BALL STATE
UNIVERSITY

Module 11 Lecture - Review of Prior Topics

Analysis of Variance

Quinton Quagliano, M.S., C.S.P

Department of Educational Psychology

Table of Contents

1 Overview and Introduction **2**
- 1.1 Textbook Learning Objectives . . . 2
- 1.2 Instructor Learning Objectives . . . 2
- 1.3 Introduction . . . 2

2 Basics of the Chi-Square Distribution **3**
- 2.1 Formula & Notation . . . 3
- 2.2 Distribution Characteristics . . . 3
- 2.3 The df for Chi-square Distribution . . . 4

3 Goodness-of-Fit Test **5**
- 3.1 Introduction . . . 5
- 3.2 Formula . . . 6
- 3.3 Null and Alternative Hypotheses Testing . . . 6
- 3.4 Application Example . . . 7

4 Test of Independence **8**
- 4.1 Introduction . . . 8
- 4.2 Formula . . . 9
- 4.3 Null and Alternative Hypotheses Testing . . . 9
- 4.4 Application Example . . . 9

5 Test of Homogeneity **10**
- 5.1 Introduction . . . 10
- 5.2 Formula . . . 11
- 5.3 Null and Alternative Hypotheses Testing . . . 11
- 5.4 Application Example . . . 12

6 Conclusion **13**
- 6.1 Recap . . . 13

Key Terms **13**

1 Overview and Introduction

1.1 Textbook Learning Objectives

- Interpret the chi-square probability distribution as the sample size changes.
- Conduct and interpret chi-square goodness-of-fit hypothesis tests.
- Conduct and interpret chi-square test of independence hypothesis tests.
- Conduct and interpret chi-square homogeneity hypothesis tests.

1.2 Instructor Learning Objectives

- Appreciate how the chi-square family of tests largely depart from the previous mean-based testing in t-tests
- Understand what scenarios each form of the chi-square test may be appropriate to
- See how the chi-square family of tests add to our toolbox of inferential tests

1.3 Introduction

- The χ^2 distribution, also known as chi-square distribution, or chi-squared distribution is a unique tool that can be _____ in inferential testing
  - Much like the _____, the chi-square distribution serves a purpose of _____
- There are 3 primary _____ for the chi-square distribution:
  - Goodness-of-fit: to compare if data is distributed as _____
  - Test of independence: to compare if two _____ are independent
  - Test of homogeneity: to test if two distribution are the _____ as one another

> **Important**
>
> Though not explicitly mentioned in the book, the chi-square distribution is particularly well-suited to helping us make sense of categorical data - but that is not the ONLY way it can be used

2 Basics of the Chi-Square Distribution

2.1 Formula & Notation

Discuss: Review: Write out the notation for Student's t-distribution

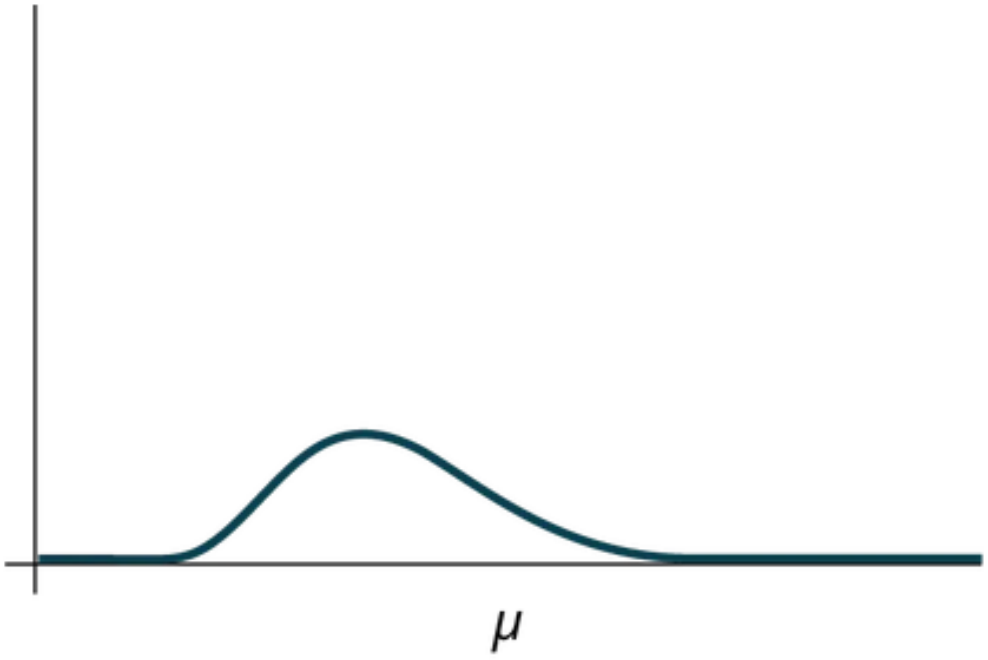
- Formula for χ^2 distribution:
  - $X \sim \chi^2_{df}$
  - Where:
    - df: degrees of _____
    - X: a _____ variable (much like other distributions, this can really be any _____ letter, X is just convention)
    - χ^2: arbitrary notation for the chi-square distribution

