Mathematics 13 : Elementary Statistics

Unit 4 : Section 3

Lesson 43 :

The Chi-Squared Test for Independence

The Chi-Squared Test for Independence

One specific application of the Chi-Squared Goodness-of-Fit Test involves testing the assumption that two events are independent of one another. That is, to determine if the categories of two qualitative variables have no affect on each other.

This adaptation of the Chi-Squared Goodness-of-Fit Test is referred to as the Chi-Squared Test for Independence.

The Chi-Squared Test for Independence

The expected frequencies used in the Chi-Squared Test for Independence are generated under the assumption that the two events are independent of one other.

Since two events E and F are considered to be independent when $p_{E \text{ given } F} = p_E$, the predicted proportions ($p_{\cdot i}$) of each category in this test are the probabilities of the corresponding categories of event E (p_E).

The Chi-Squared Test for Independence

The Chi-Squared
Test for Independence

is used to determine whether two events
are independent of one another

State hypothesis
  H_0: The events are independent
  H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into
an rxc matrix

X^2-Test
with df = (c - 1)·(r - 1)

Decision:

  Reject H_0 when p-value $\leq \alpha$

  Otherwise do not reject H_0

State conclusion

The stated hypotheses (H_0 and H_1) in the Chi-Squared Test for Independence are always expressed in the same manner.

The Chi-Squared Test for Independence

The Chi-Squared
Test for Independence

is used to determine whether two events
are independent of one another

State hypothesis
  H_0: The events are independent
  H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into
an rxc matrix

X^2-Test
with df = (c - 1)·(r - 1)

Decision:

  Reject H_0 when p-value $\leq \alpha$

  Otherwise do not reject H_0

State conclusion

Unless there exists a compelling reason to do so otherwise, hypothesis testing procedures are conducted using the customary 5% level of significance.

The Chi-Squared Test for Independence

The Chi-Squared Test for Independence

is used to determine whether two events are independent of one another

State hypothesis
H_0: The events are independent
H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into an rxc matrix

X^2-Test
with df = $(c - 1) \cdot (r - 1)$

Decision:

Reject H_0 when p-value $\leq \alpha$

Otherwise do not reject H_0

State conclusion

		Qualitative Variable 2			
		Column 1	Column 2	...	Column c
Qualitative Variable 1	Row 1	$O_{1\,1}$	$O_{1\,2}$	...	$O_{1\,c}$
	Row 2	$O_{2\,1}$	$O_{2\,2}$	...	$O_{2\,c}$
	⋮	⋮	⋮	$O_{j\,i}$	⋮
	Row r	$O_{r\,1}$	$O_{r\,2}$	...	$O_{r\,c}$

Data organized in this manner can be entered into the TI-84 calculator as an rxc matrix.

The Chi-Squared Test for Independence

The Chi-Squared Test for Independence

is used to determine whether two events are independent of one another

State hypothesis
H_0: The events are independent
H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into an rxc matrix

X^2-Test
with df = $(c - 1) \cdot (r - 1)$

Decision:

Reject H_0 when p-value $\leq \alpha$

Otherwise do not reject H_0

State conclusion

After the observed frequencies are entered into an rxc matrix, the X^2-Test command on the TI-84 calculator will calculate the desired p-value for the hypothesis test procedure.

The Chi-Squared Test for Independence

The Chi-Squared Test for Independence

is used to determine whether two events are independent of one another

State hypothesis
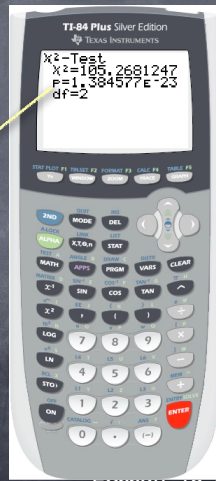H_0: The events are independent
H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into an rxc matrix

X^2-Test
with df = $(c - 1) \cdot (r - 1)$

Decision:

