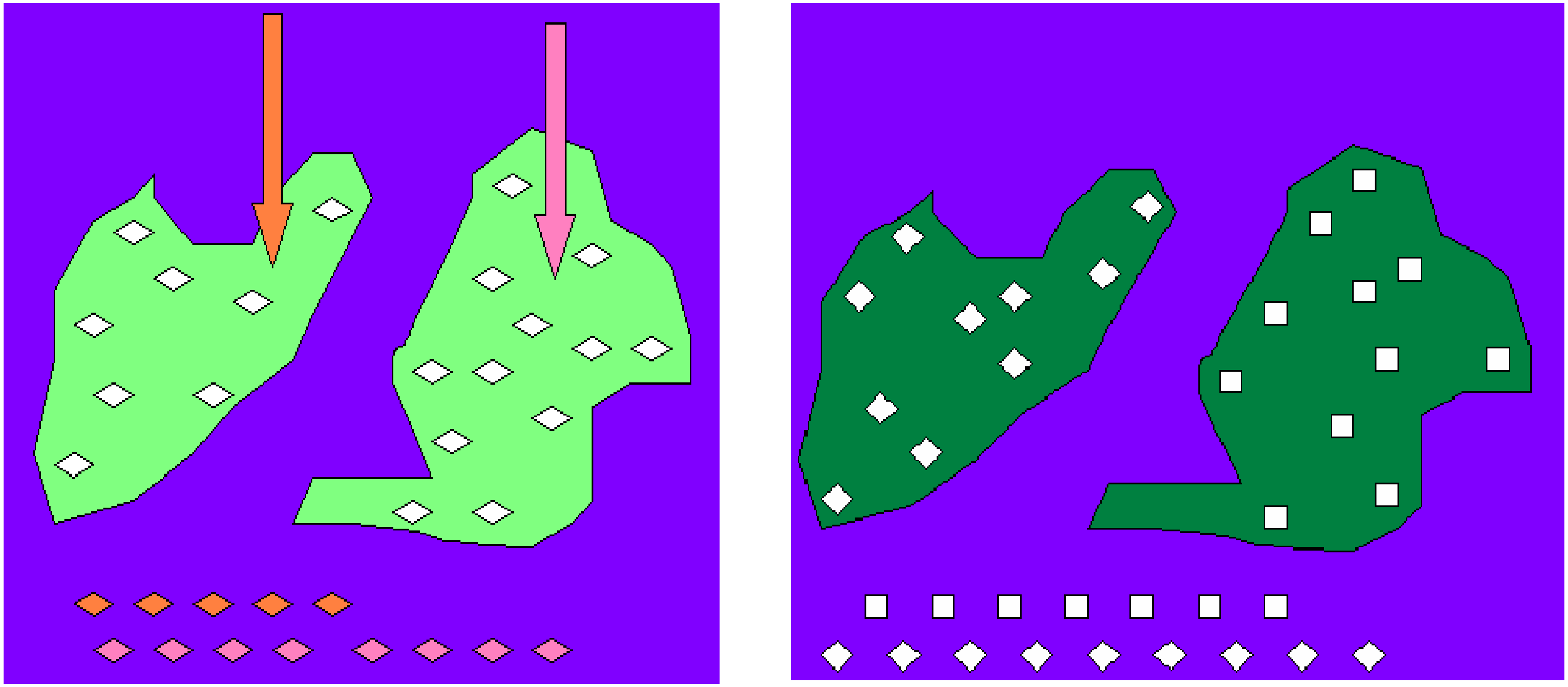
Hypothesis tests for two independent samples

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

B. Two independent samples model



Model of two groups of objects with different

a) Intervention levels,

b) Individual proper

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

Let $(X_1, X_2, ..., X_{n_1})$ be a sample of a binary variable X taking value 1 with probability p_1 and value 0 with probability $(1-p_1)$, $(Y_1, Y_2, ..., Y_{n_2})$ 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.

By Moivre-Laplace Theorem, for large sample size,

$n_1 \times p_1 \geq 5$ and $n_1 \times (1-p_1) \geq 5$,

$n_2 \times p_2 \geq 5$ and $n_2 \times (1-p_2) \geq 5$,

the sample proportions $m(p_1)/n_1$ and $m(p_2)/n_2$ of appearance of number 1 have distributions approximate to normal distribution with expectation p_1, p_2 and variance $p_1 \times (1-p_1)/n_1$, $p_2 \times (1-p_2)/n_2$, respectively. Denote $m_1 = m(p_1)$ and $m_2 = m(p_2)$.

