1 Inference for Categorical Data: Chi-Square

1.1 Chi Square Test for Goodness of Fit

Expected Counts

- A goodness-of-fit test is used to test the hypothesis than an observed frequency distribution fits to some claimed distribution.
- An example of an equal expected count might be the following:
- A fair, six-sided die is rolled 60 times. What would be the expected count of each outcome?

Outcome	1	2	3	4	5	6
Expected Count	10	10	10	10	10	10

We found the expected count by taking the total number of trials and dividing it equally among each of the outcomes.

- An example of an unequal expected count might be the folowing:
- An unfair, six-sided die is rolled 60 times. The die is loaded so that the number 1 turns up 50% of the time and the other five outcomes occur 10% of the time. What would be the expected count of each outcome?

Outcome	1	2	3	4	5	6
Probability	.50	.10	.10	.10	.10	.10
Expected Count	30	60(.10)=6	6	6	6	6

We found the expected count by taking the total number of trials and multiplying it by the probability of the outcome.

Observed Counts:

While we know what is expected, when we run a simulation of rolling a fair die 60 times, we do not expect each outcome to be observed exactly 10 times each due to sampling variability.

Let's say a die was give to you claimed as fair, and you observed the following counts when rolling the die 60 times:

Outcome	1	2	3	4	5	6
Expected Count	10	10	10	10	10	10
Observed Count	16	7	15	10	4	8

- Do you suspect the die given to you was a fair die?
- This is going to be the question that our new test helps us answer: Are the differences between the actual observed counts and the expected counts significant?

Note: The observed counts must be all whole numbers because they represent actual counts, but the expected counts do not need to be whole numbers.

To measure the difference between the observed and expected counts, and to determine if the difference is

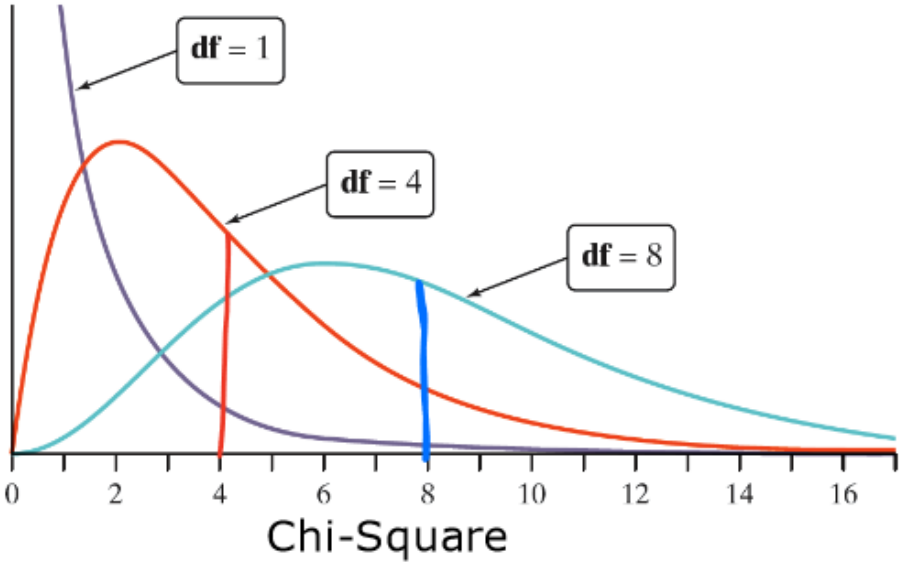
significant, we will introduce a new test statistic, called the chi-square statistic:

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

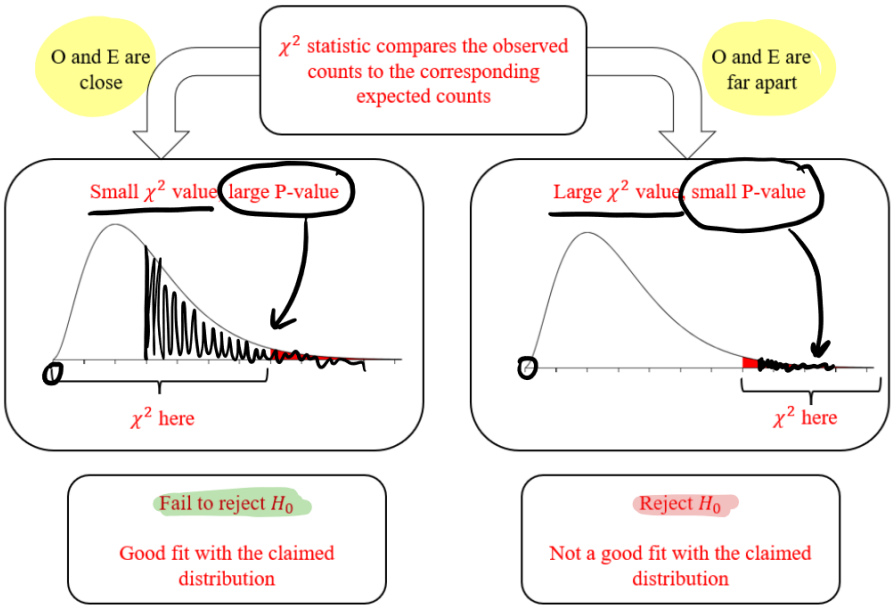
Where O represents each observed count in the distribution and E represents each corresponding expected count.

