Step 1. Taking a variable $\chi^2_{(k-1)}$ of Chi-squared distribution with (k-1) degrees of freedom calculate the probability (p-value)

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

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

* If $b \geq \alpha$ → accept hypothesis H , conclude the all proportions are equal

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


0.15
0.10
0.05
0
5
10
15
20
25
χ2
b

Method B. (Critical value)

Looking in Table of Chi-squared distribution to find critical value $\chi^2_{(k-1)}(\alpha)$ of Chi-squared distribution with *k-1* degrees of freedom (α 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

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

One-way Analysis of Variance

The one-factor analysis of variance (ANOVA) test is an extension of the T-test method that compares the mean values to the multi-independent samples' model: For normally distributed random variables $X^{(1)}, X^{(2)}, \ldots, X^{(k)}$ with common variance and expectations $\mu_1, \mu_2, \ldots, \mu_k$, consider the hypothesis:

$$H : \mu_1 = \mu_2 = \ldots = \mu_k$$

Analysis of Variance:

$$H_0\text{: } \mu_1 = \mu_2 = \mu_3 = \cdot \ \cdot \ \cdot = \mu_k$$

H_a: Not all population means are equal

- If H_0 is rejected, we cannot conclude that *all* population means are different.

- Rejecting H_0 means that at least two population means have different values.

Analysis of Variance:

- Assumptions for Analysis of Variance
  - For each population, the response (dependent) variable is normally distributed.
  - The variance of the response variable, denoted σ^2, is the same for all of the populations.
  - The observations must be independent.

One-way Analysis of Variance

Consider the sample $\{x_j^{(i)}, j = 1, 2, \ldots, n_i\}$ of observations in the i-th group, corresponding to the variable $X^{(i)}$, including n_i observation, $i = 1, 2, \ldots, k$.

Consider the model $$x_j^{(i)} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where ε_{ij} is the deviation of each observation in the group from the group mean μ_i, and μ the common mean of all observations in the data, α_i is the difference between the group mean μ_i and the overall mean, μ obviously

$$\alpha_1 + \alpha_2 + \ldots + \alpha_k = 0$$

One-way Analysis of Variance

The total volatility in the i-th group, $i = 1, 2, \ldots, k$, is equal to

$$SS_i = \varepsilon_{i1}^2 + \varepsilon_{i2}^2 \ldots + \varepsilon_{in_i}^2$$

and

$$(\varepsilon_{i1}^2 + \varepsilon_{i2}^2 \ldots + \varepsilon_{in_i}^2)/(n_i - 1)$$

is the adjusted variance of the random variable $X^{(i)}$.

$$RSS = SS_1 + \ldots + SS_k = (\varepsilon_{11}^2 + \varepsilon_{12}^2 \ldots + \varepsilon_{1n_1}^2) + \ldots + (\varepsilon_{k1}^2 + \varepsilon_{k2}^2 \ldots + \varepsilon_{kn_k}^2)$$

is the sum of all volatility in all groups.

Denote the total number of observations by $n = n_1 + \ldots + n_k$

We can take $RSS/(n-k)$

as an estimate of the common variance of all variables

$$X^{(1)}, X^{(2)}, \ldots, X^{(k)}$$

One-way Analysis of Variance

$$BSS = n_1\alpha_1^2 + n_2\alpha_2^2 + \dots + n_k\alpha_k^2$$

is the total variation between the group means ,

$$BSS/(k-1)$$

reflecting the random variation from group to group of the considered quantity .

