3. **Test Statistic:**

$$\chi_0^2 = \overbrace{\frac{(17-12.464)^2}{12.464} + \frac{(8-11.248)^2}{11.248} + \cdots + \frac{(13-11.856)^2}{11.856}}^{\text{include all 12}}$$

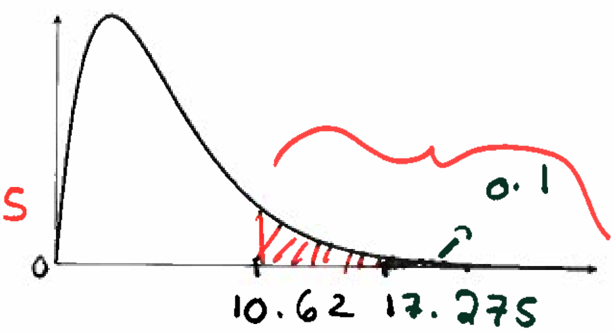
$$= 10.62$$

4. ***P*-value:** use $df = 12 - 1 = 11$

By χ^2-table

P-value

$= P(\chi^2 > 10.62) > 0.1 > 0.05$



using StatCrunch: P-value $= 0.4753 > 0.05$

Chi-Square goodness-of-fit results:

Observed: Observed Birth Month
Expected: Expected Birth Month

N	DF	Chi-Square	P-value
152	11	10.624087	0.4753

5. **Conclusion:** Since P-value $> \alpha = 0.05$, we do not reject Ho at the 0.05 significance level, that is, there is not enough statistical evidence to conclude the data doesn't match the statistics Canada model.

Section 22.4: Chi-Squared Test of Independence

Situation: We have **two categorical variables** measured on the **same population**. We want to know if there is an association between the two variables, that is, we want to test whether the two categorical variables are independent.

relationship

We use a two-way frequency table called a **contingency table** to display the **counts** of the variables. Contingency tables categorize the counts, to help us see whether the distribution of counts of one variable is contingent on the other.

In this case, the number of **degrees of freedom** is

$$df = (\#\,\text{Rows} - 1)(\#\,\text{Columns} - 1)$$

not including "total" rows/columns

The null hypotheses is that the variables are independent. Assuming the null hypothesis is true, the expected value for each cell in the table is

$$\text{Expected Cell Count} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

Note: The **Expected Cell Frequency Condition** is still required, that is, the expected count in each cell must be at least 5. If any of the expected cell counts are less than 5, categories (rows or columns) may be combined (in a sensible way) to create acceptable expected cell counts. If you combine categories, compute the *df* using the reduced number of categories.

Note: If we reject the null hypothesis and conclude that the variables are not independent, this is does **not** show a cause-and-effect relationship. There may be lurking variables. We say there is evidence of an association.

A chi-squared (hypothesis) test for independence has five steps:

1. **Assumptions/Conditions:**
   - Data must be in **counts** and **not** in proportions or percentages.
   - Responses of individuals should be independent of each other.
   - Sample must be chosen randomly (to generalize to the population).
   - Sample size must be large enough: each cell should have an **expected count** of at least 5.

2. **Hypotheses:**

$$H_0 : \text{The variables are independent.}$$
$$H_A : \text{The variables are not independent.}$$

3. **Test Statistic:**

$$\chi_0^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

   where

$$\text{Expected} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

4. ***P*-value:** The P-value $= P(\chi^2 > \chi_0^2)$ can be computed using software or using the χ^2-table with $df = (\#\,\text{Rows} - 1)(\#\,\text{Columns} - 1)$.

5. **Conclusion:** Report and interpret the P-value in context. Given a significance level α,
   - if P - value $\leq \alpha$, we reject H_0 at level α
   - if P- value $> \alpha$, we do not reject H_0 at level α

Example: A national survey was conducted to obtain information on alcohol consumption and marital status. A random sample of 1772 adults provided the following information:

Drinks per month

observed

	Abstain	1-60	Over 60	Total
Single	67	213	74	354
Married	411	633	129	1173
Widowed	85	51	7	143
Divorced	27	60	15	102
Total	590	957	225	1772

Do the data provide evidence at the 5% significance level of an association between alcohol consumption and marital status?

Drinks per month

Expected

	Abstain	1-60	Over 60	Total
Single	117.87	191.18	44.95	354
Married	390.56	633.5	148.94	1173
Widowed	47.61	77.23	18.16	143
Divorced	33.96	55.09	12.95	102
Total	590	957	225	1772

$\frac{354(590)}{1772} = 117.87$ $\quad \frac{1173(225)}{1772} = 148.94$ $\quad \frac{143(957)}{1772} = 77.23$ $\quad \frac{102(590)}{1772} = 33.96$

1. **Assumptions/Conditions:**
   - independent, random sample
   - data in counts
   - expected counts ≥ 5

2. **Hypotheses:**

H_0: The variables "alcohol consumption" and marital status are independent.