- The sampling distribution of the chi-square statistic is not a normal distribution
- It is a right-skewed distribution that allows only for positive values because the statistic cannot be negative.

When the expected counts are all at least 5, the sampling distribution of the χ^2 statistic is close to a chi-square distribution with degrees of freedom (df) equal to the number of categories minus 1.



- The chi-square distributions are a family of distributions that take only positive values and are skewed to the right.
- A particular chi-square distribution is specified by giving its degrees of freedom.
- The chi-square goodness-of-fit test uses the chi-square distribution with df = #categories - 1



Hypotheses

- The null hypothesis in a chi-square goodness-of-fit test should state a claim about the distribution of a single categorical variable in the population of interest.
- We can write this in words or symbols; both are acceptable and used on the AP exam.
- Using our fair die as an example, we would say:

Using symbols, $H_0 : p_1, p_2, p_3, p_4, p_5, p_6 = \frac{1}{6}$, where p is the proportion of outcomes of each face of die.

Using words, H_0: The proportion of dice outcomes is equally distributed.

The alternative hypothesis in a chi-square goodness-of-fit test is the categorical variable does not have the specified distribution, and is easily given in words:

H_A: At least one of the claimed proportions is incorrect.

Conditions:

- Random: The data came from a well-designed random sample or randomized experiment
- Independent: When sampling without replacement, the 10% condition is met.
- Large Counts
  - All expected counts are at least 5
  - This allows us to say that the sampling distribution will follow a Chi-Square distribution

When the conditions are met, the chi-square goodness of fit test can be performed with the hypotheses:

H_0: The claimed distribution is correct. H_A: At least one proportion in the claimed distribution is incorrect.

We find the expected counts assuming the claimed distribution is true, and then we calculate the chi-square statistic:

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

The p-value is the area to the right of χ^2 under the density curve of the chi-square distribution with k-1 degrees of freedom, where k represents the number of categories.

$$\text{P-value} = \text{P}(\chi^2 > \text{value}) = \chi^2\text{cdf(value, 1e99, df) on the TI-84}$$

Beware

1. The chi-square test statistic compares observed and expected counts. Don't try to perform calculations with the observed and expected proportions in each category.
2. When checking the Large Counts condition, be sure to examine the expected counts, not the observed.

Example

A geneticist is studying the gene pattern of eye color in a group of white mice. He observed a random sample of mice from the lab and found that 110 had red eyes, 57 had brown eyes, 32 had pink eyes, and 13 had blue eyes. His model suggest that this distribution of eye color should occur in a $9:3:3:1$ ratio. Is there evidence that the geneticist's model is not accurate? Use a 5% level of significance.

H_0: The proportion of white mice eye color occurs in a $9:3:3:1$ ratio.

H_A: At least one of the proportion is incorrect.

- Random: Random sample of 212 white mice
- Independent: $n = 212 \leq 0.10$(all white mice)
- Large Counts:

	Red	Brown	Pink	Blue
OBS	110	57	32	13
EXP	119.25	39.75	39.75	13.25

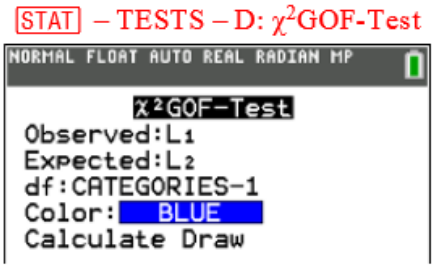
All expected counts ≥ 5.

Chi Square Test for Goodness of Fit

$\chi^2 = \frac{(110-119.25)^2}{119.25} + \frac{(57-39.75)^2}{39.75} + \frac{(32-39.75)^2}{39.75} + \frac{(13-13.25)^2}{13.25} = 9.7191.$

Doing χ^2cdf, with (lower: 9.7191, upper: ∞, df: 3) gives 0.0211 for the p value.

