Chi-squared

ANA 500 – Foundations of Data Analytics

Module 2 - week 4B

Chi-squared

- The Chi-squared distribution is a member of the normal family of distributions and often arises as the sampling distribution for a test statistic.

- We will use the Chi-squared distribution for:
  1) **Goodness of fit tests**
     - E.g.,we might be interested in whether the frequency of students across majors is equally distributed.
  2) **Tests of independence**
     - E.g.,we might be interested in whether gender and being a STEM major are related were independent.

Chi-squared

- We will use the Chi-squared distribution for:

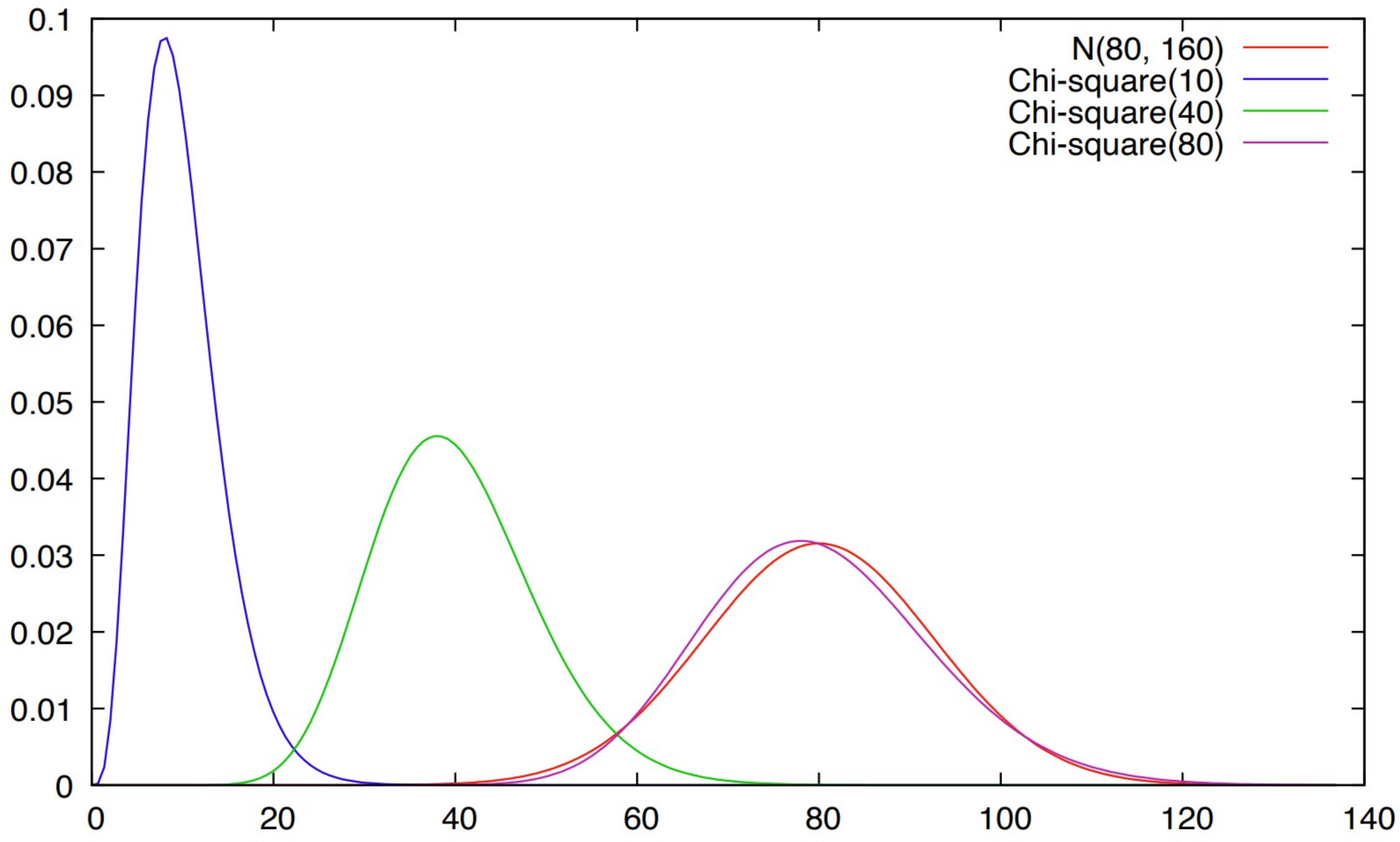
  3) Tests of homogeneity
    - Test whether the distribution of a categorical variable is the same across different populations
    - E.g., we might be interested in whether the distribution of letter grades is the same between STEM and non-STEM majors.