Reject H_0 when p-value $\leq \alpha$

Otherwise do not reject H_0

State conclusion


```
X²-Test
X²=105.2681247
P=1.384577ᴇ⁻23
df=2
```

The Chi-Squared Test for Independence

The Chi-Squared Test for Independence

is used to determine whether two events are independent of one another

State hypothesis
H_0: The events are independent
H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into an rxc matrix

X^2-Test
with df = $(c - 1) \cdot (r - 1)$

Decision:

Reject H_0 when p-value $\leq \alpha$

Otherwise do not reject H_0

State conclusion

The test statistic for this hypothesis testing procedure adapts the Chi-Squared Goodness-of-Fit test statistic to accommodate the categories of two qualitative variables.

$$X^2 = \sum_{j=1}^{r} \sum_{i=1}^{c} \frac{\left(O_{ji} - E_{ji}\right)^2}{E_{ji}}$$

The expected frequencies are determined under the assumption of independence ($E_{ji} = n_{j\cdot} \cdot p_{\cdot i}$).

The Chi-Squared Test for Independence

The Chi-Squared Test for Independence

is used to determine whether two events are independent of one another

State hypothesis
H_0: The events are independent
H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into an rxc matrix

X^2-Test
with df = $(c - 1) \cdot (r - 1)$

Decision:

Reject H_0 when p-value $\leq \alpha$

Otherwise do not reject H_0

State conclusion

In order for the Chi-Squared Test for Independence to be applied, the independent random sample must be sufficiently large. This can be accomplished by ensuring that the sample size is large enough so that all (or nearly all) of the expected frequencies are 5 or more. Also, none of the expected frequencies should equal 0.

The Chi-Squared Test for Independence

The Chi-Squared Test for Independence

is used to determine whether two events are independent of one another

State hypothesis
H_0: The events are independent
H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into an rxc matrix

X^2-Test
with df = $(c - 1) \cdot (r - 1)$

Decision:

Reject H_0 when p-value $\leq \alpha$

Otherwise do not reject H_0

State conclusion

The calculated p-value is used to reach a decision regarding the validity of H_0.

Small p-values provide sample evidence contradicting H_0.

Large p-values provide sample evidence consistent with H_0.

The Chi-Squared Test for Independence

The Chi-Squared Test for Independence

is used to determine whether two events are independent of one another

State hypothesis
H_0: The events are independent
H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into an rxc matrix

X^2-Test
with $df = (c - 1)\cdot(r - 1)$

Decision:

Reject H_0 when p-value $\leq \alpha$

Otherwise do not reject H_0

State conclusion

When H_0 is not rejected, the conclusion is that the events are independent.

This indicates that there is no relationship or association between these qualitative variables. Thus, the variables do not affect each other.

The Chi-Squared Test for Independence

The Chi-Squared Test for Independence

is used to determine whether two events are independent of one another

State hypothesis
H_0: The events are independent
H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into an rxc matrix

X^2-Test
with $df = (c - 1)\cdot(r - 1)$

Decision:

Reject H_0 when p-value $\leq \alpha$

Otherwise do not reject H_0

State conclusion

When H_0 is rejected, the conclusion is that the events are not independent.

This indicates that there is a relationship or association between these qualitative variables. Thus, the variables do affect each other.

Example 1

Based on the results of the Sierra College Elementary Statistics Student Survey, can one conclude that a student's rating of their mathematics ability is made independent of their gender?

The Chi-Squared Test for Independence is used to determine whether two events are independent of one another.

Example 1

Q1. What is your gender?		Q3. How would you rate your ability as a mathematics student?			
		Below Average	Just Average	Above Average	Total
	Male	2	11	4	17
	Female	6	17	4	27
	Total	8	28	8	44

Example 1

H_0: The events are independent

H_1: The events are not independent

Use $\alpha = 0.05$

X^2-Test

p-value ≈ 0.586

Since the p-value of 0.586 is not 0.05 or less, the decision is to not reject H_0.

