Math 1710 Class 36

Chi Square
Dr. Back

Nov. 20, 2009
Sales figures for two car brands by district.
MP as more Sensitive

2-Sample t-Test of $\mu_1 - \mu_2$

No Selector

Individual Alpha Level 0.05

$H_0: \mu_1 - \mu_2 = 0$  $H_a: \mu_1 - \mu_2 \neq 0$

Hide Results

SalesGL - SalesBW:

Test $H_0: \mu(SalesGL) - \mu(SalesBW) = 0$ vs $H_a: \mu(SalesGL) - \mu(SalesBW) \neq 0$

Difference Between Means = 13.500000  t-Statistic = 0.3

Fail to reject $H_0$ at Alpha = 0.05

$p = 0.7058$
MP as more Sensitive

Paired t-Test of \( \mu(1 - 2) \)

- **No Selector**
- **Individual Alpha Level 0.05**
- **Ho: \( \mu(1 - 2) = 0 \)** Ha: \( \mu(1 - 2) \neq 0 \)

**Hide Results**

SalesGL - SalesBW:
Test Ho: \( \mu(SalesGL-SalesBW) = 0 \) vs Ha: \( \mu(SalesGL-SalesBW) \neq 0 \)
Mean of Paired Differences = 13.500000 t-Statistic = 2.257 w/19 df
Reject Ho at Alpha = 0.05
\( p = 0.0360 \)
MP as more Sensitive

Taste Test Results on 2 teas:
Each of 16 people tasted both teas and gave them a rating from 1 to 7, 1 being best.
MP as more Sensitive

Why $(\frac{Obs - Exp}{Exp})^2$?

2 by 2 Tables

Degrees of Freedom

Chi Square Hypotheses

Goodness of Fit Example

Using Table chi

Chi Square Distribution as an RV

Chi Square

Robustness

t-Conditions

Mendel

Regression

Inference

Questions
MP as more Sensitive

Math 1710
Class 36
V3

Chi Square
Using Table chi
Chi Square
Distribution as an RV
Chi Square
Hypotheses
Goodness of Fit Example

Why \( \frac{(Obs - Exp)^2}{Exp} \)?

2 by 2 Tables

Degrees of Freedom

Mendel
Regression
Inference
Questions
MP as more Sensitive

t-Conditions
Robustness
Chi Square
Using Table
Chi Square Distribution as an RV
Chi Square Hypotheses
Goodness of Fit Example
Why \( \frac{(Obs - Exp)^2}{Exp} \)?
2 by 2 Tables
Degrees of
MP as more Sensitive

Paired t-Interval for \( \mu(1 - 2) \)

- No Selector
- Individual Confidence: 95.00%
- Bounds: Lower Bound > \( \mu(1 - 2) \) < Upper Bound

With 95.00% Confidence, 0.10507624 < \( \mu(\text{Taste 2 - Taste 1}) \) < 1.5199238

Why \( \frac{(Obs - Exp)^2}{Exp} \)?

2 by 2 Tables

Degrees of Freedom

Chi Square

Distribution as an RV

Chi Square Hypotheses

Goodness of Fit Example

Why \( \frac{(Obs - Exp)^2}{Exp} \)?

2 by 2 Tables

Degrees of Freedom

Chi Square

Distribution as an RV

Chi Square Hypotheses

Goodness of Fit Example
Conditions for t Tests

- plausible independence
- random sampling
- 10% condition
- nearly normal condition
  
  \[ n < 15 \text{ unimodal and symmetric} \]
  
  \[ 15 \leq n \leq 40 \text{ avoid strong skewness and outliers} \]
  
  \( \text{ (unimodal and sym best)} \)
  
  \[ n > 40 \text{ pretty much ok} \]
Conditions for t Tests

- plausible independence
- random sampling
- 10% condition
- nearly normal condition

\[ n < 15 \text{ unimodal and symmetric} \]
\[ 15 \leq n \leq 40 \text{ avoid strong skewness and outliers} \]
\[ n > 40 \text{ pretty much ok} \]