  4) Tests of a single variance
    - Test whether a population variance is equal to a specified value.
    - E.g., is the variants in the amount of soda dispensed by a machine at McDonald‘s, equal to a certain value as you can

Chi-squared

- A Chi-squared distribution has a mean equal to df and standard deviation equal to $\sqrt{2df}$
- A Chi-squared random variable with k degrees of freedom is the sum of k independent, squared standard normal distributions.
- A Chi-squared distribution is skewed to the right and the skew decreases as df increases.
- The test statistic is always greater than or equal to zero. The reason is because the chi square distribution arises from the sum of squared, normally distributed variables.

Chi-squared



N(80, 160)
Chi-square(10)
Chi-square(40)
Chi-square(80)
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
20
40
60
80
100
120
140

Goodness of fit test

- Do the data fit a particular distribution?
- Often used for categorical data
- H_o: data fit the expected distribution
- H_a: data do not fit the expected distribution
- The test involves comparing expected frequencies (*E*) with observed frequencies (*O*).
- Let k = the number of categories of the variable of interest

Goodness of fit test

- Test statistic = $\sum_K \frac{(O-E)^2}{E}$
- $df = k - 1$

- The test statistic is always positive
- As a general rule of thumb, we want the expected value for each category to be greater than or equal to five.

Goodness of fit test

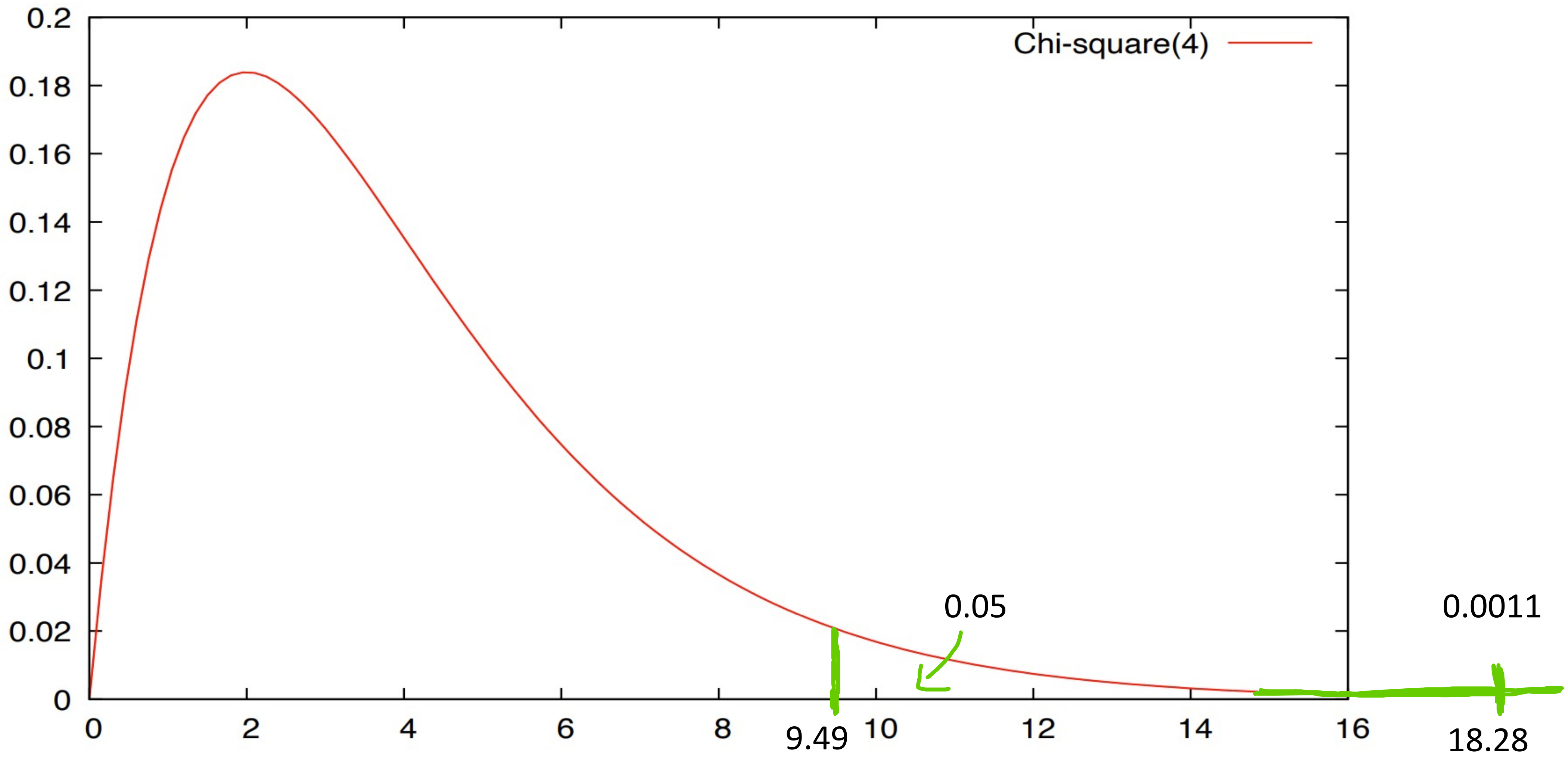
- E.g. The following table shows the distribution of grades a professor expects in a class of 100 students as well as the grades she observes.

