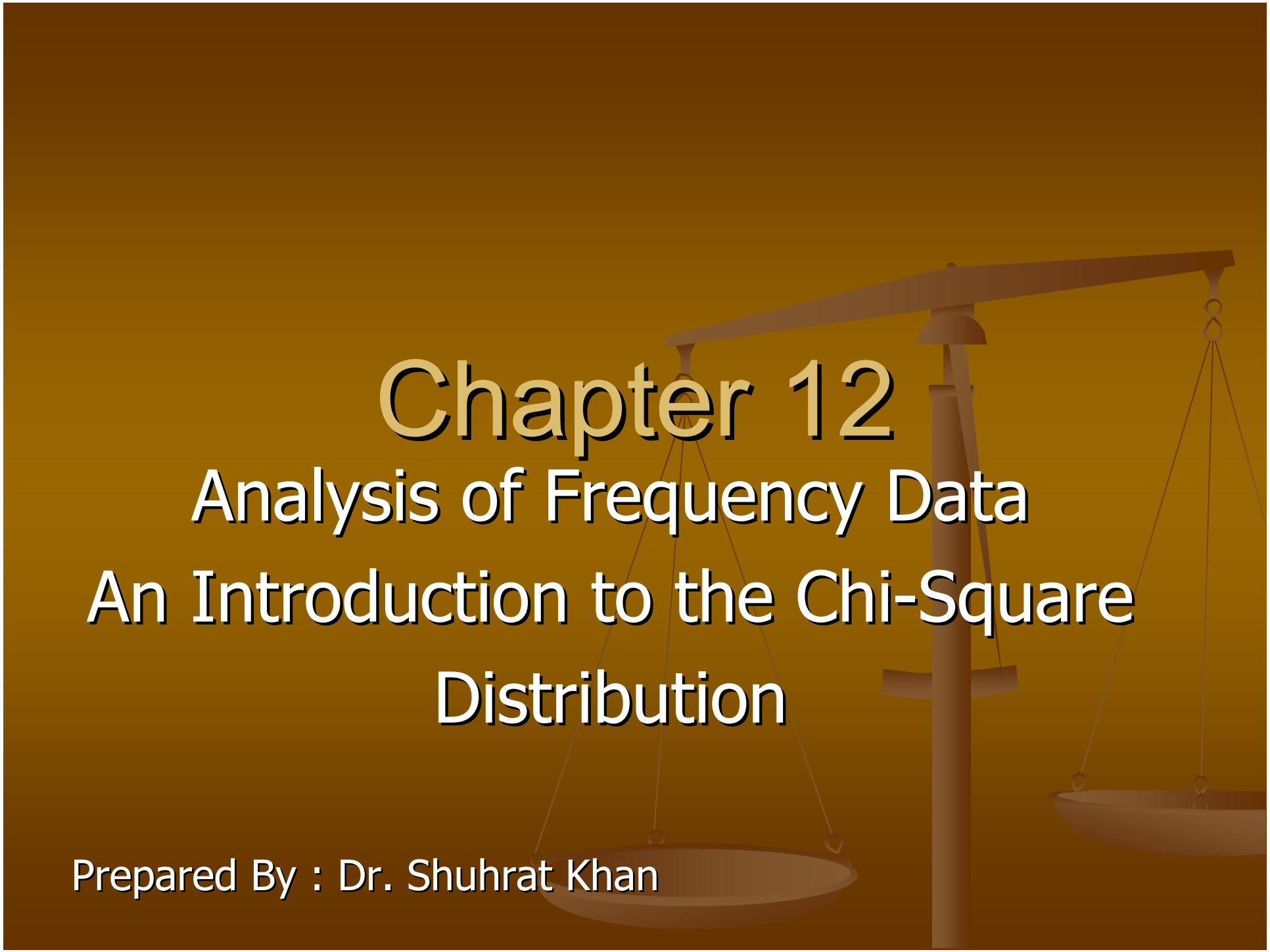

Chapter 12

Analysis of Frequency Data

An Introduction to the Chi-Square Distribution

Prepared By : Dr. Shuhrat Khan

TESTS OF INDEPENDENCE

- To test whether two criteria of classification are independent . For example socioeconomic status and area of residence of people in a city are independent.
- We divide our sample according to status, low, medium and high incomes etc. and the same samples is categorized according to urban, rural or suburban and slums etc.
- Put the first criterion in columns equal in number to classification of 1st criteria (Socioeconomic status) and the 2nd in rows, where the no. of rows equal to the no. of categories of 2nd criteria (areas of cities).

The Contingency Table

- Table Two-Way Classification of sample

First Criterion of Classification →

Second Criterion ↓	1	2	3	…..	c	Total
1	N_{11}	N_{12}	N_{13}	……	N_{1c}	$N_{1.}$
2	N_{21}	N_{22}	N_{23}	……	N_{2c}	$N_{2.}$
3	N_{31}	N_{32}	N_{33}	……	N_{3c}	$N_{3.}$
.	.	.	.	……	.	.
.	.	.	.		.	.
r	N_{r1}	N_{r2}	N_{r3}		N_{rc}	$N_{r.}$
Total	$N_{.1}$	$N_{.2}$	$N_{.3}$	……	$N_{.c}$	N

Observed versus Expected Frequencies

- O_{ij} : The frequencies in ith row and jth column given in any contingency table are called observed frequencies that result form the cross classification according to the two classifications.
- e_{ij} :Expected frequencies on the assumption of independence of two criterion are calculated by multiplying the marginal totals of any cell and then dividing by total frequency
- Formula:

$$e_{ij} = \frac{(N_{i\bullet}(N_{\bullet j})}{N}$$

Chi-square Test

- After the calculations of expected frequency,

  Prepare a table for expected frequencies and use Chi-square

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(o_i - e_i)^2}{e_i} \right]$$

  Where summation is for all values of r xc = k cells.