Ratio between two variances

$$F = \frac{BSS/(k-1)}{RSS/(n-k)}$$

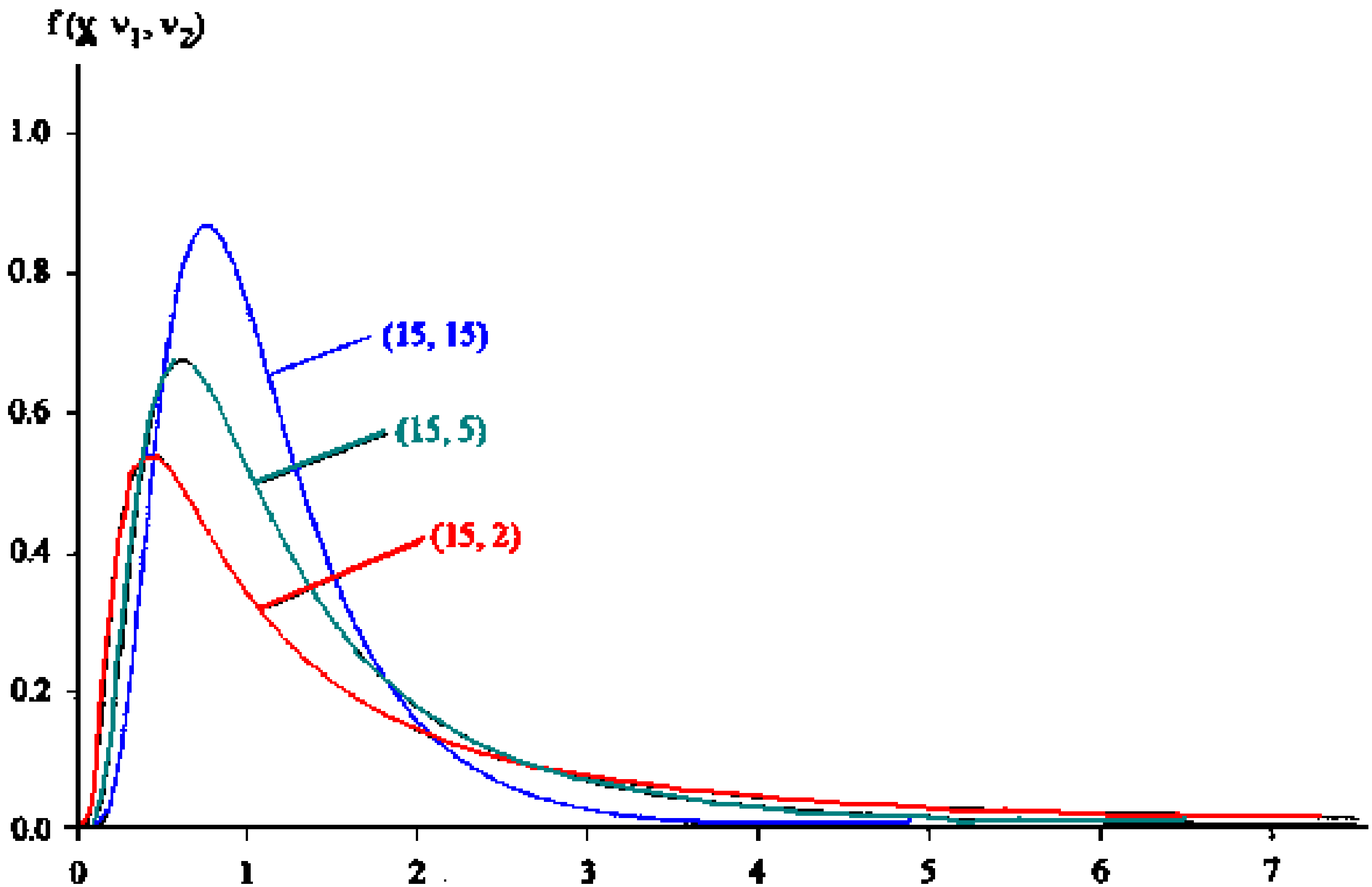
is the test statistic for hypothesis H comparing mean values.

One-way Analysis of Variance

Lemma. *If the hypothesis H is true, then the ratio (test statistic) F has a Fisher-Snedecor distribution with (k-1) and (n-k) degrees of freedom.*

With the above lemma, the testing problem has approaches of p-value or critical value.

