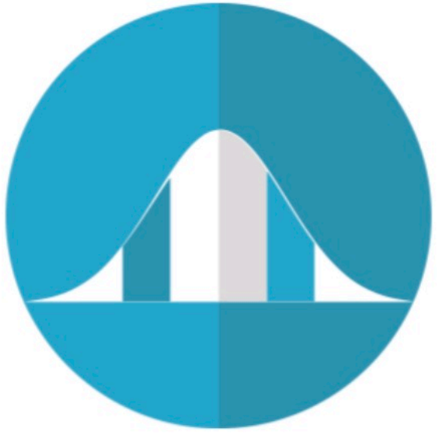
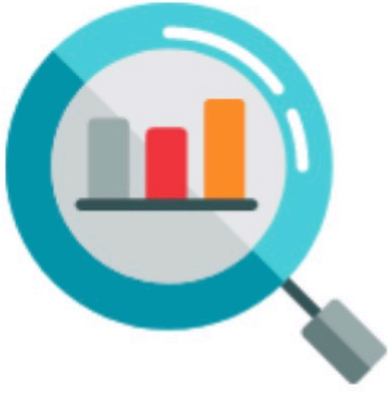


Statistics
Session-7



CLARUSWAY©
WAY TO REINVENT YOURSELF

Did you finish Statistics (Analysis of Categorical Data) pre-class activity?