What Happens if Not Satisfied:

random sampling - could be critical;
might be ok if "representative"
representative hard/impossible to define
Conditions for t Tests

- plausible independence
- random sampling
- 10% condition
- nearly normal condition
  
  \[
  n < 15 \quad \text{unimodal and symmetric}
  \]
  
  \[
  15 \leq n \leq 40 \quad \text{avoid strong skewness and outliers}
  \]
  
  \[
  (\text{unimodal and sym best})
  \]
  
  \[
  n > 40 \quad \text{pretty much ok}
  \]

What Happens if Not Satisfied:

plausible independence -
could be critical
sometimes just a working hypothesis
Conditions for t Tests

- plausible independence
- random sampling
- 10% condition
- nearly normal condition

\[ n < 15 \text{ unimodal and symmetric} \]
\[ 15 \leq n \leq 40 \text{ avoid strong skewness and outliers} \]

nearly normal condition

\[ n > 40 \text{ pretty much ok} \]

What Happens if Not Satisfied:

10% condition -
results in overestimation of samp. dist. st dev

gradual breakdown in formulas, not method
Conditions for t Tests

- plausible independence
- random sampling
- 10% condition
- nearly normal condition

\[ n < 15 \] unimodal and symmetric
\[ 15 \leq n \leq 40 \] avoid strong skewness and outliers
  (unimodal and sym best)
\[ n > 40 \] pretty much ok

What Happens if Not Satisfied:

- nearly normal - no guarantee
- progressive reduction of accuracy
For 2-sample inference, we add the

\textit{independence groups assumption}.

The chance of an individual in one of the groups assuming a certain value should be independent of the values assumed by any of the individuals in the other group.
Robustness

Formulas for t-inference and regression inference are based on assumptions of normality of the data. Yet most distributions are not normal. (Although the CLT makes averages of lots normal.) So it is perhaps remarkable that t-inference methods for moderate size data sets without outliers are typically pretty good. Why is this?
Robustness

Formulas for t-inference and regression inference are based on assumptions of normality of the data. Yet most distributions are not normal. (Although the CLT makes averages of lots normal.) So it is perhaps remarkable that t-inference methods for moderate size data sets without outliers are typically pretty good. Why is this? One answer is robustness against non-normality.
Robustness

Actual confidence level vs kurtosis

for some symmetric $n = 25$ test distributions in a 1975 Biometrika paper of Pearson and Please. (kurtosis=3 in the normal case.)
Robustness

Actual confidence level vs kurtosis

for some skewed $n = 25$ test distributions in a 1975 Biometrika paper of Pearson and Please.
Chi Square Test of Independence

Flu Shots (categorical) vs Getting the Flu

<table>
<thead>
<tr>
<th></th>
<th>0-shot</th>
<th>1-shot</th>
<th>2-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>no flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
</tr>
</tbody>
</table>
Chi Square Test of Independence

Hypotheses:

- **$H_0$**: The categorical variables FluShot? and GetTheFlu? are independent.
- **$H_a$**: The categorical variables are not independent. *(i.e. there is an association.)*
### Chi Square Test of Independence

<table>
<thead>
<tr>
<th></th>
<th>0-shot</th>
<th>1-shot</th>
<th>2-shot</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>flu</strong></td>
<td>24</td>
<td>9</td>
<td>13</td>
<td>46</td>
</tr>
<tr>
<td><strong>no flu</strong></td>
<td>289</td>
<td>100</td>
<td>565</td>
<td>954</td>
</tr>
<tr>
<td></td>
<td>313</td>
<td>109</td>
<td>578</td>
<td>1000</td>
</tr>
</tbody>
</table>

