Chi-Square Test for

Goodness of fit — one variable, one population

Time of Birthday NHL players

Homogeneity — one variable but 2+ populations

Independence — two variables but only 1 population

Require Two-Way Tables

Suppose we wanted to know if the gender of an interviewer could affect the **responses to a survey question**. The subjects in their experiment were 100 males from their school.

Half of the males were randomly assigned to be asked, "Would you vote for a female president?" by a female interviewer. The other half of the males were asked the same question by a male interviewer.

Gender of interviewer

Response to question		Male	Female	Total
	Yes	30	39	69
	No	8	3	11
	Maybe	12	8	20
	Total	50	50	100

Suppose we wanted to know if the gender of an interviewer could affect the **responses to a survey question**. The subjects in their experiment were 100 males from their school.

Half of the males were randomly assigned to be asked, "Would you vote for a female president?" by a female interviewer. The other half of the males were asked the same question by a male interviewer.

Population

Population 2

Gender of interviewer

		Male	Female	Total
Response to question	Yes	30	39	69
	No	8	3	11
	Maybe	12	8	20
	Total	50	50	100

Variable

Lesson 11.2 (Day 1): **Does gummy bear brand matter?**

Is the distribution of gummy bear color the same for **Haribo** gummy bears and **Great Value** gummy bears? We'll collect data as a class and determine if we have convincing evidence

1. Add your data to the board and fill in the table below with the class totals.

	Counts of Haribo	Total
Red	12, 9, 13 9 13	64
Green	6 2 4 5 7 6	30
Yellow	3 5 5 6 7 7	33
Orange	3 5 3 5 1 1	18
White	3 3 8 3 5 7	29
		174

Haribo

	Counts of **Great Value**	Total
Red	10 5 6 7 9 3 6 8 8 12 8 7	89
Green	4 5 3 3 7 3 1 4 6 4 8	50
Yellow	3 3 2 6 6 4 6 5 2 6 6 5	54
Orange	3 4 4 3 4 6 3 4 3 4 3 3	44
White	3 3 2 4 4 2 5 5 6 5 4 1	44
		281

Great Value

1. Add your data to the board and fill in the table below with the class totals.

Observed:

Brand

Color	Haribo	Great Value	Total
Red	64	89	153
Green	30	50	80
Yellow	33	54	87
Orange	18	44	62
White	29	44	73
Total	174	281	455

2. **How many samples do we have? What population are they from?**

2 samples

1 from Haribo
1 from Great Value

3. **How many variables are we examining?** 1 variable - Color

4. **As a class, write down hypotheses for a significance test.**

H_0: There is no difference in the true distributions of color between Haribo an Great Value.

H_a: There is a difference in the true

5. Now we will use a chi-square test (of Homogeneity) to test if there is a difference between the two populations. We first need to find the expected values.

Expected:

Expected percent of total

Fill in totals first

Color	Brand: Haribo	Great Value	Total
Red	58.5	94.5	153
Green	30.6	49.4	80
Yellow	33.3	53.7	87
Orange	23.7	38.3	62
White	27.9	45.1	73
Total	174	281	455

$$\text{Expected Counts} = \frac{\text{Row total} \times \text{Column Total}}{\text{Table Total}}$$

6. On the back side, continue with a 4-step significance test.

STATE: Hypotheses:

Significance level:

$\alpha = 0.5$

PLAN: Name of procedure: Chi-square test for homogeneity

Check conditions:

Random: We randomly selected gummies

10%: Haribo total sample $174 < \frac{1}{10}$ (all Haribo pop)

Great Value total samp. $281 < \frac{1}{10}$ (all Great Value pop)

Large Counts: All expected counts ≥ 5 (see table)

Chi-Square Test for Homogeneity

Conditions for Performing a Chi-Square Test for Homogeneity

Random: The data come from *independent* random samples OR from groups in a randomized experiment.

10%: When sampling without replacement, $n < 0.10N$ for each sample.

Large Counts: All *expected* counts are at least 5.

Chi-Square Test for Homogeneity

Suppose the conditions are met. To perform a test of

H_0: There is no difference in the distribution of a categorical variable for several populations or treatments

compute the chi-square test statistic

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

where the sum is over all cells (not including totals) in the two-way table. The *P*-value is the area to the right of χ^2 under the chi-square density curve with degrees of freedom = (num. of rows – 1)(num. of columns – 1).