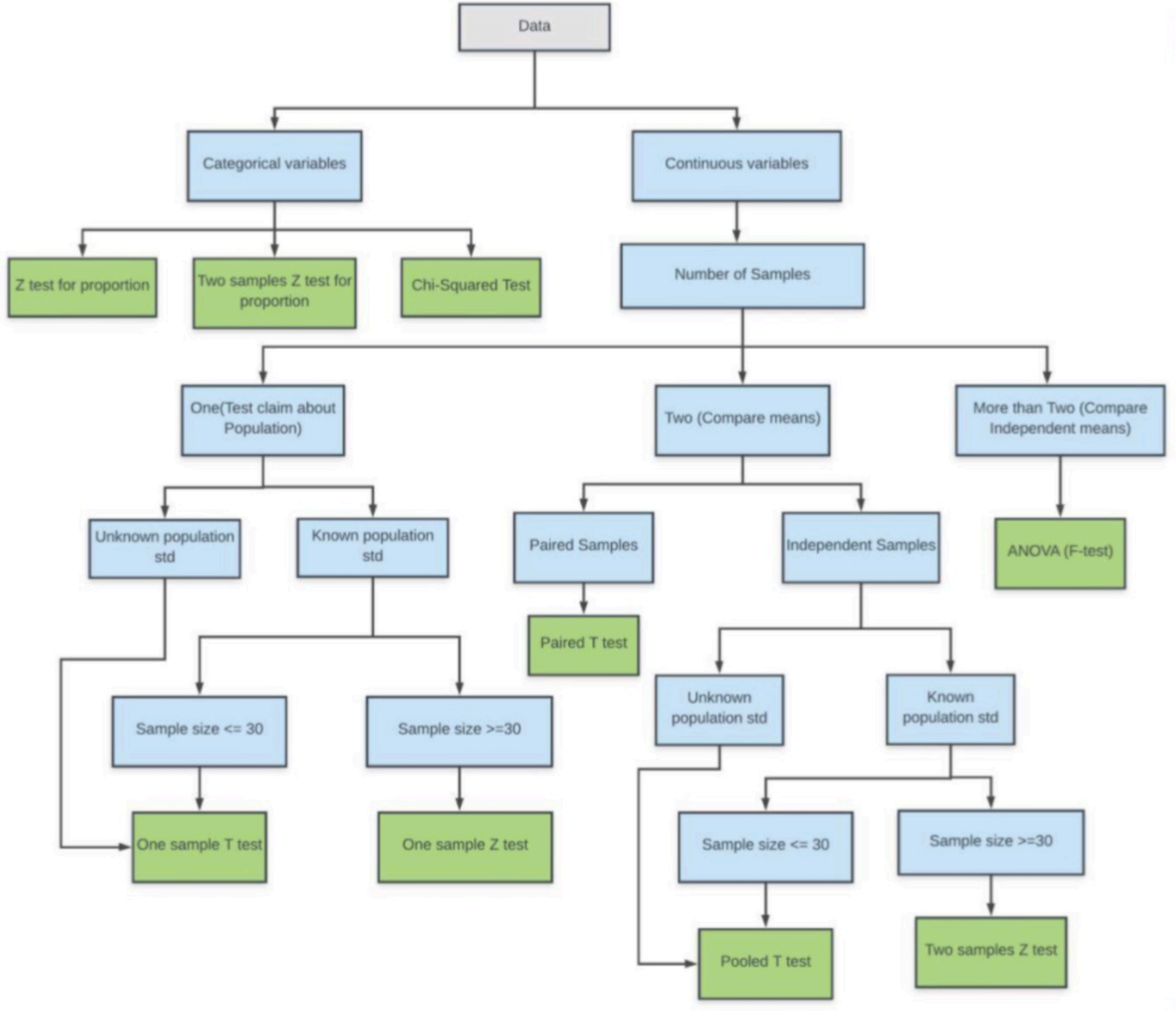
Students choose an option

Pear Deck Interactive Slide
Do not remove this bar

No Multiple Choice Response
You didn't answer this question

Table of Contents

Test Types


Data
Categorical variables
Continuous variables
Z test for proportion
Two samples Z test for proportion
Chi-Squared Test
Number of Samples
One(Test claim about Population)
Two (Compare means)
More than Two (Compare Independent means)
Unknown population std
Known population std
Paired Samples
Independent Samples
ANOVA (F-test)
Paired T test
Sample size <= 30
Sample size >=30
Unknown population std
Known population std
One sample T test
One sample Z test
Sample size <= 30
Sample size >=30
Pooled T test
Two samples Z test

The Steps of a Significance Test

A significance test has five steps.

- **Step 1** : Assumptions
- **Step 2** : Hypotheses
- **Step 3** : Test Statistic
- **Step 4** : P-Value
- **Step 5** : Conclusion

Association

- Examining relationship between 2 categorical variables
- Some examples of association:
  - Smoking and lung cancer
  - Ethic group and coronary heart disease

- Questions of interest when testing for association between two categorical variables
  - Does the presence/absence of one factor (variable) influence the presence/absence of the other factor (variable)?
- Caution
  - Presence of an association does not necessarily imply causation

Contingency Tables

A contingency table (also called crosstab) is used to reveal the association between categorical variables.

Suppose that we wish to classify defects found on furniture produced in a manufacturing plant according to

(1) the type of defect
(2) the production shift

Contingency Tables

	Type of Defect				
Shift	A	B	C	D	*Total*
1	15	21	45	13	*94*
2	26	31	34	5	*96*
3	33	17	49	20	*119*
Total	*74*	*69*	*128*	*38*	*309*

- n=309 furniture defect was recorded
- The defects were classified as one of four types: A, B, C, or D
- The shifts were classified as one of three types: 1, 2, or 3

Association

	Type of Defect				
Shift	A	B	C	D	*Total*
1	15	21	45	13	*94*
2	26	31	34	5	*96*
3	33	17	49	20	*119*
Total	*74*	*69*	*128*	*38*	*309*

- Is there an association between type of defect and production shift?
- How can we see whether there is a relationship between those two variables?

Chi square

To analyze relationship between two categorical variables: Pearson's χ2 test.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Karl Pearson

Founder of Modern Statistics

Hypotheses

Null Hypothesis: The two categorical variables are independent.

- There **is no** association between type of defect and production shift in the population.

Alternative Hypothesis: The two categorical variables are dependent.

- There **is an** association between type of defect and production shift in the population.