With Marginals

**Why** $(\frac{(Obs - Exp)^2}{Exp})$?

**2 by 2 Tables**

**Degrees of Freedom**

**Chi Square Hypotheses**

**Goodness of Fit Example**

**Robustness**

**t-Conditions**

**Sensitive**

**MP as more**

**Math 1710 Class 36**

**V3**

**313**

**Chi Square**

**Using Table chi**

**Chi Square Distribution as an RV**

**Chi Square**

**Hypotheses**

**With Marginals**

**Goodness of Fit Example**

**Chi Square**

**Hypotheses**

**Using Table chi**

**Chi Square**

**Distribution as an RV**

**Chi Square**

**Hypotheses**

**Goodness of Fit Example**

**Why** $(\frac{(Obs - Exp)^2}{Exp})$?

**2 by 2 Tables**

**Degrees of Freedom**

**Chi Square Hypotheses**

**Goodness of Fit Example**

**Chi Square**

**Hypotheses**

**Using Table chi**

**Chi Square**

**Distribution as an RV**

**Chi Square**

**Hypotheses**

**Goodness of Fit Example**

**Why** $(\frac{(Obs - Exp)^2}{Exp})$?

**2 by 2 Tables**

**Degrees of Freedom**

**Chi Square Hypotheses**

**Goodness of Fit Example**

**Chi Square**

**Hypotheses**

**Using Table chi**

**Chi Square**

**Distribution as an RV**

**Chi Square**

**Hypotheses**

**Goodness of Fit Example**

**Why** $(\frac{(Obs - Exp)^2}{Exp})$?

**2 by 2 Tables**

**Degrees of Freedom**

**Chi Square Hypotheses**

**Goodness of Fit Example**

**Chi Square**

**Hypotheses**

**Using Table chi**

**Chi Square**

**Distribution as an RV**

**Chi Square**

**Hypotheses**

**Goodness of Fit Example**

**Why** $(\frac{(Obs - Exp)^2}{Exp})$?

**2 by 2 Tables**

**Degrees of Freedom**
Chi Square Test of Independence

With Marginals

<table>
<thead>
<tr>
<th></th>
<th>0-shot</th>
<th>1-shot</th>
<th>2-shot</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
<td>46</td>
</tr>
<tr>
<td>no flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
<td>954</td>
</tr>
<tr>
<td></td>
<td>313</td>
<td>109</td>
<td>578</td>
<td>1000</td>
</tr>
</tbody>
</table>

If independent, the expected number of 0-shot people getting the flu would be

\[
\text{overall fraction getting flu} \times \text{number 0-shot people} = \frac{46}{1000} \times 313 = 14.4
\]
Chi Square Test of Independence

<table>
<thead>
<tr>
<th></th>
<th>0-shot</th>
<th>1-shot</th>
<th>2-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>no flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
</tr>
<tr>
<td></td>
<td>313</td>
<td>109</td>
<td>578</td>
</tr>
</tbody>
</table>

If independent, the expected number of 0-shot people getting the flu would be

\[
\text{overall fraction getting flu} \times \text{number 0-shot people} = \frac{46}{1000} \times 313 = 14.4
\]
Chi Square Test of Independence

If independent, the expected number of 0-shot people getting the flu would also be

\[
\text{overall fraction getting 0-shots} \times \text{number getting flu} = \frac{313}{1000} \times 46 = 14.4
\]
Chi Square Test of Independence

If independent, the expected number of 0-shot people getting the flu would be

\[ \text{overall fraction getting flu} \times \text{number 0-shot people} = \frac{46}{1000} \times 313 = 14.4 \]

This comes down to

\[ \text{expected value} = \frac{\text{row sum} \times \text{col sum}}{\text{total}} \]
## Chi Square Test of Independence

### With Expected Values

<table>
<thead>
<tr>
<th></th>
<th>0-shot</th>
<th>1-shot</th>
<th>2-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>flu</strong></td>
<td>24</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>(14.4)</td>
<td>(5.0)</td>
<td>(26.6)</td>
</tr>
<tr>
<td><strong>no flu</strong></td>
<td>289</td>
<td>100</td>
<td>565</td>
</tr>
<tr>
<td></td>
<td>(298.6)</td>
<td>(104)</td>
<td>(551.4)</td>
</tr>
<tr>
<td></td>
<td>313</td>
<td>109</td>
<td>578</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1000</td>
</tr>
</tbody>
</table>
Chi Square Test of Independence