Density function of Fisher-Snedecor distribution (F)

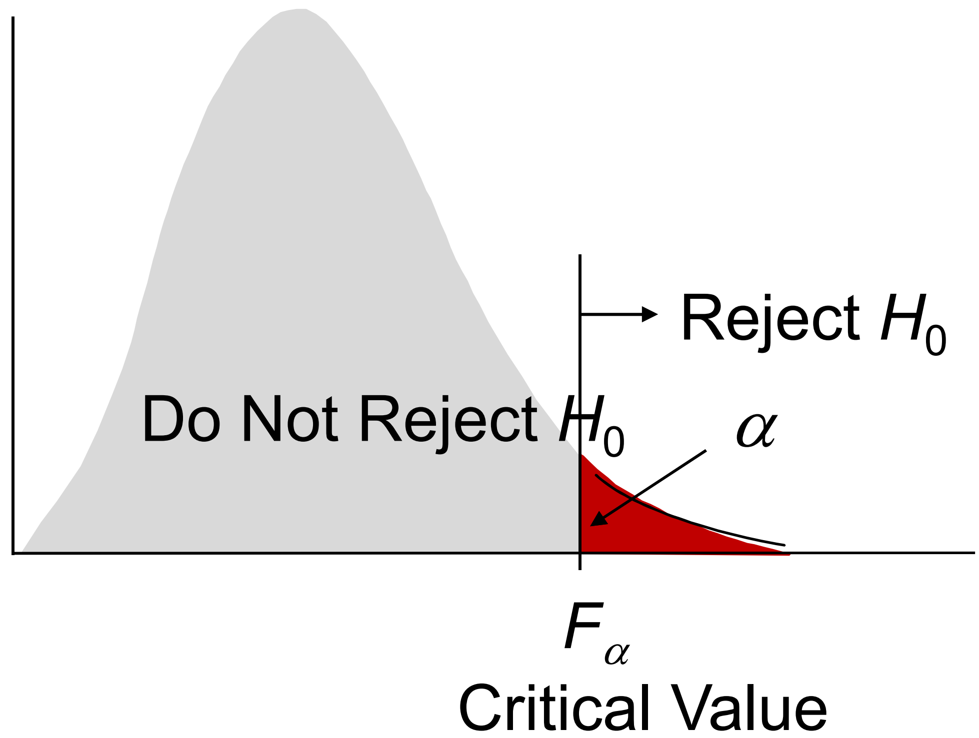


One-way Analysis of Variance

a) Compare the value of the test statistic F with the critical value $F_{cr} = F_{\alpha}^{(k-1,n-k)}$, which is the (1 - α) percentile of the Fisher-Snedecor distribution with $(k-1)$ and $(n-k)$ degrees of freedom: -

- If $F \geq F_{cr}$ $\rightarrow$ rejects hypothesis H,
- If $F < F_{cr}$ $\rightarrow$ accepts H .

The F Test



One-way Analysis of Variance

b) With ***FS*** being a random variable with a Fisher-Snedecor distribution with $(k-1)$ and $(n-k)$ degrees of freedom , calculate the probability of significance