Chi square test statistic

To test this, we **compare** the observed frequencies in each category to the frequencies we would expect in the categories if there would be no association

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

fo = observed frequencies
fe = expected frequencies (if there would be no association = expected under the null hypothesis)

Expected Frequencies

	Type of Defect				
Shift	A	B	C	D	*Total*
1					*94*
2					*96*
3					*119*
Total	*74*	*69*	*128*	*38*	*309*

- We can calculate the expected frequency of the marked cell via this formula:

$$f_e = \frac{n_r n_c}{n}$$

- n_r = total number in the row
- n_c = total number in the column
- n = total sample size

Expected Frequencies

	Type of Defect				
Shift	A	B	C	D	*Total*
1	22.51	20.99	38.94	11.56	*94*
2	22.99	21.44	39.77	11.81	*96*
3	28.50	26.57	49.29	14.63	*119*
Total	*74*	*69*	*128*	*38*	*309*

- We can calculate the expected frequency of the marked cell via this formula:

$$f_e = \frac{n_r n_c}{n}$$

- n_r = total number in the row
- n_c = total number in the column
- n = total sample size

Observed vs Expected Frequencies

	Type of Defect				
Shift	A	B	C	D	*Total*
1	15 (22.51)	21 (20.99)	45 (38.94)	13 (11.56)	*94*
2	26 (22.99)	31 (21.44)	34 (39.77)	5 (11.81)	*96*
3	33 (28.50)	17 (26.57)	49 (49.29)	20 (14.63)	*119*
Total	*74*	*69*	*128*	*38*	*309*

- Using this we can calculate our $\chi 2$ test statistic!

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Do not forget the square!

Chi-square test for statistical association

- For the given problem:

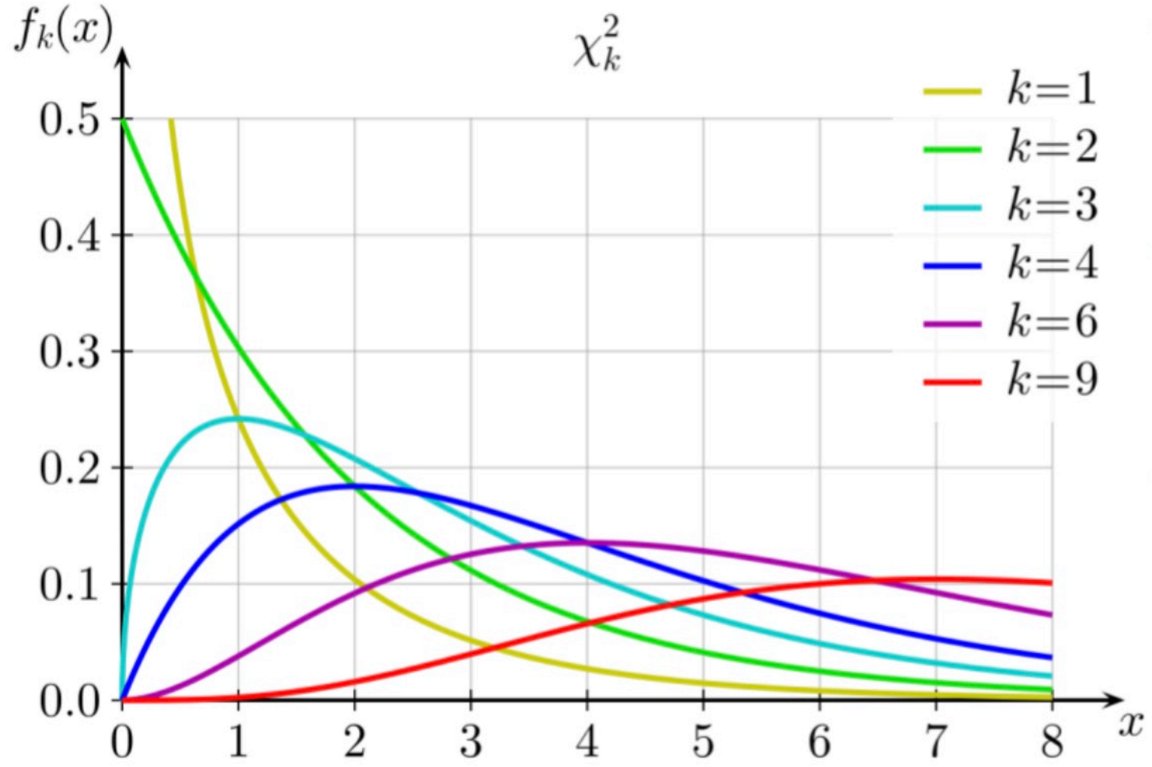
$$X^2 = \sum_{j=1}^{4}\sum_{i=1}^{3} \frac{[n_{ij} - \widehat{E(n_{ij})}]^2}{\widehat{E(n_{ij})}}$$