Grade	Expected (E)	Observed (O)	(O-E)^2	((O-E)^2)/E
A	20	11	81	4.05
B	30	38	64	2.13
C	30	42	144	4.8
D	10	7	9	0.90
F	10	2	64	6.4

$$\sum_{K} \frac{(O-E)^2}{E}$$

- Test-stat = 4.05 + 2.13 + 4.8 + 0.90 + 6.4 = 18.28
- df = k – 1 = 5 – 1 = 4
- p-value = 0.0011
- 0.0011 < 0.05 --- reject Ho, the observed grades do not fit the expected distribution

Goodness of fit test



Test of independence

- Are two variables independent?
- Often used for nominal variables
- Ho: the two variables are independent
- Ha: the two variables are not independent
- Test statistic = $\sum_K \frac{(O-E)^2}{E}$
- $df = (\mathit{number\ of\ rows} - 1) * (\mathit{number\ of\ columns} - 1)$
- Expected frequencies (E) won't always be obvious.
  - Calculate using (row total)*(column total) / (total number of observations)
  - Row and column totals are often referred to as row and column marginals

Test of independence

- E.g. Is there a relationship between gender and being a STEM major?

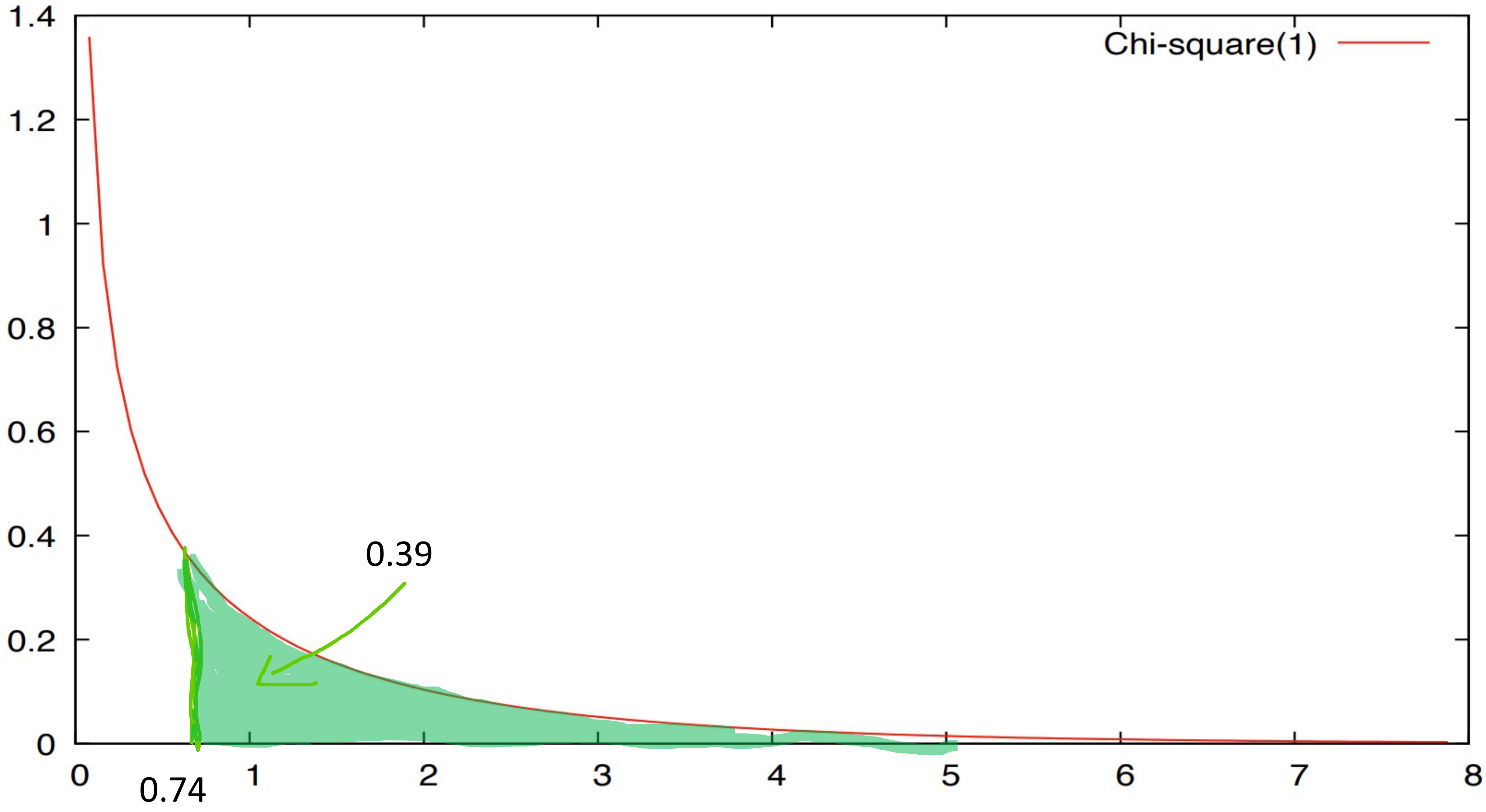
	STEM	Non-STEM	Total
Male	22	47	69
Female	18	53	71
Total	40	100	140

e.g., (40*69)/140=19.7

Observed (O)	Expected (E)	(O – E)	(O – E)^2	((O – E)^2)/E
22	19.7	2.3	5.3	0.27
18	20.3	-2.3	5.3	0.26
47	49.3	-2.3	5.3	0.11
53	50.7	2.3	5.3	0.10

- Test-stat = 0.27 + 0.26 + 0.11 + 0.10 = 0.74
- df = (2 – 1)*(2 – 1) = 1
- p-value = 0.39
- 0.39 > 0.05 --- fail to reject Ho, gender and being a STEM major are independent.

Test of independence



Tests of homogeneity

- The goodness of fit test is used to determine if the data fit a particular distribution, but it won't suffice for determining whether two variables follow the same unknown distribution.
- Test for homogeneity
  - H_o: distributions are the same
  - H_a: distributions are different
  - Test statistic is calculated in same way as for the goodness of fit test.
  - *df* = number of columns – 1
  - Comparing a single qualitative variable with more than 2 categories across two populations
  - All values in table must be greater than or equal to 5

Tests of homogeneity

- E.g. Is there a difference in favorite professional sport to watch between those living on the east coast versus the west coast?

	Baseball	Football	Basketball	Total
East coast	18	22	14	54
West coast	20	19	17	56
Total	38	41	31	110