<table>
<thead>
<tr>
<th></th>
<th>0-shot</th>
<th>1-shot</th>
<th>2-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>flu</td>
<td>24 (14.4)</td>
<td>9 (5.0)</td>
<td>13 (26.6)</td>
</tr>
<tr>
<td>no flu</td>
<td>289 (298.6)</td>
<td>100 (104)</td>
<td>565 (551.4)</td>
</tr>
<tr>
<td></td>
<td>313</td>
<td>109</td>
<td>578</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}
\]
### Chi Square Test of Independence

#### With Expected Values

<table>
<thead>
<tr>
<th></th>
<th>0-shot</th>
<th>1-shot</th>
<th>2-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>(14.4)</td>
<td>(5.0)</td>
<td>(26.6)</td>
</tr>
<tr>
<td>no flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
</tr>
<tr>
<td></td>
<td>(298.6)</td>
<td>(104)</td>
<td>(551.4)</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp} = \frac{(24 - 14.4)^2}{14.4} + \ldots (5 \text{ more terms})
\]
Chi Square Test of Independence

<table>
<thead>
<tr>
<th></th>
<th>0-shot</th>
<th>1-shot</th>
<th>2-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>(14.4)</td>
<td>(5.0)</td>
<td>(26.6)</td>
</tr>
<tr>
<td>no flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
</tr>
<tr>
<td></td>
<td>(298.6)</td>
<td>(104)</td>
<td>(551.4)</td>
</tr>
<tr>
<td></td>
<td>313</td>
<td>109</td>
<td>578</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1000</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp} = \frac{(24 - 14.4)^2}{14.4} + \ldots = 17.35
\]
### Chi Square Test of Independence

#### With Expected Values

<table>
<thead>
<tr>
<th></th>
<th>0-shot</th>
<th>1-shot</th>
<th>2-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>(14.4)</td>
<td>(5.0)</td>
<td>(26.6)</td>
</tr>
<tr>
<td>no flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
</tr>
<tr>
<td></td>
<td>(298.6)</td>
<td>(104)</td>
<td>(551.4)</td>
</tr>
<tr>
<td></td>
<td>313</td>
<td>109</td>
<td>578</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \sum \frac{(Obs - Exp)^2}{Exp} = \frac{(24 - 14.4)^2}{14.4} + \ldots = 17.35 \]

\(\chi\) square tests also have a degrees of freedom. For a test of independence, the number of rows (nr) and the number of columns (nc) in the original contingency table are the numbers of categories of the resp. cat. vars.

\[ df = (nr - 1)(nc - 1) \]
### Chi Square Test of Independence

#### With Expected Values

<table>
<thead>
<tr>
<th></th>
<th>0-shot</th>
<th>1-shot</th>
<th>2-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>(14.4)</td>
<td>(5.0)</td>
<td>(26.6)</td>
</tr>
<tr>
<td>no flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
</tr>
<tr>
<td></td>
<td>(298.6)</td>
<td>(104)</td>
<td>(551.4)</td>
</tr>
<tr>
<td>total</td>
<td>313</td>
<td>109</td>
<td>578</td>
</tr>
<tr>
<td></td>
<td>(313)</td>
<td>(109)</td>
<td>(578)</td>
</tr>
</tbody>
</table>

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp} = \frac{(24 - 14.4)^2}{14.4} + \ldots = 17.35$$

$$df = (nr - 1)(nc - 1)$$

Here

$$df = (2 - 1)(3 - 1) = 2.$$
Using Table $\chi$

<table>
<thead>
<tr>
<th>df</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.706</td>
<td>3.841</td>
<td>5.024</td>
<td>6.635</td>
<td>7.879</td>
</tr>
<tr>
<td>2</td>
<td>4.605</td>
<td>5.991</td>
<td>7.378</td>
<td>9.210</td>
<td>10.597</td>
</tr>
</tbody>
</table>

These are all critical values $\chi^*$. 