- D.F.: the degrees of freedom for using the table are (r-1)(c-1) for α level of significance
- Note that the test is always one-sided.

Example 12.401(page 613)

The researcher are interested to determine that preconception use of folic acid and race are independent. The data is:

Observed Frequencies Table

frequencies Table

	Use of Folic Acid		total
	Yes	No	
White	260	299	559
Black	15	41	56
Other	7	14	21
Total	282	354	636

Expected

	Yes	no	Total
White	(282)(559)/636 =247.86	(354)(559)/636 =311.14	559
Black	(282)(56)/636 =24.83	(354)(559) = 31.17	56
Others	(282)((21) =9.31	21x354/636 =11.69	21

Calculations and Testing

- Data: See the given table
- Assumption: Simple random sample
- Hypothesis: H_0: race and use of folic acid are independent
  HA: the two variables are not independent. Let α = 0.05
- The test statistic is Chi Square given earlier
- Distribution when H_0 is true chi-square is valid with (r-1)(c-1) = (3-1)(2-1)= 2 d.f.
- Decision Rule: Reject H0 if value of χ^2 is greater than $\chi^2_{\alpha,(r-1)(c-1)}$ = 5.991
- Calculations:

$$\chi^2 = (260 - 247.86)^2 / 247.86 + (299 - 311.14)^2 / 311.14 + \ldots + (14 - 11.69)^2 / 11.69 = \underline{9.091}$$

Conclusion

- Statistical decision. We reject H_0 since 9.08960> 5.991
- Conclusion: we conclude that H_0 is false, and that there is a relationship between race and preconception use of folic acid.
- P value. Since 7.378< 9.08960< 9.210, 0.01<p <0.025
- We also reject the hypothesis at 0.025 level of significance but do not reject it at 0.01 level.
- Solve Ex12.4.1 and 12.4.5 (p 620 & P 622)

ODDS RATIO

- In a retrospective study, samples are selected from those who have the disease called '*cases*' and those who do not have the disease called '*controls*' . The investigator looks back (have a *retrospective look)* at the subjects and determines which one have (or had) and which one do not have (or did not have) the risk factor.
- The data is classified into 2x2 table, for comparing cases and controls for risk factor *ODDS RATIO* IS CALCULATED
- ODDS are defined to be the ratio of probability of success to the probability of failure.
- The estimate of population odds ratio is $\therefore \hat{OR} = \frac{a/b}{cld} = \frac{ad}{bc}$

ODDS RATIO

- Where a, b, c and d are the numbers given in the following table:

Risk Factor ↓	Sample		Total
	Cases	Control	
Present	a	b	a + b
Absent	c	d	c + d
Total	a + c	b + d	

- We may construct 100(1-α)%CI for OR by formula:
-

$$R^{1 \pm (z_{\alpha/2} / \sqrt{X^2})}$$

Example 12.7.2 for Odds Ratio

- Example 12.5.7.2 page 640: Data relates to the obesity status of children aged 5-6 and the smoking status of their mothers during pregnancy
- Hence OR for table
- is :

$$OR = \frac{(64)(3496)}{(342)(68)} = 9.62$$

Obesity status

Smoking status(during Pregnancy)	cases	Non-cases	Total
Smoked throughout	64	342	406
Never smoked	68	3496	3564
Total	132	3838	3970

Confidence Interval for Odds Ratio

The (1-α) 100% Confidence Interval for Odds Ratio is:

$$\hat{OR}^{1\pm(z_\alpha/\sqrt{X^2})}$$

Where

$$X^2 = \frac{n(ad-bc)^2}{(a+c)(a+d)(b+c)(b+d)}$$

For Example 12.5.7.2 we have: a=64, b=342, c=68, d=3496 , therefore:

$$X^2 = 3970(64\times3496-342\times68)^2 \Big/ (132)(3833)(406)(3564) = 217.68$$

Its 95% CI is:

$$\hat{OR}^{1\pm(z_\alpha/\sqrt{X^2})} = 9.62^{1\pm(1.96/\sqrt{217.6831})}$$

- or (7.12, 13.00)

Interpretation of Example 12.7.2 Data

- The 95% confidence interval (7.12, 13.00) mean that we are 95% confident that the population odds ratio is somewhere between 7.12 and 13.00
- Since the interval does not contain 1, in fact contains values larger than one, we conclude that, in Pop. Obese children (cases) are more likely than non-obese children (non-cases) to have had a mother who smoked throughout the pregnancy.
- Solve Ex 12.7.4 (page 646)

Interpretation of ODDS RATIO

- The sample odds ratio provides an estimate of the relative risk of population in the case of a rare disease.
- The odds ratio can assume values between 0 to ∞.
- A value of 1 indicate no association between risk factor and disease status.
- A value greater than one indicates increased odds of having the disease among subjects in whom the risk factor is present.