We can also use a calculator:



Since the p-value of 0.0211 is less than $\alpha = 0.05$, we reject the null. There is convincing evidence that the color of white mice eyes does not match a $9:3:3:1$ ratio.

1.2 Chi Square Test for Homogeneity

- Recall: The Chi Square Test for Goodness of Fit is testing if one population "fits" a given claim.
- The Chi Square Test for Homogeneity compares the distributions of one categorical variable across two or more populations to see if they are the same or different

Constructing a Chi-Square Test for Homogeneity:

- State the Hypotheses:
  - H_0: There is no difference in the distribution of categorical variable among two or more groups.
  - H_A: There is a difference in the distribution of categorical variable among two or more groups.
- Check the Assumptions and Conditions
  - Randomness: The individuals whose counts are avaliable for analysis should be a random sample of the population.
  - 10% Condition: The sample size, n, must be no larger than 10% of the population.
  - Large Counts: We should expect to see at least 5 counts in each category of the categorical variable. List the counts and state "all exp counts ≥ 5"

  Note: When performing an experiment, Randomness is satisfied by randomly assigning treatments to subjects and the 10% condition is satisfied if we can assume independence among the individuals in the study.
- Name the Inference Method: Chi-Square Test for Homogeneity
- Calculate the Test Statistic

$$\chi^2 = \sum \frac{(\text{observed-expected})^2}{\text{expected}}$$

Observed Counts - Actual frequencies of the variable from your sample

Expected Counts - Projected frequencies of the variable if the null hypothesis is true

$$\text{EXP} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}}$$

- Obtain the P-value 2nd-Vars-8:χ^2cdf()
  - Lower: χ^2
  - Upper: ∞
  - df: (#Rows-1)(#Columns-1)

  Calculator Steps:

  1. Enter Observed Values into Matrix [A]
     - 2nd-x^{-1}(Matrix)-EDIT
     - #Rows x # Columns
  2. Enter Expected Values into Matrix [B]
     - Will populate when you run the χ^2 Test
  3. STAT-Tests-C:χ^2Test
     - Observed: [A]
     - Expected: [B]
- Make a decision: This is the same as always
- State your conclusion in context: This is also the same

Example

Andrea is addicted to TikTok and she believes that most students in her high school are too. She wants to investigate if there is a difference in TikTok usage among Juniors and Seniors. She randomly samples 65 Juniors and 85 Seniors. The data she collected is given in the table below. Is there convincing evidence of a difference in the distribution of TikTok among the Juniors and Seniors in her school?

↓ TikTok Use ↓	Juniors	Seniors	Total
Once a Month or Less	4	10	14
Once a Week	15	21	36
At Least Once a Day	46	54	100
Total	65	85	150

H_0: TikTok usage among juniors and seniors is the same

H_A: TikTok usage among juniors and seniors is different.

- Random: Randomly sampled 65 juniors and 85 seniors.
- Independent: $65 \leq 0.10$(all juniors in her HS), $85 \leq 0.10$(all seniors in her HS)
- Large Counts: $\begin{bmatrix} 6.07 & 7.93 \\ 15.6 & 20.4 \\ 43.33 & 56.67 \end{bmatrix}$ All exp ≥ 5
- χ^2 Test for Homogeneity
- $\chi^2 = 1.5727$, $p = 0.4555$, df $= 2$: $(\#\text{col} - 1)(\#\text{rows} - 1) = 2$.
- Since the p-value of 0.4555 is greater than $\alpha = 0.05$, we fail to reject the null. There is not convincing evidence that TikTok usage among juniors and seniors is different at Andrea's high school.

Example

Aspirin prevents blood from clotting which helps prevent strokes. A medical study (we will assume this is a well-designed experiment) asked whether adding another anti-clotting drug named Dipyridamole would be more effective for patients who already had a stroke. Here are the data on strokes during the two years of the study.

Group	Treatment	Number of Patients	Number of Patients w/Stroke
1	Placebo	1649	250
2	Aspirin	1649	206
3	Dipyridamole	1654	211
4	Both	1650	157

(a) Summarize the data into a two-way table.

	Placebo	Aspirin	Dipyridamole	Both	Total
Stroke	250	206	211	157	824
No Stroke	1399	1443	1443	1493	5778
Total	1649	1649	1654	1650	6602

(b) Is there convincing evidence of a difference in the effectiveness of the four treatments at the $\alpha = 0.05$ significance level?