![Chi Square Distribution Table](image)
Using Table $\chi$

These are all critical values $\chi^*$. For example $P(\chi > 5.991) = .05$ for the $\chi$ square distribution with 2 df.
Our $\chi$ square-statistic of 17.35 is more extreme than any on the df=2 row of the table.
The picture shows what the critical value $\chi^* = 10.597$ for a tail prob. of .005 means.
So our tail probability and P-value are both less than .005 and we reject the null. Flu shots are making a difference.
What we just did is actually a hypothesis test.
Hypotheses:

- $H_0$: The categorical variables FluShot? and GetTheFlu? are independent.
- $H_a$: The categorical variables are not independent. (i.e. there is an association.)
Let $X_1, \ldots, X_d$ be $d$ independent standard normal random variables.
Chi Square Distribution as an RV

Let $X_1, \ldots, X_d$ be $d$ independent standard normal random variables.

The **Chi Square distribution with $d$ degrees of freedom** is given by:

$$
\chi^2 = X_1^2 + X_2^2 + \ldots X_d^2
$$
Chi Square Distribution as an RV

Let $X_1, \ldots, X_d$ be $d$ independent standard normal random variables.

The **Chi Square distribution with $d$ degrees of freedom** is given by:

$$\chi^2 = X_1^2 + X_2^2 + \ldots + X_d^2$$

### Chi Square Distribution Formula

$$f(x) = \frac{1}{\Gamma\left(\frac{d}{2}\right)2^{\frac{d}{2}}} x^{\frac{d}{2}-1} e^{-\frac{x}{2}}$$

where $d$ is the number of degrees of freedom.
Let $X_1, \ldots X_d$ be $d$ independent standard normal random variables. The **Chi Square distribution with $d$ degrees of freedom** is given by:

$$
\chi^2 = X_1^2 + X_2^2 + \ldots X_d^2
$$

### Chi Square Distribution Formula

$$
f(x) = \frac{1}{\Gamma\left(\frac{d}{2}\right)\left(2\right)^{d/2}} x^{d/2-1} e^{-x/2}
$$

where $d$ is the number of degrees of freedom. The mean is $d$ and the variance is $2d$. 

Chi Square Hypotheses

**Test of Independence**: Two Categorical Variables.
- $H_0$: The two categorical variables are independent.
- $H_a$: There is an association between the variables.
Chi Square Hypotheses

Test of Homogeneity: Multiple samples/populations. and Another Categorical Variable.

- $H_0$: The categorical variable has the same distribution within each population.
- $H_a$: The distributions differ among some of the populations.

(Which Population? could be viewed as a cat. var.)
Chi Square Hypotheses

Same Chi Square statistic,
Degrees of Freedom $= (\text{numrows}-1)(\text{numcols}-1)$ for both indep. and homog.
Chi Square Hypotheses

Same Chi Square statistic,
 Degrees of Freedom = (numrows-1)(numcols-1) for both indep. and homog.

The concept of “what is a population” does not have a clear answer, so we will not require you to distinguish between indep. and homog. on exams or homework.
(You need to know lots about the design to truly tell these apart.)
Chi Square Hypotheses

The concept of “what is a population” does not have a clear answer, so we will not require you to distinguish between indep. and homog. on exams or homework. (You need to know lots about the design to truly tell these apart.)

In fact indep. and homog. have different exact mathematical models, but both are approximated by the same \( \chi^2 \).
Chi Square Hypotheses

**Goodness of Fit**: Frequencies of One Cat Var AND a hypothesized distribution.

- $H_0$: The cat. var. follows the hypothesized distribution.
- $H_a$: The cat. var. doesn’t.

Degrees of Freedom $=$ number of cells - 1.
Suppose a maze has 3 exits. 90 rats run the maze and choose the following exits:

<table>
<thead>
<tr>
<th></th>
<th>door 1</th>
<th>door 2</th>
<th>door 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>23</td>
<td>36</td>
<td>31</td>
</tr>
</tbody>
</table>
Goodness of Fit Example

Suppose a maze has 3 exits. 90 rats run the maze and choose the following exits:

<table>
<thead>
<tr>
<th>frequency</th>
<th>door 1</th>
<th>door 2</th>
<th>door 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>23</td>
<td>36</td>
<td>31</td>
</tr>
</tbody>
</table>