H_A: The variables are not independent.

3. **Test Statistic:**

$$\chi_0^2 = \frac{(67-117.87)^2}{117.87} + \frac{(213-191.8)^2}{191.8} + \cdots + \frac{(15-12.95)^2}{12.95}$$

$$= 94.269$$

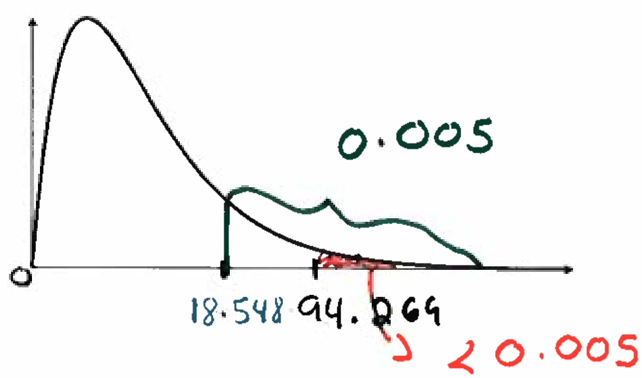
4. ***P*-value:** df: $(4-1)(3-1) = 6$

By the χ^2-table:

P-value

$= P(\chi^2 > 94.269) < 0.0005$

≤ 0.05



Using Statcrunch,

p-value < 0.0001

≤ 0.05

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	6	94.268801	<0.0001

5. **Conclusion:** Since p-value $\leq \alpha = 0.05$, we reject H_0 at the 0.05 significance level, that is, there is enough statistical evidence to conclude that "alcohol consumption" and "marital status" are not independent.

(there is an association)

Section 22.2: Chi-Squared Test of Homogeneity

Situation: We have **one categorical variables** measured on **two or more groups or populations**. We want to compare the distributions of the variable between the groups.

The word "homogeneity" means the quality or state of being all the same or all of the same kind. If the groups have the same distribution for a variable, they are **homogeneous** with respect to that variable; otherwise they are **nonhomogeneous** with respect to the variable.

do not change from group to group.

The null hypothesis is that the category proportions / distributions are the same for all groups. (The groups/populations are homogeneous with respect to the variable.)

The alternative hypothesis is that the distributions are not all the same. (The groups/populations are nonhomogenous with respect to the variable.) The alternative hypothesis means that the distributions differ for at least two of the groups.

Note: A test for homogeneity is the generalization of the two-tailed z-test for two proportions.

for 2x2 table, tests are equivalent $z_0^2 = \chi_0^2$, Same p-value.

Note: The mechanics of the test of homogeneity are the same as the test for independence.

Note: The expected count in each cell must be at least 5. If any of the expected cell counts are less than 5, categories (rows or columns) may be combined (in a sensible way) to create acceptable expected cell counts. If you combine categories, compute the df using the reduced number of categories.

A chi-squared (hypothesis) test for homogeneity has five steps:

1. **Assumptions/Conditions:**
   - Data must be in **counts** and **not** in proportions or percentages.
   - Responses of individuals should be independent of each other. The groups should be independent of each other.
   - Sample must be chosen randomly.
   - Sample size must be large enough: each cell should have an **expected count** of at least 5.

2. **Hypotheses:**

   category proportions

   H_0 : The distributions are the same for all groups.

   H_A : The distributions are not the same.

3. **Test Statistic:**

$$\chi_0^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

   where

$$\text{Expected} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

4. ***P*-value:** The P-value $= P(\chi^2 > \chi_0^2)$ can be computed using software or using the χ^2-table with $df = (\#\,\text{Rows} - 1)(\#\,\text{Columns} - 1)$.

5. **Conclusion:** Report and interpret the P-value in context. Given a significance level α,
   - if P - value $\leq \alpha$, we reject H_0 at level α
   - if P- value $> \alpha$, we do not reject H_0 at level α

Example: A random survey of vehicles parked in the student lot and in the staff lot at a large university classified the brands by manufacturing region.

Observed

	North American	European	Asian	Total
Student	107	33	55	195
Staff	105	12	47	164
Total	212	45	102	359

Are there differences in the distribution of vehicle origin for staff and students? Use $\alpha = 0.05$.

Expected

	North American	European	Asian	Total
Student	115.15	24.44	55.4	195
Staff	96.85	20.56	46.6	164
Total	212	45	102	359

$$\frac{(195)(45)}{359} = 24.44 \qquad \frac{(164)(212)}{359} = 96.85 \qquad \frac{(164)(102)}{359} = 46.6$$

1. **Assumptions/Conditions:**
   - data in counts
   - independent responses/groups
   - random sample
   - expected counts ≥ 5
2. **Hypotheses:**

H_0: The distribution of vehicle origin is the same for staff and students

H_A: The distribution of vehicle origin is different for staff and students.