Example 1

Therefore, based on the results of the Sierra College Elementary Statistics Student Survey, one can conclude that a student's rating of their mathematics ability is made independent of their gender.

So, it appears that gender does not affect the way in which a Sierra College Elementary Statistics student would rate their ability as a mathematics student.

Example 1

That is, the students' responses to the question "How would you rate your ability as a mathematics student?" were essentially the same for both males and females.

Q1. What is your gender?	Q3. How would you rate your ability as a mathematics student?			
	Below Average	Just Average	Above Average	Total
Male	2	11	4	17
Female	6	17	4	27
Total	8	28	8	44

Example 2

The following contingency table was constructed using the results obtained in the Movie Database Sample by counting separately for each of the MPAA movie ratings the number of movies that received multiple (more than one) Oscar nominations and the number of movies that did not.

Example 2

Multiple Oscar Nominations?	MPAA Movie Rating			
	G	PG	PG-13	R
No	3	4	11	5
Yes	6	7	12	12

Example 2

Conduct a hypothesis testing procedure, at the 5% level of significance, to determine whether a movie receiving multiple Oscar nominations or not is independent of the movie's MPAA rating.

Based on this result, what can one conclude about the relationship between a movie's MPAA rating and receiving multiple Oscar nominations?

Example 2

The Chi-Squared Test for Independence

is used to determine whether two events are independent of one another

State hypothesis
H_0: The events are independent
H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into an r×c matrix

X^2-Test
with $df = (c - 1) \cdot (r - 1)$

Decision:
Reject H_0 when p-value $\leq \alpha$
Otherwise do not reject H_0

State conclusion

> The Chi-Squared Test for Independence is used to determine whether two events are independent of one another.

Example 2

H_0: The events are independent

H_1: The events are not independent

Use $\alpha = 0.05$

X^2-Test

   p-value ≈ 0.668

Since the p-value of 0.668 is not 0.05 or less, the decision is to not reject H_0.

Example 2

Therefore, at the 5% level of significance, a movie receiving multiple Oscar nominations or not is independent of the movie's MPAA rating.

Based on this result, one can conclude that there is no relationship between a movie's MPAA rating and receiving multiple Oscar nominations.

Example 2

That is, whether a movie is rated "G", "PG", "PG-13", or "R" makes no difference with respect to the movie receiving multiple Oscar nominations or not.

Example 3

The National Highway Traffic Safety Administration (NHTSA) sponsored a study investigating seat belt usage. As part of this study, a random sample of vehicles were selected on roadway sites nationwide. Information regarding the driver's age and whether they were wearing their seat belt or not was collected. The results are shown in the following contingency table.

Example 3

		Age of Driver		
		18 to 24	25 to 34	35 to 64
Wearing Seat Belt?	Yes	52	85	121
	No	19	20	18

Example 3

Is a driver's age independent of their seat belt usage?

Based on this result, what should the NHTSA report about the association between a driver's age and their seat belt usage?

Example 3

The Chi-Squared Test for Independence

is used to determine whether two events are independent of one another

State hypothesis
H_0: The events are independent
H_1: The events are not independent

Use $\alpha = 0.05$
(unless stated otherwise)

Enter the observed frequencies (data) into an r×c matrix

X^2-Test
with df = (c - 1)·(r - 1)

Decision:

Reject H_0 when p-value ≤ α

Otherwise do not reject H_0

State conclusion

The Chi-Squared Test for Independence is used to determine whether two events are independent of one another.

Example 3

H_0: The events are independent

H_1: The events are not independent

Use $\alpha = 0.05$

X^2-Test

p-value $\approx$ 0.046

Since the p-value of 0.046 is 0.05 or less, the decision is to reject H_0.

Example 3

Therefore, a driver's age is not independent of their seat belt usage.

Based on this result, the NHTSA should report that a higher percentage of seat belt usage is associated with older drivers (35 to 64 years) and a lower percentage of seat belt usage is associated with younger drivers (18 to 24 years).