The null hypothesis is that all doors are equally likely. Is this data strong evidence that not all doors are equally likely?
Goodness of Fit Example

<table>
<thead>
<tr>
<th>frequency</th>
<th>door 1</th>
<th>door 2</th>
<th>door 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>23</td>
<td>36</td>
<td>31</td>
</tr>
</tbody>
</table>

In this test, “equally likely” could be replaced by any hypothesized distribution of interest; e.g. there might be a theory to be evaluated that 38% exit door 1, 42% exit door 2, and 20% exit door 3.
Goodness of Fit Example

<table>
<thead>
<tr>
<th>frequency</th>
<th>door 1</th>
<th>door 2</th>
<th>door 3</th>
</tr>
</thead>
</table>

23 36 31

In this test, expected values come from the hypothesized distribution, NOT \( \frac{\text{rowsum} \cdot \text{colsum}}{\text{total}} \).
Goodness of Fit Example

(The row sum 90 is the total number of rats.)
The expected values are all $\frac{1}{3} \cdot 90$.

<table>
<thead>
<tr>
<th>frequency</th>
<th>door 1</th>
<th>door 2</th>
<th>door 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>23</td>
<td>36</td>
<td>31</td>
</tr>
<tr>
<td>expected values</td>
<td>(30)</td>
<td>(30)</td>
<td>(30)</td>
</tr>
</tbody>
</table>
Goodness of Fit Example

(The row sum 90 is the total number of rats.)

The expected values are all \( \frac{1}{3} \cdot 90 \).

<table>
<thead>
<tr>
<th>frequency</th>
<th>door 1</th>
<th>door 2</th>
<th>door 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>expected values</td>
<td>(30)</td>
<td>(30)</td>
<td>(30)</td>
</tr>
<tr>
<td>frequency</td>
<td>23</td>
<td>36</td>
<td>31</td>
</tr>
<tr>
<td>expected values</td>
<td>(30)</td>
<td>(30)</td>
<td>(30)</td>
</tr>
</tbody>
</table>

The Chi Square statistic is again

\[
\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}
\]

but the degrees of freedom are \( \text{numcells} - 1 \).
Goodness of Fit Example

(The row sum 90 is the total number of rats.)
The expected values are all \( \frac{1}{3} \cdot 90 \).

<table>
<thead>
<tr>
<th></th>
<th>door 1</th>
<th>door 2</th>
<th>door 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>23</td>
<td>36</td>
<td>31</td>
</tr>
<tr>
<td>expected values</td>
<td>(30)</td>
<td>(30)</td>
<td>(30)</td>
</tr>
</tbody>
</table>

Here \( df = 3 - 1 = 2 \) and

\[
\chi^2 = \frac{(23 - 30)^2}{30} + \frac{(36 - 30)^2}{30} + \frac{(31 - 30)^2}{30} = \frac{86}{30} = 2.87.
\]
Goodness of Fit Example

(The row sum 90 is the total number of rats.)
The expected values are all \( \frac{1}{3} \cdot 90 \).

<table>
<thead>
<tr>
<th>door 1</th>
<th>door 2</th>
<th>door 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>23</td>
<td>36</td>
</tr>
<tr>
<td>expected values</td>
<td>(30)</td>
<td>(30)</td>
</tr>
</tbody>
</table>

Our P-value is greater than .1. We retain the null. All exits might be equally likely.
Focus on 1 cell in our contingency table for a moment. Each cell corresponds to a particular value of each of the two categorical variables.
Focus on 1 cell in our contingency table for a moment. Notation:

\( n \) The total number of observed subjects.
\( \hat{p} \) The observed proportion in the cell of interest.
\( p \) The true proportion in the cell of interest.
Focus on 1 cell in our contingency table for a moment. Then

**Observed** \( n\hat{p} \).

**Expected** \( np \).

\[
\frac{(Obs - Exp)^2}{Exp} = \frac{(n\hat{p} - np)^2}{np} = \left( \frac{\hat{p} - p}{\sqrt{pq/n}} \right)^2
\]
Why $\frac{(Obs - Exp)^2}{Exp}$?