$$= \frac{(15 - 22.51)^2}{22.51} + \frac{(26 - 22.99)^2}{22.99} + \cdots + \frac{(20 - 14.63)^2}{14.63}$$

$$= 19.17.$$

- Chi-square degree of freedom is given by: (no. of rows-1)*(no. of cols-1) = (4-1)*(3-1) = 6

Chi-square Distribution



- Distribution depends on ***degrees of freedom*** (like the t-distribution)
- Chi-square is always≥ 0, because we used the squared differences.
- To determine the degrees of freedom:
  df (degrees of freedom) = (r-1)*(c-1)

p-value and Conclusion

```
[1]  from scipy import stats
```

```
[2]  1 - stats.chi2.cdf(19.17, 6)
```

```
0.003885957910757742 4
```

- P-value < α
- We reject the null hypothesis.

- Therefore, we can reject H_0 at 0.05 significance level and conclude that defect type and manufacturing shift are dependent.

p-value and Conclusion

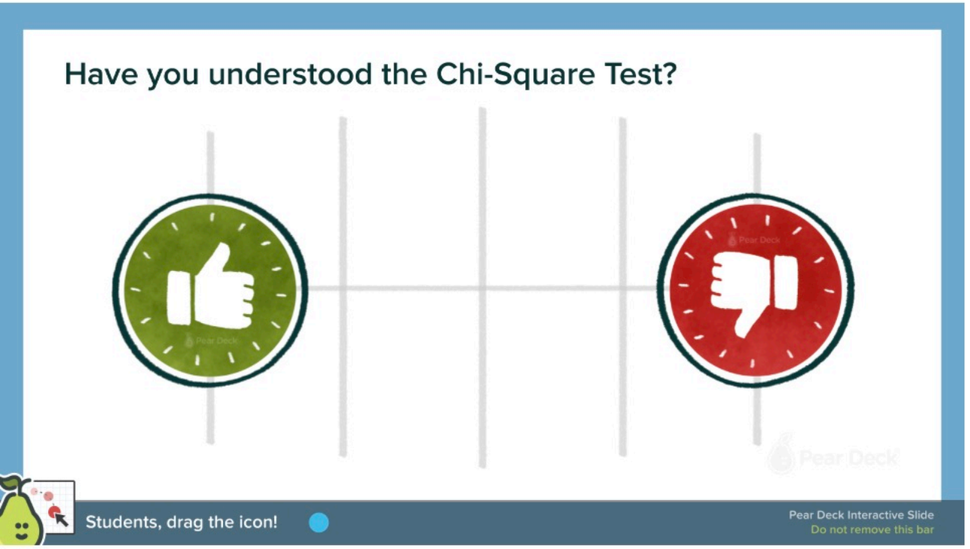
```
[5]  table = [[15, 21, 45, 13],[26, 31, 34, 5], [33, 17, 49, 20]]
```

```
[6]  stat, p, dof, expected = chi2_contingency(table)
```

```
stat=19.178, p=0.0039
```

- P-value < α
- We reject the null hypothesis.

- Therefore, we can reject H_0 at 0.05 significance level and conclude that defect type and manufacturing shift are dependent.



No Draggable™ Response

You didn't answer this question

THANKS!

Any questions?

You can find me at:

- jason@clarusway.com