Your solutions should be clear, complete, and sufficiently detailed in order to demonstrate your understanding and communicate your reasoning and method of solving the problem.

Exercise 1

A CBS News poll asked a random sample of 1025 Americans how they identified themselves politically and whether they thought the news media was too liberal. The results of this poll are summarized in the table below.

		Political Identification		
		Conservative	Moderate	Liberal
Is the news media too liberal?	Yes	232	165	41
	No	154	247	186

(a) Is a person's response to this question independent of their political identification? Use $\alpha = 0.05$.

(b) Based on this result, does a person's response to this question depend on how they identify themselves politically?

Exercise 2

In a recent Gallup poll, a random sample of 1014 adults nationwide were asked "How often do you worry about your weight?" and "How would you describe your own personal weight situation?". The poll results are presented in the table below.

		How often do you worry about your weight?			
		All of the time	Some of the time	Not too often	Never
How would you describe your own personal weight situation?	Overweight	102	198	91	25
	About right	49	126	213	159
	Underweight	6	3	11	31

(a) Is how often a person worries about their weight independent of how they describe their own personal weight situation?

(b) Based on this conclusion, does a person's personal weight situation affect how often they worry about their weight?

Exercise 3

A health club manager was interested in whether gender is related to the number of days per week that its members utilize the club's facilities. A random sample of 300 members of this health club was selected, and a report of these members' activity over the past week was generated using the health club's information management system. The following results were obtained.

		Number of Days per Week Member Utilizes Club's Facilities			
		0 - 1	2 - 3	4 - 5	6 - 7
Gender	Male	43	91	32	7
	Female	35	68	21	3

(a) Is there enough evidence, at the 5% level of significance, to conclude that the number of days per week a member utilizes the health club's facilities is independent of their gender?

(b) Based on this result, what should the health club manager conclude about the relationship between gender and the number of days per week that its members utilize the club's facilities?

Exercise 4

A music industry marketing firm asked a random sample of 150 consumers if they have purchased any music from an online music store in the last year. The results, grouped by age, are displayed in the following table.

		Consumer's Age		
		18 to 29	30 to 49	50 to 69
Have you purchased any music from an online music store in the last year?	Yes	37	12	5
	No	46	22	28

(a)   Is purchasing music online independent of the consumer's age? Use α = 5%.

(b)   Based on this result, what should the music industry marketing firm report about the association between purchasing music online and the consumer's age?

> Your solutions should be clear, complete, and sufficiently detailed in order to demonstrate your understanding and communicate your reasoning and method of solving the problem.

Exercise 1

A CBS News poll asked a random sample of 1025 Americans how they identified themselves politically and whether they thought the news media was too liberal. The results of this poll are summarized in the table below.

		Political Identification		
		Conservative	Moderate	Liberal
Is the news media too liberal?	Yes	232	165	41
	No	154	247	186

(a) Is a person's response to this question independent of their political identification? Use $\alpha = 0.05$.

H_0 : The events are independent.

H_1 : The events are not independent.

Use $\alpha = 0.05$

```
MATRIX[A]  2 ×3
[ 232      165      41     ]
[ 154      247      186    ]
```

```
χ²-Test
 Observed: [A]
 Expected: [B]
 Calculate  Draw
```

```
χ²-Test
 χ²=105.2681247
 P=1.384577E-23
 df=2
```

p-value $\approx 1.4E - 23 \approx 0.000$

Since the p-value of 0.000 is 0.05 or less, the decision is to reject H_0.

Therefore, a person's response to this question is not independent of their political identification.

(b) Based on this result, does a person's response to this question depend on how they identify themselves politically?

Based on this result, a person's response to this question does depend on how they identify themselves politically.

Exercise 2

In a recent Gallup poll, a random sample of 1014 adults nationwide were asked "How often do you worry about your weight?" and "How would you describe your own personal weight situation?". The poll results are presented in the table below.

