χ^2 Contingency Test:

• Tests goodness-of-fit to the data of the null hypothesis of independence of variables

• Two categorical variables but, unlike the Odds Ratio, each variable can have more than 2 categories

A chi-squared test statistic in a test of a contingency table that is equal to zero means:

A. The two nominal variables have values consistence with independence.
B. The two nominal variables have values that are consistent with equality.
C. The two nominal variables have the same proportions listed in Ho.
D. All of these choices.

χ^2 Contingency Test:

• Tests goodness-of-fit to the data of the null hypothesis of independence of variables

• Two categorical variables but, unlike the Odds Ratio, each variable can have more than 2 categories

When is it appropriate to use Chi-Squared tests?

----

a. When you are determining if two categorical variables are associated.

b. When you are directly comparing proportions

c. When your number of independent data points is less than 5

d. When you are looking for an exact P value.

<u>Example:</u> *Is there a relationship between age at first birth and the development of breast cancer?*

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	**3220**
No Cancer	1422	7325	1092	406	**10245**
Column Total	**1742**	**9542**	**1555**	**626**	**13465**

<u>STEP 1: Formulate null hypothesis</u>

Example: *Is there a relationship between age at first birth and the development of breast cancer?*

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	3220
No Cancer	1422	7325	1092	406	10245
Column Total	1742	9542	1555	626	13465

Step 1:

H_0: The development of breast cancer is ***independent*** of the age at first birth

H_A: The development of breast cancer is ***dependent*** of the age at first birth

Step 2: Identify the test statistic

χ^2 expectation under independence

With independence,

P[Age at first birth AND breast cancer] = ?

Example: *Is there a relationship between age at first birth and the development of breast cancer?*

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	3220
No Cancer	1422	7325	1092	406	10245
Column Total	1742	9542	1555	626	13465

Step 1:

H_0: The development of breast cancer is ***independent*** of the age at first birth

H_A: The development of breast cancer is ***dependent*** of the age at first birth

Step 2: Identify the test statistic

χ^2 expectation under independence

With independence,

P[Particular Age at first birth AND breast cancer] = P[Particulat Age at first birth]P[Breast cancer]

Calculating the expectations under H_0:

	<20	20-29	30-34	>=35	Row total
Cancer	320	2217	463	220	3220
No Cancer	1422	7325	1092	406	10245
Column Total	1742	9542	1555	626	13465

$$P[Age < 20 Birth] = \frac{1742}{13465} = 0.13$$

$$P[Cancer] = \frac{3220}{13465} = 0.24$$

$$P[NoCancer] = \frac{10245}{13465} = 0.76$$

If H_0 is true, then:

P[< 20 Age at first birth AND breast cancer] = 0.13*0.24 = 0.031

Calculating the expected **COUNTS** under H_0:

EXPECTED values Under Ho	<20	20-29	30-34	>=35	Row total
Cancer	**416.6** 320	**2281.9** 2217	**371.9** 463	**149.7** 220	3220
No Cancer	**1325.6** 1422	**7260.2** 7325	**1183.2** 1092	**477** 406	10245
Column Total	1742	9542	1555	626	13465

χ^2 Contingency Test

$$\chi^2 = \sum_i \frac{(Observed_i - Expected_i)^2}{Expected_i} = 104.76$$

$$= \frac{(416.6-320)^2}{416.6} + \frac{(2281.9-2217)^2}{2281.9} + \frac{(371.9-463)^2}{371.9} + \frac{(149.7-220)^2}{149.7} + \frac{(1325.6-1422)^2}{1325.6} + \frac{(7260.2-7325)^2}{7260.2} + \frac{(1183.2-1092)^2}{1183.2} + \frac{(477-406)^2}{477}$$

Degrees of Freedom:

dof = (row - 1)(column - 1)

For the Birth age/cancer example,

dof = (2-1)(4-1)=3

What would a chi-square contingency test resulting in a significance value of P > 0.05 suggest?

A. We cannot reject the hypothesis of independence between the two variables
B. We cannot reject the hypothesis of dependency between the two variables
C. There is a significant relationship between the two variables
D. We can reject the hypothesis of dependency between the two variables

Conclusion:

$$\chi^2 = 104.76 \quad >> \quad \chi_3^2 = 7.81$$

We can reject the null hypothesis of independence with a significance level of at least 0.05 and say that the age of first birth Was not independent on whether or not breast cancer eventually Developed.

Assumptions:

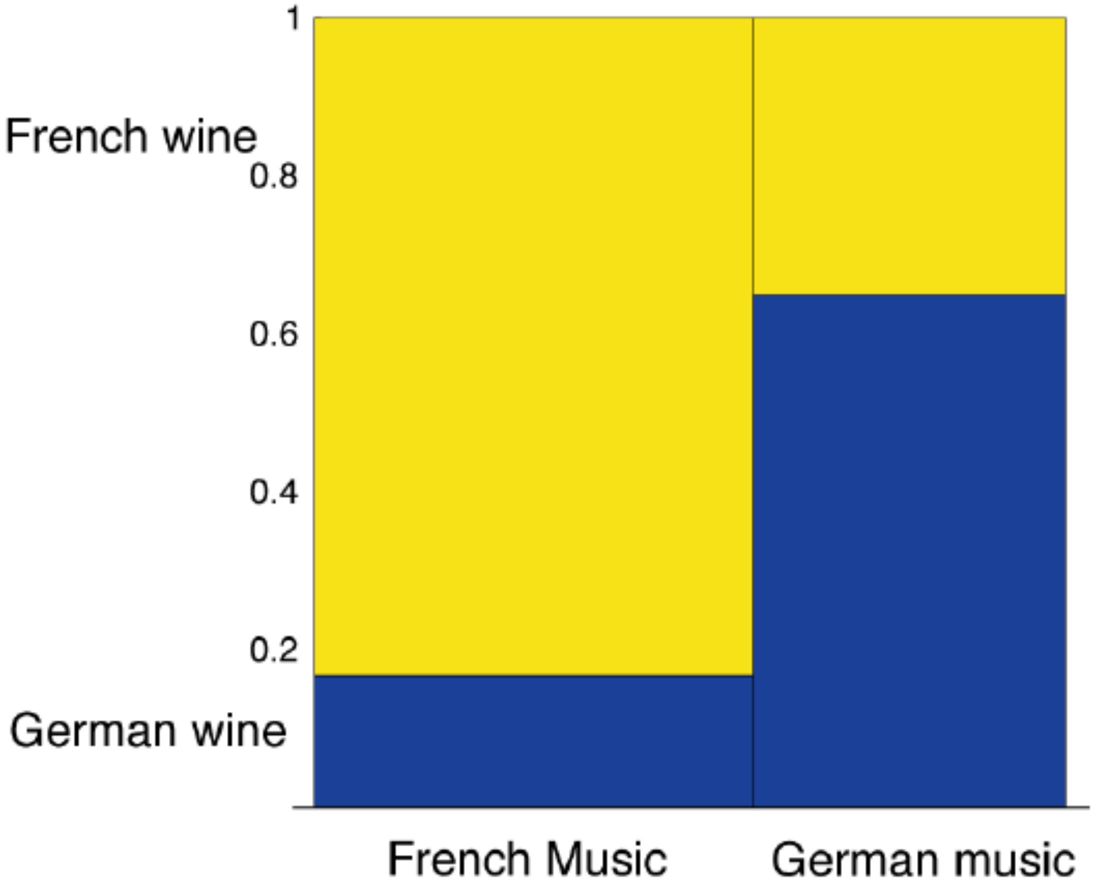
- The χ^2 test is a special case of the χ^2 goodness-of-fit test and so it has the same assumptions

- You can't have any **expectation** < 1 and no more than 20% of the **expected** categories < 5

Example: Does the nationality of background music effect the nationality of wine that is bought?

Observed	French Music	German Music	Row Totals
Bottles of French Wine	40	12	**52**
Bottles of German Wine	8	22	**30**
Column Totals	**48**	**34**	**82**

<u>Example:</u> Does the nationality of background music effect the nationality of wine that is bought?



Example: Does the nationality of background music effect the nationality of wine that is bought?

1. Hypotheses:

   H_0: The nationality of the purchased bottle of wine is *independent* of the nationality of the music played when it is sold

   H_A: The nationality of the bottle of wine sold *depends* on the nationality of the music played when it is sold

Example: Does the nationality of background music effect the nationality of wine that is bought?

1. Hypotheses:

   H_0: The nationality of the purchased bottle of wine is *independent* of the nationality of the music played when it is sold

   H_A: The nationality of the bottle of wine sold *depends* on the nationality of the music played when it is sold

2. Test statistic: χ^2 expectation under independence

   With independence,

   P[French wine AND French music] = ?

Example: Does the nationality of background music effect the nationality of wine that is bought?

1. Hypotheses:

   H_0: The nationality of the purchased bottle of wine is *independent* of the nationality of the music played when it is sold

   H_A: The nationality of the bottle of wine sold *depends* on the nationality of the music played when it is sold

2. Test statistic: χ^2 expectation under independence

With independence,

P[French wine AND French music] =

P[French Wine]xP[French Music]

Calculating the expectations under H_0:

Obs'd	French Music	German Music	Row Totals
Bottles of French Wine	40	12	52
Bottles of German Wine	8	22	30
Column Totals	48	34	82

$$P[FrenchWine] = \frac{52}{82} = 0.634$$

$$P[FrenchMusic] = \frac{48}{82} = 0.585$$

If H_0 is true, then:

P[Fr Music and Fr Wine]
= (0.634)(0.585)
= 0.37112

Calculating the expectations under H_0:

Expected	French Music	German Music	Row Totals
Bottles of French Wine	0.37(82) = 30.4 40	21.6 12	52
Bottles of German Wine	17.6 8	12.4 22	30
Column Totals	48	34	82

χ^2 Contingency Test

$$\chi^2 = \sum_i \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

$$= \frac{(40 - 30.4)^2}{30.4} + \frac{(12 - 21.6)^2}{21.6} + \frac{(8 - 17.6)^2}{17.6} + \frac{(22 - 12.4)^2}{12.4}$$

= 20

Degrees of Freedom:

dof = (row - 1)(column - 1)

For the music/wine nationality example,

dof = (2-1)(2-1)=1

Conclusion:

$$\chi^2 = 20 \;\gg\; \chi^2_{1,\alpha=0.05} = 3.85$$

We can reject the null hypothesis of independence with a significance level of at least 0.05 and say that the nationality of the wine sold was not independent on what music was played.

Assumptions:

- The χ^2 test is a special case of the χ^2 goodness-of-fit test and so it has the same assumptions

- You can't have any expectation < 1 and no more than 20% of the categories < 5