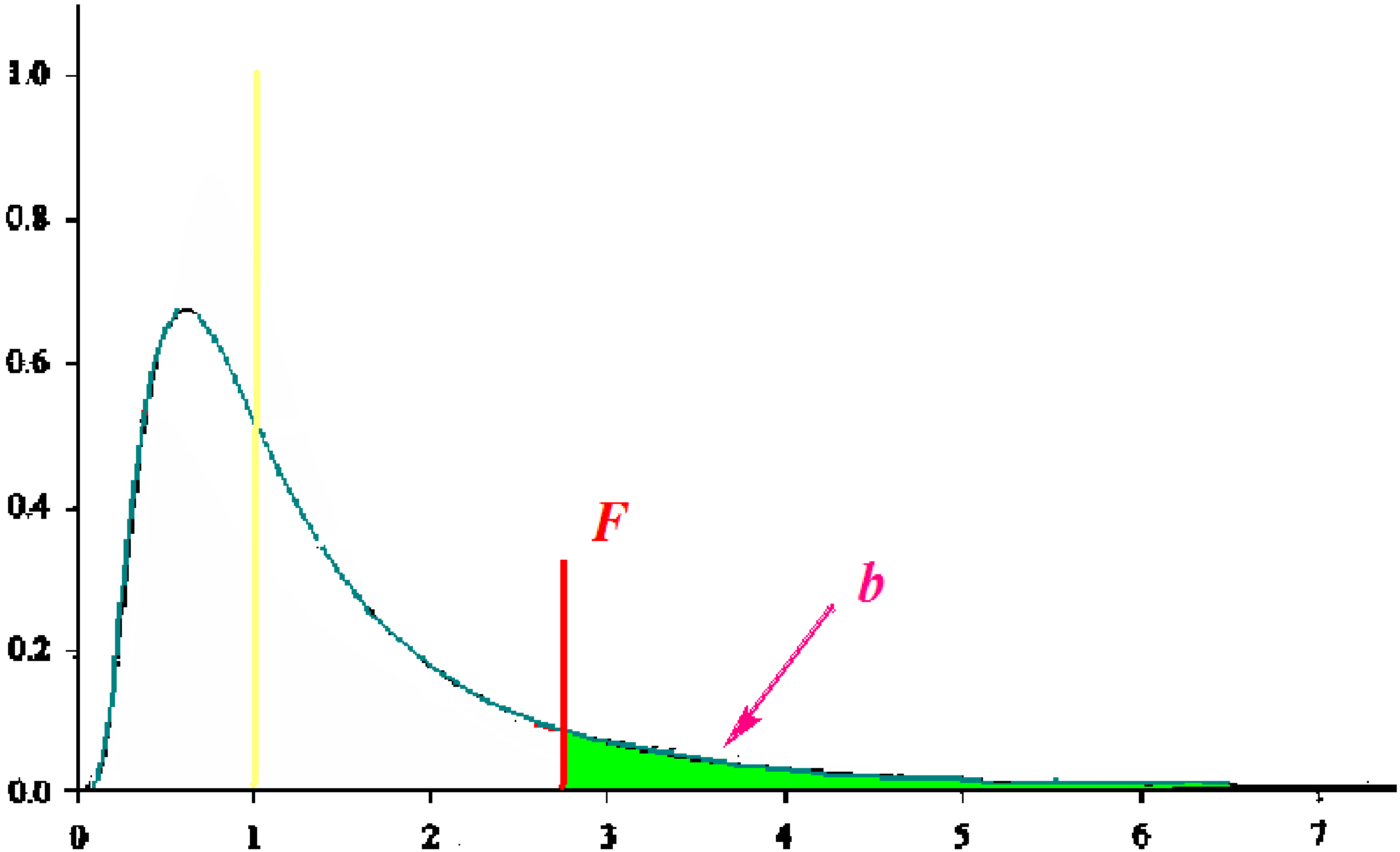
$$p_{sig} = \mathrm{P}\{FS > F\}$$

Compare the probability of significance of the test statistic with the significance level α (=5%):

-If $p_{sig} \leq \alpha$ ➔ rejects hypothesis H,-

- If $p_{sig} > \alpha$ ➔ accepts H.

The right-tailed probability of Fisher-Snedecor distribution (F)



Using Excel to Compute *Fisher - Snedecor* Distribution

- Excel has two functions for computing cumulative probabilities and *x* values for <u>any</u> *Fisher - Snedecor* distribution:
  - <u>F.DIST</u> is used to compute the cumulative probability given an *x* value, p-value.
  - <u>F.INV</u> is used to compute the *x* value given a cumulative probability, critical value.