3. **Test Statistic:**

$$\chi_0^2 = \frac{(107 - 115.15)^2}{115.15} + \frac{(33 - 24.44)^2}{24.44} + \dots + \frac{(47 - 46.6)^2}{46.6}$$

$$= 7.828$$

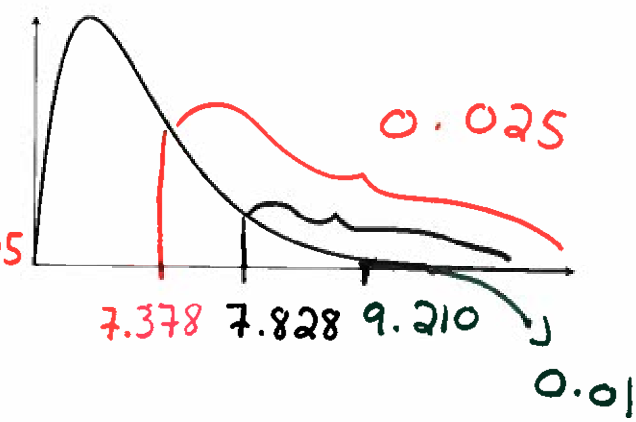
4. ***P*-value:** df: $(2-1)(3-1) = 2$

By the χ^2- table,

$0.01 <$ P-value < 0.025 ≤ 0.05

$\quad\quad\quad \parallel$

$P(\chi^2 > 7.828)$



using StatCrunch,

P-value $= 0.02 \leq 0.05$

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	2	7.8278067	0.02

5. **Conclusion:** Since P-value $\leq \alpha = 0.05$, we reject H_0 at the 0.05 significance level, that is, there is enough statistical evidence to conclude that the distribution of vehicle origin differs between staff and students.

Example: To determine the effects of a chemical treatment on the rate of seed germination, 100 chemically treated seeds and 150 untreated seeds are planted. The number of seeds that germinated are recorded below:

	Germinated	Not Germinated	Total
Treated	84 (86.4)	16 (13.6)	100
Untreated	132(129.6)	18 (20.4)	150
Total	216	34	250

a) Conduct a chi-squared test to determine if the rate of germination is different for the treated and untreated seeds. Use $\alpha = 0.05$.

1. **Assumptions/Conditions:**
   - data in counts
   - independent results / groups
   - random sample ??
   - expected counts ≥ 5
2. **Hypotheses:**

   H_0: The proportions in each category are the same for treated and untreated seeds.

   H_A: The proportions are not the same.
3. **Test Statistic:**

$$\chi_0^2 = \frac{(84-86.4)^2}{86.4} + \frac{(16-13.6)^2}{13.6} + \frac{(132-129.6)^2}{129.6}$$

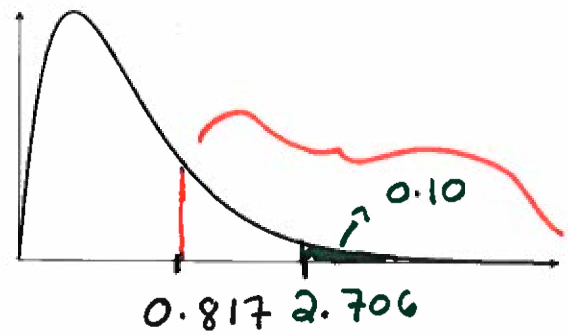
$$+ \frac{(18-20.4)^2}{20.4} = 0.81699$$

4. ***P*-value:** $df = (2-1)(2-1) = 1$

By the χ^2- table,

P-value $= P(\chi^2 > 0.817) > 0.1$

> 0.05



Using StatCrunch, **Chi-Square test:**

P-value $= 0.3661$

> 0.05

Statistic	DF	Value	P-value
Chi-square	1	0.81699346	0.3661

5. **Conclusion:** Since P-value $> \alpha = 0.05$, we do not reject H_0 at the 0.05 significance level, that is, there is not enough statistical evidence to conclude that the rate of germination is different between the treated and untreated seeds.

b) Conduct a two-proportion z-test to determine if the rate of germination is different for the treated and untreated seeds. Use $\alpha = 0.05$.

p_1 = proportion of treated seeds that germinate

p_2 = proportion of untreated seeds that germinate

2. **Hypotheses:**

$H_0: p_1 - p_2 = 0 \quad (p_1 = p_2)$

$H_A: p_1 - p_2 \neq 0 \quad$ (two-tailed test)