CHI-SQUARE DISTRIBUTION
&

THE ANALYSIS OF FREQUENCIES

Nguyen Quang Vinh - Nguyen Thi Tu Van

CHI-SQUARE (χ^2)

Karl Pearson (1857-1936)

- One of the most widely used distributions
- Testing hypothesis where data in form of frequencies: to test differences between proportions
- Most appropriate for use with categorical variables

$N(\mu, \sigma^2)$ *transformed to* $N(0, 1)$ *by:*

$$z = \frac{x - \mu}{\sigma}$$

$$z^2 = \left(\frac{x - \mu}{\sigma}\right)^2 \text{ follows a } \chi^2 \text{distribution}$$

with $df = 1$, *or* :

$$z^2 = \chi^2_{(1)}$$

$$\chi^2_{(2)} = \left(\frac{x_1 - \mu}{\sigma}\right)^2 + \left(\frac{x_2 - \mu}{\sigma}\right)^2 = z_1^2 + z_2^2$$

follows a χ^2 distribution *with df* $= 2$

In general:

$$\chi^2_{(n)} = z_1^2 + z_2^2 + \ldots + z_n^2$$

follows a χ^2 distribution *with df* $= n$

The mathematical form of the χ^2 *distribution:*

$$f(u) = \frac{1}{\left(\frac{k}{2} - 1\right)!} \frac{1}{2^{k/2}} u^{(k/2)-1} e^{-(u/2)}, \qquad u > 0$$

$$\text{where } e = 2.71828, \; k = df$$

APPLICATIONS OF THE χ^2 *STATISTIC*

- Observed frequencies (OBSERVATION)

  vs.

  Expected frequencies (HYPOTHESIS)

- (1) Test of *goodness-of-fit*

  (2) Test of *independence*

  (3) Test of *homogeneity*

χ^2 test of goodness-of- fit

- A two-tailed test on p
- For Binomial situation:

$$H_O: p = p_0$$
$$H_A: p \neq p_0$$

- For Multinomial situation:

$$H_O: p_1 = p_{10},\ p_2 = p_{20},\ \ldots,\ p_k = p_{k0}$$

H_A: at least one of the p_i*'s is incorrect*

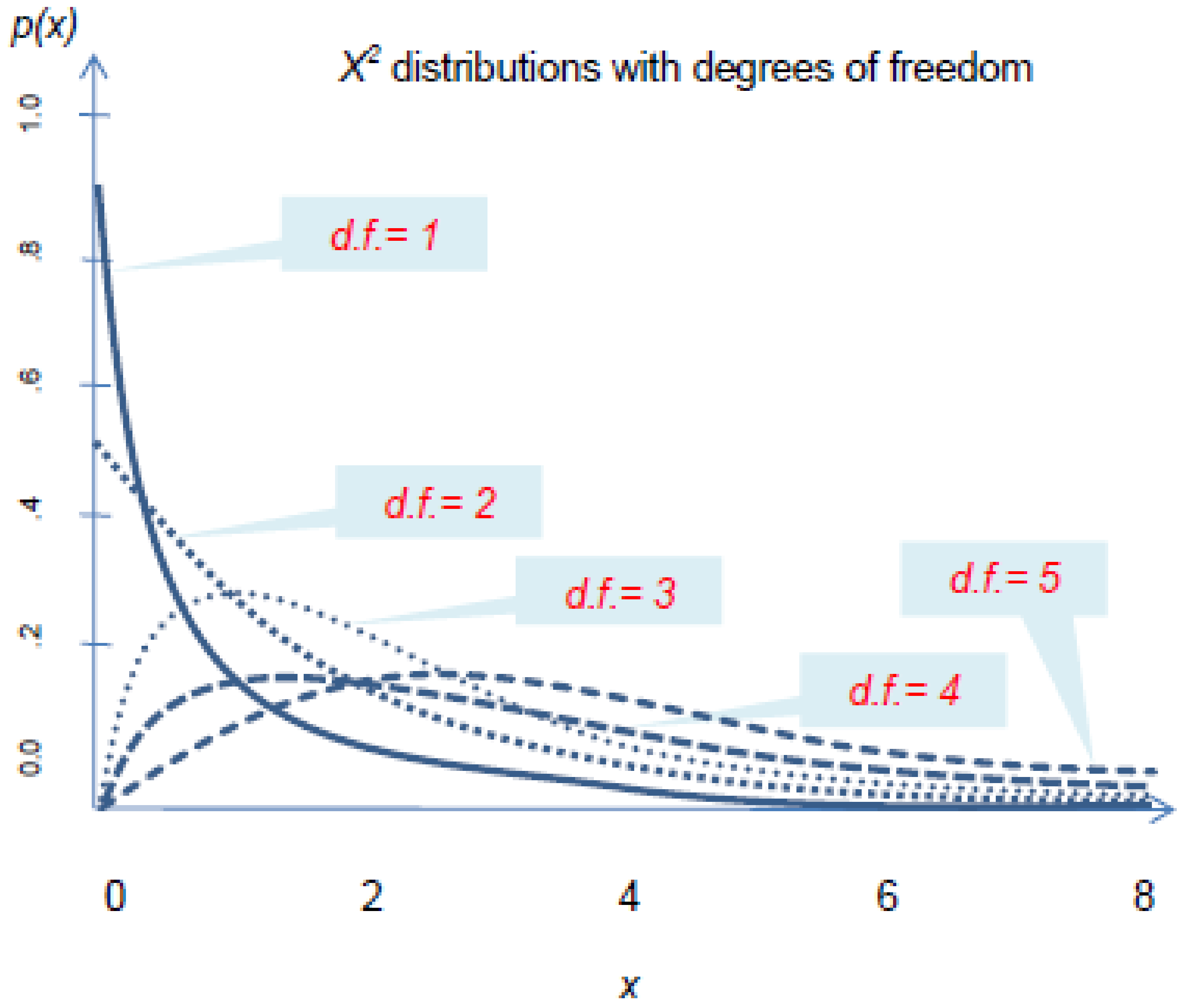
$$\chi^2 \ test\ of\ goodness-of-fit$$

Test statistic :

$$\chi_c^2 = \sum \frac{(O-E)^2}{E}$$

$df =$ number of categories - 1

reject H_o if $\chi_c^2 > \chi_{\alpha,df}^2$


p(x)
X² distributions with degrees of freedom
1.0
.8
.6
.4
.2
0.0
d.f.= 1
d.f.= 2
d.f.= 3
d.f.= 4
d.f.= 5
0
2
4
6
8
x

χ^2 test of goodness-of- fit

- *How well the distribution of sample data conforms to some theorical distribution**

- *Small expected frequencies: 5*

  - Combining adjacent categories $\rightarrow$ to achieve the suggested minimum.

**Kolmogorov-Smirnov test $\rightarrow$ continuous distributions*

χ^2 test of Independence

- Most frequent use of χ^2
- A *single* population, where each member was classified according to 2 criteria:

  1st criteria: row

  2nd criteria: column
- Contingency table: r rows, c columns
- H_O: 2 criteria of classification *are independent*
  H_A: 2 criteria of classification *are not independent*
- $df = (r - 1)(c - 1)$

χ^2 *test of Independence*

Small expected frequencies

- χ^2 test should not be used if *any* E_i < 5

χ^2 test of Homogeneity

To determine whether the *distinct* groups can be viewed as belonging to the *same* population.

χ^2 test of Independence

- row and column totals are not under the control of the investigator
- ? *independent* (the 2 criteria)

χ^2 test of Homogeneity

- either row or column totals may be under the control of the investigator
- ? *homogeneous* (the samples drawn from the *same* population)

mathematically ***equivalent*** *but conceptually* ***different***

FISHER'S EXACT TEST

	Treatment	Control	**Total**
O+	x	$K - x$	**K**
O-	$n - x$	$(N\text{-}K)\text{-}(n\text{-}x)$	**N - K**
Total	**n**	**N-n**	**N**

$$N \rightarrow \left\{ \begin{matrix} K & & x \\ & & \\ N-K & & n\text{-}x \end{matrix} \right\} \leftarrow n$$

$$P(x) = \frac{{}_K C_x \cdot {}_{N-K} C_{n-x}}{{}_N C_n}$$

We have a result from a trial as follow:

	Treatment	Control	**Total**
O+	6	1	**7**
O-	2	4	**6**
Total	**8**	**5**	**13**

Listing all possible tables in the sample of size 13, which have:

- 7 positive outcomes &
- 8 subjects in treatment group.

We have 6 tables as follow:

	Treatment	Control	**Total**
O+	*7*	0	**7**
O-	1	5	**6**
Total	**8**	**5**	**13**

$$P(x=7) = \frac{{}_7C_7 \cdot {}_6C_1}{{}_{13}C_8}$$

$$= \frac{6}{1287} = .0047$$

	Treatment	Control	**Total**
O+	*6*	1	**7**
O-	2	4	**6**
Total	**8**	**5**	**13**

$$P(x=6) = \frac{{}_7C_6 \cdot {}_6C_2}{{}_{13}C_8}$$

$$= .0816$$

	Treatment	Control	**Total**
O+	***5***	2	**7**
O-	3	3	**6**
Total	**8**	**5**	**13**

$$P(x=5)=\frac{{}_{7}C_{5}\cdot{}_{6}C_{3}}{{}_{13}C_{8}}$$

$$=.3262$$

	Treatment	Control	**Total**
O+	***4***	3	**7**
O-	4	2	**6**
Total	**8**	**5**	**13**

$$P(x=4)=\frac{{}_{7}C_{4}\cdot{}_{6}C_{4}}{{}_{13}C_{8}}$$

$$=.4070$$

	Treatment	Control	**Total**
O+	***3***	4	**7**
O-	5	1	**6**
Total	**8**	**5**	**13**