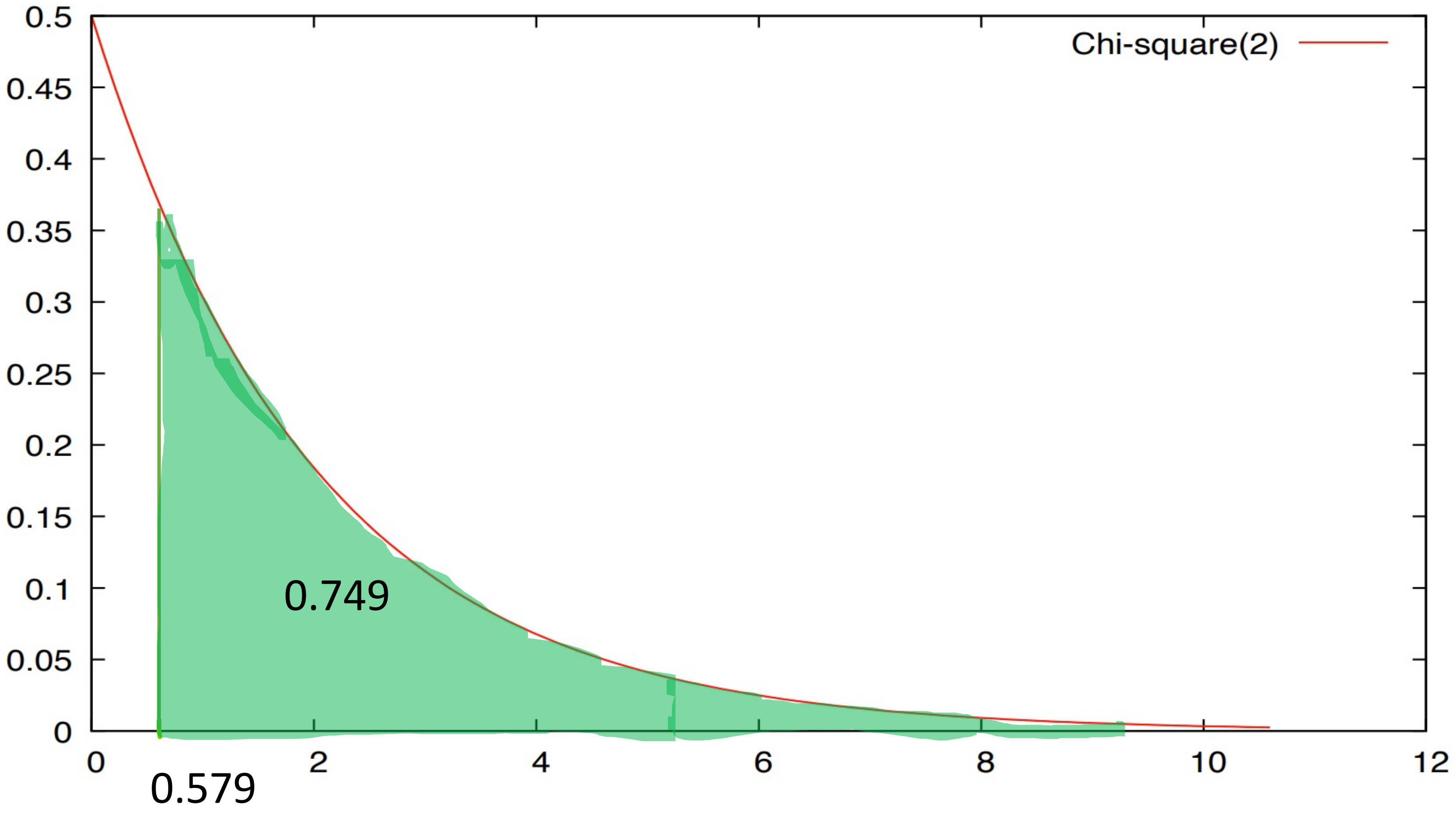
Tests of homogeneity

e.g., (38*54)/110=18.65

Observed	Expected	(O – E)^2	((O – E)^2)/E
18	18.65	0.42	0.023
20	19.35	0.42	0.022
22	20.13	3.5	0.174
19	20.87	3.5	0.168
14	15.22	1.49	0.098
17	15.78	1.49	0.094

- Test-statistic = 0.023 + 0.022 + 0.174 + 0.168 + 0.098 + 0.094 = 0.579
- p-value = 0.749
- 0.749 > 0.05 --- fail to reject Ho, there is not a statistically significant difference between east and west coast in the distribution of favorite pro sports

Tests of homogeneity



Test of a single variance

- Assume underlying population is normal
- H_o: the population variance is equal to a specified value
- H_a: the population variance is not equal to or greater than or less than the specified value

- Test-statistic $= \frac{(n-1)s^2}{\sigma^2}$
- $df = n - 1$

Test of a single variance

- E.g. Is the variance in waiting time at the DMV greater than 10 minutes?
- Ho: $\sigma^2 = 10$
- Ha: $\sigma^2 > 10$
- Suppose in a random sample of 30 people at the DMV the sample variance is calculated and equal to 12.
- Test-statistic $= \frac{(n-1)s^2}{\sigma^2} = \frac{(30-1)*12}{10} = 34.8$
- p-value = 0.21
- 0.21 > 0.05 --- fail to reject Ho, the variance in wait times is not greater than 10 minutes.

Test of a single variance