DO step

DO: Specific Formula: $\chi^2 = \sum \frac{(O-E)^2}{E}$

df = (rows – 1)(columns – 1)

$= (5-1)(2-1)$

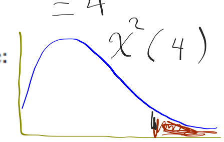
$= 4$

Work:

$$= \frac{(64-58.5)^2}{58.5} + \frac{(89-94.5)^2}{94.5} + \dots$$

$$= 3.15$$

Picture:



Test statistic: 3.15

P-value:

$$P(\chi^2 > 3.15) = \chi^2\text{cdf}[\underset{L}{3.15}, \underset{U}{10000}, \underset{df}{4}] = .533$$

Add p. 733 T.Tip

CONCLUDE:

Because the P-Value of .53 > $\alpha = 0.05$
we fail to ~~reject~~ reject H_0

∴ There IS NOT convincing evidence that there is a difference between the true distributions of color of Haribo an Great Value Gummies.

7. Explain how this test is different from a chi-square test for goodness of fit?

8. Interpret the P-Value you calculated above:

Assuming ______________________________ in the true distributions of the color for Haribo and Great Value, there is a _______ probability of observing differences in the distributions of responses as ___________________________________ than the ones in the study.

7. Explain how this test is different from a chi-square test for goodness of fit?

We have two samples from two populations.
(Haribo and Great Value)

χ^2-GOF has one sample from 1 population
(compared to a known distrib)

8. Interpret the P-Value you calculated above:

Assuming no difference *in the true distributions of the color for Haribo and Great Value, there is a* .52 *probability of observing differences in the distributions of responses as* large or larger *than the ones in the study.*

Important ideas:

Hypotheses:

Expected counts

χ^2 Homogeneity

χ^2 GOF

Important ideas:

Hypotheses: H_0: There is no difference in the distribution in the categorical variable distribution for pop. 1 and pop. 2.

H_a: There is a difference...

Expected Counts

χ^2 Homogeneity

χ^2 GOF

Important ideas:

Hypotheses: H_0: There is no difference in the distribution in the categorical variable distribution for pop. 1 and pop. 2.

H_a: There is a difference...

Expected Counts $= \dfrac{\text{Row Total} \times \text{Column Total}}{\text{Table Total}}$ $df = (\text{rows}-1)(\text{columns}-1)$

χ^2 Homogeneity

χ^2 GOF

Important ideas:

Hypotheses: H_o: There is no difference in the distribution in the categorical variable distribution for pop. 1 and pop. 2.

H_a: There is a difference...

Expected Counts $= \dfrac{\text{Row Total} \times \text{Column Total}}{\text{Table Total}}$ $df = (\text{rows}-1)(\text{columns}-1)$

χ^2 Homogeneity 2 samples, 1 variable

χ^2 GOF 1 sample, 1 variable

Music Preferences between

Californians and Michigonians ?

Michiganites ?

What is your music preference? *The chi-square test for homogeneity*

Do high school students in Michigan and California have the same music preferences? We used the Census at School® website to select separate random samples of 100 high school students from Michigan and 100 high school students from California. Students were asked, "What is your favorite music genre?" The two-way table summarizes their responses.

		State		
		Michigan	California	Total
Favorite music genre	Country	12	4	16
	Pop	15	14	29
	Rap	21	22	43
	Rock	7	10	17
	Other	45	50	95
	Total	100	100	200

Do these data provide convincing evidence at the $\alpha = 0.05$ level that the distributions of favorite music genre differ for high school students in Michigan and California?

STATE

STATE

H_0: There is no difference in the distributions of favorite music genre for high school students in Michigan & Calif.

H_a: There is a difference in the distributions of favorite music ...

$\alpha = 0.05$

PLAN Chi-Square test for Homogeneity

Random: Independent random samples of students from Michigan and Calif.

10%: $100 < 10\%$ of all Michigan high school students

$100 < 10\%$ of all Calif. High School Students

Large Counts: All expected Counts ≥ 5 (See table)

Expected Values

State

Fav. Music Genre

	Mich.	Calif	Total
Country			16
Pop			29
Rap			43
Rock			17
Other			95
Total	100	100	200

Expected Values

State

Fav. Music Genre

	Mich.	Calif	Total
Country	8	8	16
Pop	14.5	14.5	29
Rap	21.5	21.5	43
Rock	8.5	8.5	17
Other	47.5	47.5	95
Total	100	100	200

Can use Matrices

page 735 -736

For this test to run properly, there must be at least 2 rows and at least 2 columns

(• will not work for χ^2 for Goodness of fit)

DO

Test Statistic

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

$$\chi^2 = \frac{(12-8)^2}{8} + \frac{(4-8)^2}{8} + \frac{(15-14.5)^2}{14.5} + \ldots$$

$$= 4.85$$

df $= (5-1)(2-1) = 4$

P-Value $\chi^2_{cdf}[\underset{\text{Lower}}{4.85}, \underset{\text{Upper}}{10000}, \underset{\text{df}}{4}] = 0.303$

CONCLUDE

Because the P-Value of $0.303 > \alpha = .05$ we fail to reject Ho.

There is not convincing evidence of a difference in the distributions of fav. music genre for H.S. Students in Michigan and California.

11.2.....27-35 (odds)