2.2 Distribution Characteristics

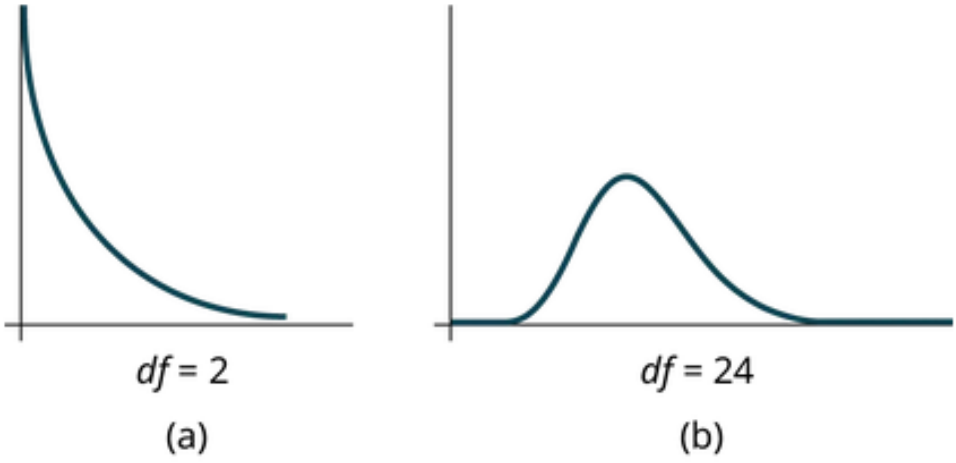
Discuss: Review: Think back to our initial discussions on the uniform and exponential distributions - write out the formulas for the find long-term mean and standard deviation in those distributions?

- Characteristics of the distribution:
  - $\mu = df$
  - $\sigma = \sqrt{2(df)}$
  - The curve is _____ and skewed to the right
  - The chi-square _____ for each df

- The test statistic for any test is always _____ than or equal to 0
- When $df > 90$, the chi-square distribution is approximately normal
- The μ is located to the near right of the "peak" of the curve



- The χ^2 distribution thus _____ dependent on the degrees of freedom



2.3 The df for Chi-square Distribution

- So how does one _____ the important df for this distribution?

- Depends on which of the 3 use cases we are _____

Discuss: Try explaining, in your own words, what degrees of freedom are again?

Important

The fact that the chi-square distribution changes dependent on the df is a unique change compared to the normal or t-distributions!

3 Goodness-of-Fit Test

3.1 Introduction

- The goodness-of-fit test is an application where the _____ distribution is used to determine whether data "fits" a certain distribution

Discuss: Why might we want to know if the data we have fit a certain distribution? hint: think about 'assumptions'

3.2 Formula

$$\chi^2 = \sum_k \frac{(O-E)^2}{E}$$

$$df = k - 1$$

- Where:
  - O = observed values are those that we actually _____ from our sample
  - E = expected values are those that we _____ our observed values to be if they follow the same distribution, i.e., if the _____ hypothesis is true
  - k = the number of different _____, you can think of this as the "pairs" of observed and expected data points