If the Hypothesis H is true then use the two samples $(X_1, X_2, ..., X_{n_1})$ and $(Y_1, Y_2, ..., Y_{n_2})$ as samples collected from one variable and estimate the common variance of X and Y by

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

then perform a statistic

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

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 (p-value)

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

Step 3. Compare the probability **b** (p-value) to the given ahead significance α

* If **b** ≥ ***α*** → Accept Hypothesis **H** , confirm the equality of two proportions

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

(One-tail tests can be done similarly)

Version B. Using Normal critical value

Looking in Table of Normal distribution find out critical value $u_{\alpha/2}$ of Normal distribution (the critical value for α = 5% equals 1.96)

Decide

- Reject Hypothesis **H** if

$$|u| \geq u_{\alpha/2}$$

- Accept Hypothesis **H** if

$$|u| < u_{\alpha/2}$$

Version C. Using confidence intervals

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

$$\left[\frac{m_1}{n_1} - Z_{1-\alpha/2} * \sqrt{\frac{m_1}{n_1}(1-\frac{m_1}{n_1}) / n_1}\,;\frac{m_1}{n_1} + Z_{1-\alpha/2} * \sqrt{\frac{m_1}{n_1}(1-\frac{m_1}{n_1}) / n_1}\right]$$

$$\left[\frac{m_2}{n_2} - Z_{1-\alpha/2} * \sqrt{\frac{m_2}{n_2}(1-\frac{m_2}{n_2}) / n_2}\,;\frac{m_2}{n_2} + Z_{1-\alpha/2} * \sqrt{\frac{m_2}{n_2}(1-\frac{m_2}{n_2}) / n_2}\right]$$

Decide

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

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

Compare two mean values

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

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

Problem: Compare two expectations μ_1 and μ_2.

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

The problem can be solved by using the following Theorem:

Theorem. *Let* $(X_1, X_2, ..., X_n)$ *and* $(Y_1, Y_2, ..., Y_m)$ *be two samples of n independent observations selected correspondingly from a variable X with sample mean* $\bar{X}$ *and sample variance* S_X^2 *and from a variable Y with sample mean* $\bar{Y}$ *and sample variance* S_Y^2 *(both variables are* ***normal distributed with common variance****). If the hypothesis H is true* ($\mu_1 = \mu_2$) *then the variable (statistic)*

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

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

Hypothesis Tests

Hypothesis

$$H: \quad \mu_1 = \mu_2$$

Alternative Hypothesis

$$K: \quad \mu_1 \neq \mu_2$$

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 (p-value approach). Taking a variable $T^{(n+m-2)}$ of Student distribution with $(n + m - 2)$ degrees of freedom calculate the p-value (probability)

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

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

+ If ***b*** ≥ ***α*** → accept Hypothesis **H** and conclude

$$\mu_1 = \mu_2$$

+ If ***b*** < ***α*** → reject Hypothesis **H** and confirm

$$\mu_1 \neq \mu_2$$

Version B. Using Student critical value

Calculate the critical value $T^{(n+m-2)}_{(1-\alpha/2)}$ of Student distribution with $n+m-2$ degrees of freedom (α is a given ahead significance level =5%, 1% or 0.5%)

Decide

- Reject Hypothesis **H** if

$$|t| \geq T^{(n+m-2)}_{(1-\alpha/2)}$$

- Accept Hypothesis **H** if

$$|t| < T^{(n+m-2)}_{(1-\alpha/2)}$$

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:

$$(Mean(X) - T^{(n-1)}_{(1-\alpha/2)} \cdot \frac{SD(X)}{\sqrt{n-1}}; Mean(X) + T^{(n-1)}_{(1-\alpha/2)} \cdot \frac{SD(X)}{\sqrt{n-1}}),$$

$$(Mean(Y) - T^{(m-1)}_{(1-\alpha/2)} \cdot \frac{SD(Y)}{\sqrt{m-1}}; Mean(Y) + T^{(m-1)}_{(1-\alpha/2)} \cdot \frac{SD(Y)}{\sqrt{m-1}})$$