		How often do you worry about your weight?			
		All of the time	Some of the time	Not too often	Never
How would you describe your own personal weight situation?	Overweight	102	198	91	25
	About right	49	126	213	159
	Underweight	6	3	11	31

(a) Is how often a person worries about their weight independent of how they describe their own personal weight situation?

H_0 : The events are independent.

H_1 : The events are not independent.

Use $\alpha = 0.05$

```
MATRIX[A]  3 ×4
[ 102      198      91     -
[ 49       126      213    -
[ 6        3        11     -
```

```
X²-Test
 Observed: [A]
 Expected: [B]
 Calculate  Draw
```

```
X²-Test
 X²=215.8538717
 P=7.965343E-44
 df=6
```

p-value $\approx 8.0E - 44 \approx 0.000$

Since the p-value of 0.000 is 0.05 or less, the decision is to reject H_0.

Therefore, how often a person worries about their weight is not independent of how they describe their own personal weight situation.

(b) Based on this conclusion, does a person's personal weight situation affect how often they worry about their weight?

Based on this conclusion, a person's personal weight situation does affect how often they worry about their weight. The data indicate that an overweight person is more likely to worry about their weight either some or all of the time while an underweight person is more likely to worry about their weight either never or not too of often.

Exercise 3

A health club manager was interested in whether gender is related to the number of days per week that its members utilize the club's facilities. A random sample of 300 members of this health club was selected, and a report of these members' activity over the past week was generated using the health club's information management system. The following results were obtained.

		Number of Days per Week Member Utilizes Club's Facilities			
		0 - 1	2 - 3	4 - 5	6 - 7
Gender	Male	43	91	32	7
	Female	35	68	21	3

(a) Is there enough evidence, at the 5% level of significance, to conclude that the number of days per week a member utilizes the health club's facilities is independent of their gender?

H_0 : The events are independent.

H_1 : The events are not independent.

Use $\alpha = 5\%$

```
MATRIX[A] 2 ×4
[ 43      91      32    _
[ 35      68      21    _
```

```
X²-Test
 Observed: [A]
 Expected: [B]
 Calculate Draw
```

```
X²-Test
 X²=1.000771633
 P=.8010652443
 df=3
```

p-value ≈ 0.801

Since the p-value of 0.801 is not 0.05 or less, the decision is to not reject H₀.

Therefore, there is enough evidence, at the 5% level of significance, to conclude that the number of days per week a member utilizes the health club's facilities is independent of their gender.

(b) Based on this result, what should the health club manager conclude about the relationship between gender and the number of days per week that its members utilize the club's facilities?

Based on this result, the health club manager should conclude that there is no relationship between gender and the number of days per week that its members utilize the club's facilities. In other words, there are no gender differences in how frequently the members utilize the club's facilities.

Exercise 4

A music industry marketing firm asked a random sample of 150 consumers if they have purchased any music from an online music store in the last year. The results, grouped by age, are displayed in the following table.

		Consumer's Age		
		18 to 29	30 to 49	50 to 69
Have you purchased any music from an online music store in the last year?	Yes	37	12	5
	No	46	22	28

(a) Is purchasing music online independent of the consumer's age? Use $\alpha = 5\%$.

H_0 : The events are independent.

H_1 : The events are not independent.

Use $\alpha = 5\%$

```
MATRIX[A]  2 ×3
[ 37      12      5    ]
[ 46      22      28   ]
```

```
X²-Test
 Observed: [A]
 Expected: [B]
 Calculate Draw
```

```
X²-Test
 X²=8.883879248
 P=.011773081
 df=2
```

p-value ≈ 0.012

Since the p-value of 0.012 is 0.05 or less, the decision is to reject H₀.

Therefore, purchasing music online is not independent of the consumer's age.

(b) Based on this result, what should the music industry marketing firm report about the association between purchasing music online and the consumer's age?

Based on this result, the music industry marketing firm should report that there is an association between purchasing music online and the consumer's age. The data indicate that consumers who purchase music online tend to be younger (18 to 29) rather than older (50 to 69).