Discuss: A chi-square statistic characteristic is always equal to or greater than 0; explain why this characteristic is true by looking at the formula

Important

Due to some mathematical oddities, don't use this unless expected values are at least 5 or greater in each category - can inflate type I error

3.3 Null and Alternative Hypotheses Testing

Important

Null and alternative hypotheses can and should be stated when doing any and all inferential testing! Still applies here

- Hypothesis pair:
  - Null hypothesis: The _____ values/proportions are the same as the _____ values

  - Alternative hypothesis: The _____ values/proportions are different than the _____ values
- Goodness-fit-tests are usually done in _____ manner, as a greater statistic suggest a greater divergence from expected distribution

Discuss: Draw a normal curve and shade it in such a manner to show a right-tailed test

3.4 Application Example

- I have a class of 20 students, who I expect will have a uniform distribution of As, Bs, Cs, Ds, and Fs
  - $E : \{4, 4, 4, 4, 4\}$
  - $O : \{2, 4, 12, 1, 1\}$
- Hypothesis pair:
  - H_0 : the observed letter grades fits the expected distribution of 4 of each A, B, C, D, F
  - H_A : the observed letter grades do not fit the expected distribution of 4 of each A, B, C, D, F

Grade	Expected (E)	Observed (O)	(O − E)	' ' / E
A	4	2	-2	(−2)² / 4 = 1.00
B	4	4	0	0.00
C	4	12	8	8² / 4 = 16.00
D	4	1	-3	(−3)² / 4 = 2.25
F	4	1	-3	(−3)² / 4 = 2.25
Total	20	20	—	**21.50**

- $\chi^2 = 21.50, df = 5 - 1 = 4$
  - p to be determined by computer and compared against arbitrary α

Discuss: Do the same process as demonstrated above, but now using 1,1,1,2,15 as the observed data

4 Test of Independence

4.1 Introduction

- The test of independence test is used when we want to determine if two factors or _____ are independent or not
  - This test will use a contingency table that shows _____ of two variable in data

Discuss: Think back to our probability units, what did 'independent' mean in that context and how did it relate to sampling?

Important

Much like goodness-of-fit, don't use this unless expected values are at least 5 or greater in each category - can inflate type I error

4.2 Formula

$$\chi^2 = \sum_{i*j} \frac{(O-E)^2}{E}$$

$$df = (i-1) * (j-1)$$

- Where:
  - O : observed values
  - E : expected values*
  - i : number of _____ in the data
  - j : number of _____ in the data
- Expected value (E) for each cell of the table is calculated as:
  - $E = \frac{\sum_{row} * \sum_{column}}{\sum_{total}}$

4.3 Null and Alternative Hypotheses Testing

- Hypothesis pair:
  - Null hypothesis: The two factors are _____
  - Alternative hypothesis: The two factors are _____

4.4 Application Example

- I have a class of 40 students, half of whom are female and half are male, and below are their counts of of As, Bs, Cs, Ds, and Fs
  - F : $\{6, 5, 3, 5, 1\}$
  - M : $\{3, 4, 4, 7, 2\}$
- Hypothesis pair:
  - H_0 : The two factors of sex and grade are independent
  - H_A : The two factors of sex and grade are dependent
- Contingency Table:

Grade	Female (F)	Male (M)	Total
A	6	3	9
B	5	4	9
C	3	4	7
D	5	7	12
F	1	2	3
Total	**20**	**20**	**40**

- Expected Counts:

Grade	Female (E)	Male (E)
A	9×20/40 = 4.5	4.5
B	9×20/40 = 4.5	4.5
C	7×20/40 = 3.5	3.5
D	12×20/40 = 6.0	6.0
F	3×20/40 = 1.5	1.5

- Chi-square statistic:

Grade	Female $(O-E)^2/E$	Male $(O-E)^2/E$
A	$(6-4.5)^2/4.5 = 0.50$	$(3-4.5)^2/4.5 = 0.50$
B	$(5-4.5)^2/4.5 = 0.06$	$(4-4.5)^2/4.5 = 0.06$
C	$(3-3.5)^2/3.5 = 0.07$	$(4-3.5)^2/3.5 = 0.07$
D	$(5-6.0)^2/6.0 = 0.17$	$(7-6.0)^2/6.0 = 0.17$
F	$(1-1.5)^2/1.5 = 0.17$	$(2-1.5)^2/1.5 = 0.17$
Total $\chi^2 = 1.94$		