- H_0: No difference in effectiveness among treatments
- H_A: Difference in effectiveness among treatments.
- Random: Treatments randomly assigned - medical study
- Independent: Assume patient results are independent
- Large Counts: EXP $= \begin{bmatrix} 205.81 & 205.81 & 206.44 & 205.94 \\ 1433.19 & 1433.19 & 1447.56 & 1444.06 \end{bmatrix}$ All EXP ≥ 5
- χ^2 Test for Homogeneity
- $\chi^2 = 24.2428$, $p = 0.00002$, df $= 3$
- Since the p-value of 0.00002 is less than $\alpha = 0.05$, we reject the null. There is convincing evidence of a difference in effectiveness among the four treatments for preventing strokes.

1.3 Chi Square Test for Independence

- Recall: The Chi Square Test for Goodness of Fit is testing is one population "fits" a given claim.
- Recall: The Chi Square Test for Homogeneity compares the distributions of one categorical variable across two or more populations to see if they are the same or different
- The Chi Square Test for Independence compares the distribution of two categorical variables across one population to see if they are independent (not associated)

Constructing a Chi-Square test for Independence:

- State the Hypotheses:
  - H_0: Categorical Variable 1 and Categorical Variable 2 are independent (not associated) for population
  - H_A: Categorical Variable 1 and Categorical Variable 2 are not independent (associated) for population
- Check Assumptions and Conditions: Same as the test for Homogeneity

- Name the Inference Method: Chi-Square Test for Independence
- Calculate the Test Statistic: Same as Homogeneity
- Obtain the P-Value: This is also the same as Homogeneity
- Make Decision: Same as always
- State your conclusion in context: This remains the same

Don't Forget: Association does not imply causation

- A small p-value is not proof of causation.
- The Chi Square Test for Independence treats the two variables symmetrically, we cannot differentiate the direction of any possible causation even if it existed.
- There is no way to eliminate the possibility that a lurking variable is responsible for the lack of independence.
- Don't say that one variable "depends" on the other just because they are not independent.

Example

Andrew thinks there might be a relationship between angry students and GPA. He asks, "Do students who are prone to sudden bursts of anger have lower GPAs?" He took an SRS of 300 students at his high school at the beginning of the year. He had each student take the Spielberger Trait Anger Scale Test which measures how prone a person is to sudden anger. At the end of the school year, Andrew collected the data on student GPAs. Here are the results:

Anger Scale Test Results

		Low Anger	Moderate Anger	High Anger
G	**Low GPA (0.1 – 1.9)**	4	14	37
P	**Mid GPA (2.0 – 2.9)**	122	33	3
A	**High GPA (3.0 – 4.0)**	79	7	1

= OBS

Does the data provide convincing evidence of an association between anger level and GPAs at Andrew's HS?

H_0: Anger level and GPA are independent for students at Andrew's HS.

H_A: Anger level and GPA are not independent for students at Andrew's HS.

- Random: SRS Of 300 students at Andrew's HS
- Independent: $n = 300 \leq 0.10$(all students at Andrew's HS)
- Large Counts: $\begin{bmatrix} 37.58 & 9.90 & 7.52 \\ 107.97 & 28.44 & 21.59 \\ 59.45 & 15.66 & 11.89 \end{bmatrix}$ All exp counts ≥ 5

Chi Square Test for Independence

$\chi^2 = 187.1097$, $p \approx 0$ df $= 4$

Since the p-value is approx. 0 is less than $\alpha = 0.05$, we reject the null. There is convincing evidence that anger level and GPA are not independent for students at Andrew's HS.

Example

Is your index finger longer than your ring finger? Or is it the other way around? It isn't the same for everyone! To investigate if there is a relationship between gender and relative finger length, we selected a random sample of 460 U.S. high school students who completed a survey. The results are shown in the table below:

		Gender	
		Male	**Female**
Relative Finger Length	**Index Longer**	85	73
	Same Length	42	44
	Ring Longer	100	116

= OBS

H_0: Gender and finger length are not associated for US HS Students

H_A: Gender and finger length are associated for US HS Students

- Random Sample of 460 US HS Students
- Independent: $n = 460 \leq 0.10$(all US HS Students)
- Large Counts: EXP $= \begin{bmatrix} 77.97 & 80.03 \\ 42.44 & 43.56 \\ 106.59 & 109.41 \end{bmatrix}$ All exp counts ≥ 5
- Chi Square Test for Independence
- $\chi^2 = 2.0652$, $p = 0.3561$, df $= 2$
- Since the p-value of 0.3561 is greater than $\alpha = 0.05$, we fail to reject the null. There is not convincing evidence that gender and finger length are associated for US HS students.