Decide

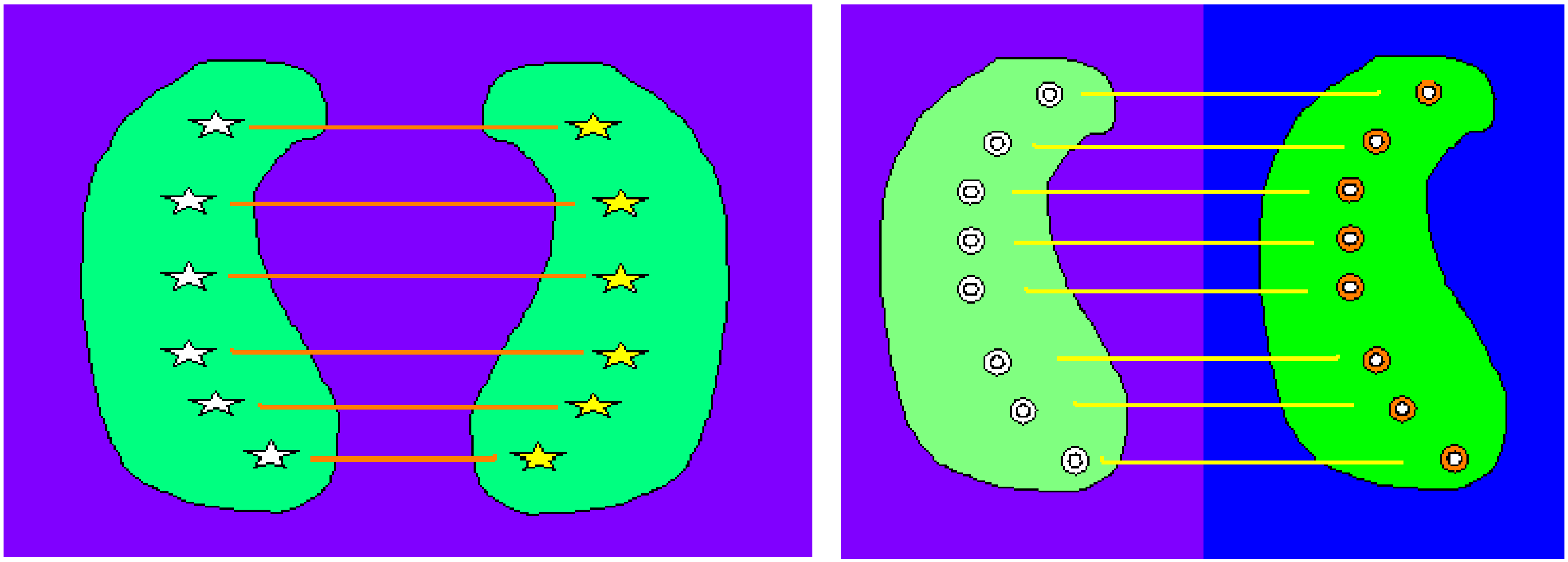
Reject Hypothesis **H** **if** the two intervals disjoin

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

Test for two related (paired) samples

- Compare two mean values

C. Model of two dependent (paired) samples



- Two dependent samples model is used in a study when
- A) Each object in the first sample is chosen together with a **similar** (paired) object in the second sample, or
- B) Any object in the second sample is the **same** one in the first sample, but the measures in the two samples are taken under **different conditions**.

Compare mean values of two related samples

For related variables ***X*** and ***Y***, the comparison of mean values is equivalent to the comparison the mean value of the difference variable ***X* – *Y*** to value ***0*** → the problem reduces to ***one-sample model***.

Compare mean values of two related samples

Hypothesis $H: \mu_1 = \mu_2$

Alternative Hypothesis $K: \mu_1 \neq \mu_2$

where μ_1 and μ_2 are the expectations of X and Y

With $\xi = X - Y$,

➔comparing expectations μ_ξ of ξ to 0:

Hypothesis $H: \mu_\xi = 0$

Alternative Hypothesis $K: \mu_\xi \neq 0$

Compare mean values of two related samples

With the empirical value of the test statistic

$$t = \frac{\sqrt{n-1}}{s_{\xi}}\bar{\xi}_n$$

a) Compare the empirical value of the t-test statistic with the critical value $t_{cr} = T_{1-\alpha/2}^{(n-1)}$, which is the $1-\alpha/2$ percentile of the Student distribution with *n*-1 degrees of freedom:

- \- If $|t| \geq t_{cr}$ ➔ reject the hypothesis *H* ,
- \- If $|t| < t_{cr}$ ➔ accept *H* .

Compare mean values of two related samples

b) Taking a random variable ***T*** having Student distribution with *n*-1 degrees of freedom, calculate the probability of significance

$$p_{sig} = \mathrm{P}\{|T| > |t|\}$$

Compare the probability of significance p_{sig} with the significance level α :

- \- If $p_{sig} \leq \alpha$ ➔ reject *H* ,
- \- If $p_{sig} > \alpha$ ➔ accept *H* .

Compare mean values of two related samples

c) Determine the 95% confidence interval of the estimation $\bar{\xi}_n$:

$$a\%CI(\bar{\xi}_n) = (\bar{\xi}_n - T_{1-\alpha/2}^{(n-1)} \cdot \frac{s_\xi}{\sqrt{n-1}}; \bar{\xi}_n + T_{1-\alpha/2}^{(n-1)} \cdot \frac{s_\xi}{\sqrt{n-1}})$$

Compare 0 to the confidence interval:

- \- If $0 \notin a\%CI(\bar{\xi}_n)$ ➔ reject *H*,
- \- If $0 \in a\%CI(\bar{\xi}_n)$ ➔ accept *H* .