- $df = (5-1)(2-1) = 4$, p to be calculated by computer

Discuss: Do the same process as demonstrated above, but now using F: 1,1,1,2,15 and M: 2,2,2,10,4 as the observed data

5 Test of Homogeneity

5.1 Introduction

- Whereas the _____ test can tell us whether data matches a specified, _____ distribution, we have a different test for testing whether two different sets of data follow the same, unknown distribution
  - We call these are test for homogeneity

Discuss: In what hypothesis testing scenarios might we want to know whether two sets of data seem to have equal distributions? hint: think independent-samples t-test

5.2 Formula

Important

The chi-square statistic here is calculated the same as with the test of independence, but the df is different!

$$\chi^2 = \sum_{i*j} \frac{(O-E)^2}{E}$$

$$df = (i-1) * (j-1)$$

- Where:
  - O : observed values
  - E : expected values*
  - i : number of _____ in the data
  - j : number of _____ in the data
  - The _____ in this setup should be the categorical variable whose _____ we are interested in
  - The _____ should be the _____ categorical variable, or the two groups we compare against one another

5.3 Null and Alternative Hypotheses Testing

- Hypothesis pair:
  - Null hypothesis: The _____ of the two populations are the same

  - Alternative hypothesis: The distributions of the two populations are _____ the same

5.4 Application Example

- I have a group of students, half of whom are female and half are male, and below are their counts of what class they are in:
  - Freshman, Sophomore, Junior, Senior
  - $F : \{6, 14, 8, 12\}$
  - $M : \{5, 7, 7, 8\}$
- Hypothesis pair:
  - H_0 : The distributions of class between male and female students are the same
  - H_A : The distributions of class between male and female students are not the same
- Data setup

Class	Female (F)	Male (M)	Total
Freshman	6	5	11
Sophomore	14	7	21
Junior	8	7	15
Senior	12	8	20
Total	**40**	**27**	**67**

- Expected counts

Class	Expected F	Expected M
Freshman	6.57	4.43
Sophomore	12.54	8.46
Junior	8.96	6.04
Senior	11.94	8.06

- Chi-square calculation

Class	Group	(O)	(E)	((O-E)^2/E)
Freshman	F	6	6.57	0.049
Freshman	M	5	4.43	0.074
Sophomore	F	14	12.54	0.170
Sophomore	M	7	8.46	0.252
Junior	F	8	8.96	0.103
Junior	M	7	6.04	0.151

Class	Group	(O)	(E)	((O-E)^2/E)
Senior	F	12	11.94	0.0003
Senior	M	8	8.06	0.0004

6 Conclusion

6.1 Recap

- The chi-squared distribution is well suited to dealing with categorical counts and proportions, but the 3 forms of the chi-square distribution have somewhat different application and formulas
- The shape of the chi-square distribution changes based upon the dfs, and the hypothesis testing done with it is usually right-tailed, among other consistent characteristics
- Like with the other distributions we have discussed, be mindful of how the null and alternative hypotheses change

Key Terms

C

chi-square distribution A unique distribution in which the mean is equal to the degrees of freedom (df) and the standard deviation $\sqrt{2(df)}$ 2
contingency table A table with a variable on either axis to show counts of occurrence for each level of those variables 8

E

expected values Those values which we arbitrarily expect or test against 6

G

goodness-of-fit test An inferential statisical test using the χ^2 distribution, tests whether data appears to fit or be the same as a pre-defined distribution 5

O

observed values Those which we gather/measure as part of an study or sample 6

T

test for homogeneity An inferential statistical test using the χ^2 distribution, tests whether data appears to have the same distribution as another set of data 10
test of independence test An inferential statisical test using the χ^2 distribution, tests whether data appears to independent of another set of data 8

The instructor-provided glossary may not include all terms worth memorizing, make sure you consider using the vocabulary list in you book and your own judgment to make sure you have all relevant terms