$$P(x=3) = \frac{{}_7C_3 \cdot {}_6C_5}{{}_{13}C_8}$$

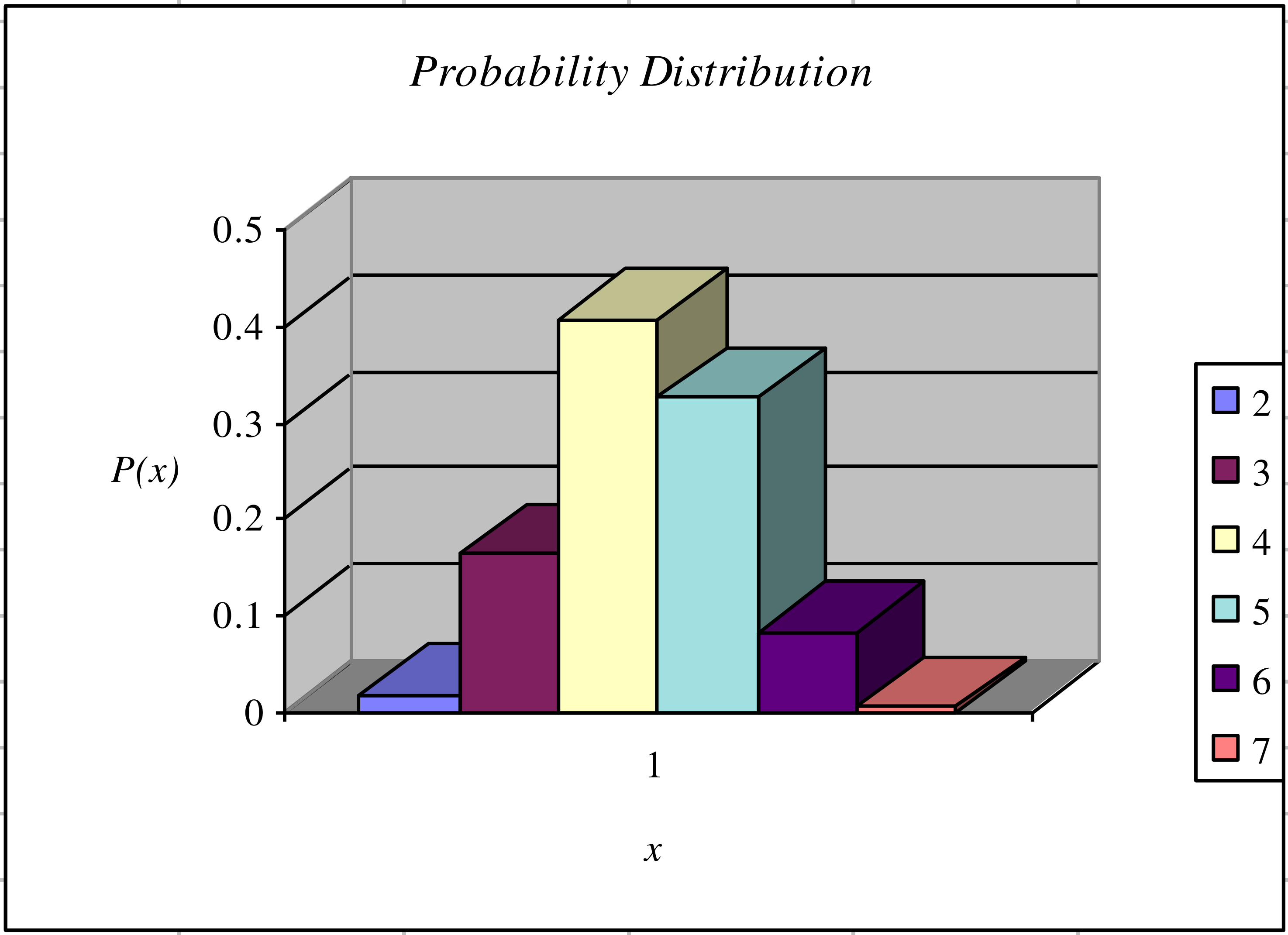
$$= .1632$$

	Treatment	Control	**Total**
O+	***2***	5	**7**
O-	6	0	**6**
Total	**8**	**5**	**13**

$$P(x=2) = \frac{{}_7C_2 \cdot {}_6C_6}{{}_{13}C_8}$$

$$= .0163$$

A useful check is that all the probabilities should sum to one (within the limits of rounding)



Probability Distribution
0.5
0.4
0.3
0.2
0.1
0
P(x)
1
x
2
3
4
5
6
7

Hypothesis

- H_O: $\pi_T = \pi_C$
  (no difference between treatment & control group)

- H_A: $\pi_T > \pi_C$ *(1-tailed)*

  or,

  H_A: $\pi_T \neq \pi_C$ *(2-tailed)*

Calculate P value

- The observed set has a probability of 0.0816
- The P value is the probability of getting the observed set, or one more extreme.

One tailed P value:

$P(x \geq 6) = P(x=6) + P(x=7) = 0.0816 + 0.0047 = 0.0863$

Calculate P value

Two tailed P value:

(1) $P(x \geq 6 \text{ or } x \leq 2) = P(x=2)+P(x=6)+P(x=7) =$
$0.0816 + 0.0047 + 0.0163 = 0.1026$

(2) Double the one tailed result*, thus:

$P = 2 \times 0.0863 = 0.1726$

**approximation*