Focus on 1 cell in our contingency table for a moment. Then

- **Observed** $n\hat{p}$.
- **Expected** $np$.

$$\frac{(Obs - Exp)^2}{Exp} = \frac{(n\hat{p} - np)^2}{np} = \left(\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}\right)^2 q$$

Thus except for the extra $q$ (often near 1), we have the square of a Z-statistic!
Why $\frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$?

Focus on 1 cell in our contingency table for a moment. Then

- **Observed** $n\hat{p}$.
- **Expected** $np$.

$$\frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}} = \frac{(n\hat{p} - np)^2}{np} = \left( \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \right)^2 q$$

Thus except for the extra $q$ (often near 1), we have the square of a Z-statistic!

The degrees of freedom being less than the number of cells may be viewed as the result of careful analysis of all those extra $q$ factors.
2 × 2 Tables
Why \( df = (\text{numrows}-1)(\text{numcols}-1) \) for test of independence? (a 3 by 4 example: \( df = 2*3 = 6 \))
Degrees of Freedom

Why $df=(\text{numrows}-1)(\text{numcols}-1)$ for test of independence? (a 3 by 4 example: $df=2*3=6$)

Imagine these marginals . . .

<table>
<thead>
<tr>
<th>known</th>
<th>known</th>
<th>known</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>known</td>
<td>known</td>
<td>known</td>
<td>300</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>300</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>300</td>
<td>400</td>
</tr>
</tbody>
</table>
Idea: Consider all 3 by 4 contingency tables with the same marginal totals. These determine the expected values. How well the observed values match the expected tests the null hypothesis of independence.
The idea is that if we specify the observed values in the 6 known cells, then the rest of the cells are determined by the marginal totals. In other words $df=6$ corresponds to 6 cells can vary freely, the rest are determined.
Degrees of Freedom

Suppose we know the values in the $2 \times 3$ upper left:

<table>
<thead>
<tr>
<th></th>
<th>40</th>
<th>80</th>
<th>100</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>110</td>
<td>150</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>300</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>300</td>
<td>400</td>
<td>1000</td>
</tr>
</tbody>
</table>
Degrees of Freedom

Then the remaining values are determined by subtraction

<table>
<thead>
<tr>
<th></th>
<th>40</th>
<th>80</th>
<th>100</th>
<th>80</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>110</td>
<td>150</td>
<td>10</td>
<td>300</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>10</td>
<td>50</td>
<td>310</td>
<td>400</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>400</td>
<td>1000</td>
</tr>
</tbody>
</table>
MP as more Sensitive

Chi Square

Hypotheses

Goodness of Fit Example

Why \( \frac{(Obs - Exp)^2}{Exp} \)?

2 by 2 Tables

Degrees of Freedom

Chi Square Distribution as an RV

Chi Square Hypotheses

Using Table chi

Robustness
t-Conditions
Regression Inference Questions

Basic Setup:

1) Data \((x_i, y_i), 1 \leq i \leq n\) leads to line of regression

\[
\hat{y} = b_0 + b_1 x
\]

2) Assume an ideal line

\[
\hat{y} = \beta_0 + \beta_1 x
\]

3) Together with an error process \(\epsilon_i\) following an \(N(0, \sigma)\) law (independent for each \(i\).)

4) So that individual observations come from

\[
y_i = \beta_0 + \beta_1 x_i + \epsilon_i.
\]
Regression Inference Questions

Natural Questions in Regression:

1) Estimate $\beta_0$ and $\beta_1$.
2) Estimate the accuracy of $b_0$ and $b_1$ as estimators of $\beta_0$ and $\beta_1$.
3) Estimate $\sigma$, the standard deviation of the error process.

For a given value $x^*$ of $x$:

4) How accurately does the regression estimate $b_0 + b_1x^*$ approximate an actual $y$ observation when $x = x^*$.
5) How accurately does the regression estimate $b_0 + b_1x^*$ approximate the average of a lot of $y$ observations when $x = x^*$. 

\[ \frac{(OBS - EXP)^2}{EXP} \]