STA221

Neil Montgomery

Last edited: 2017-01-18 14:55

applications of χ^2 goodness-of-fit testing to two-way tables

contingency tables

Recall the gas pipelines data:

```
## # A tibble: 1,000 × 4
##      Leak  Size Material Pressure
##    <fctr> <ord>   <fctr>    <ord>
## 1      No  1.75  Aldyl A     High
## 2      No  1.75  Aldyl A      Med
## 3      No     1  Aldyl A      Low
## 4     Yes   1.5    Steel      Med
## 5      No     1    Steel     High
## 6     Yes     1    Steel     High
## 7     Yes  1.75  Aldyl A      Low
## 8      No  1.75    Steel      Med
## 9      No   1.5  Aldyl A     High
## 10     No  1.75    Steel     High
## # ... with 990 more rows
```

two-way table for “Leak” and “Pressure”

(Technically the two-way table doesn’t include the Sum row and column.)

	Pressure			
Leak	Low	Medium	High	Sum
No	278	247	277	802
Yes	66	61	71	198
Sum	344	308	348	1000

the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

Question 2: are the rows and columns *independent*?

the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

Question 2: are the rows and columns *independent*?

This question is answered using a *test of independence*.

the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

Question 2: are the rows and columns *independent*?

This question is answered using a *test of independence*.

The mechanics of both tests are identical. Only the interpretation is (slightly) different.

two-way table again

Count version:

	Pressure			
Leak	Low	Medium	High	Sum
No	278	247	277	802
Yes	66	61	71	198
Sum	344	308	348	1000

Proportion version. The six proportions at each combination of level of the two factor variables is the *joint distribution* of those two variables.

	Pressure			
Leak	Low	Medium	High	Sum
No	0.278	0.247	0.277	0.802
Yes	0.066	0.061	0.071	0.198
Sum	0.344	0.308	0.348	1.000

the marginal distrbutions

	Pressure			
Leak	Low	Medium	High	Sum
No	0.278	0.247	0.277	0.802
Yes	0.066	0.061	0.071	0.198
Sum	0.344	0.308	0.348	1.000

The *marginal* distributions of `Pressure` and `Leak` are:

Low	Med	High
0.344	0.308	0.348

No	Yes
0.802	0.198

the conditional distributions

There are lots of conditional distributions. For example, the conditional distributions for the `Pressure` *given* `Leak` equals `No` and *given* `Leak` equals `Yes` are in the two rows of this table:

	Low	Med	High
No	0.347	0.308	0.345
Yes	0.333	0.308	0.359

The conditional distributions for `Leak` given `Pressure` is equal to, respectively, `Low`, `Med`, and `High`, are in these three columns:

	Low	Med	High
No	0.808	0.802	0.796
Yes	0.192	0.198	0.204

diversion - if the marginal totals are *fixed*. . .

At some point there will be a "degrees of freedom" to consider, so let's do it now.

In all χ^2 goodness-of-fit tests, **the overall sample sizes are considered to be *fixed***. This includes all the row and column totals in these two-way table analyses.

Consider the following table with fixed "marginal" totals. How many cells am I "free" to play around with?

	Factor A			
Factor B	1	2	3	Sum
1				10
2				20
Sum	5	10	15	30

diversion - if the marginal totals are *fixed*. . .

At some point there will be a "degrees of freedom" to consider, so let's do it now.

In all χ^2 goodness-of-fit tests, **the overall sample sizes are considered to be *fixed***. This includes all the row and column totals in these two-way table analyses.

Consider the following table with fixed "marginal" totals. How many cells am I "free" to play around with?

	Factor A			
Factor B	1	2	3	Sum
1				10
2				20
Sum	5	10	15	30

Answer: only **two**. With fixed marginal totals I have two "degrees of freedom". The formula is $(r-1)(c-1)$ when there are r rows and c columns.

χ^2 test of homegeneity

Do the rows (or columns) come from the *same* distribution?

Specifically: do the rows (or columns) have the *same conditional disributions*?

χ^2 test of homegeneity

Do the rows (or columns) come from the *same* distribution?

Specifically: do the rows (or columns) have the *same conditional disributions*?

Let's compare the rows from before. They look pretty close, but not identical.

	Low	Med	High
No	0.347	0.308	0.345
Yes	0.333	0.308	0.359

χ^2 test of homegeneity

Do the rows (or columns) come from the *same* distribution?

Specifically: do the rows (or columns) have the *same conditional disributions*?

Let's compare the rows from before. They look pretty close, but not identical.

	Low	Med	High
No	0.347	0.308	0.345
Yes	0.333	0.308	0.359

The null hypothesis is "H_0: The rows have the same (conditional) distributions", and we keep all the marginal totals fixed.

some technical details. . .

Let's get rid of the numbers from the tables and use some more general symbols.

The conditional distributions, which H_0 says are the same:

	1	2	3
1	p_{11}	p_{12}	p_{13}
2	p_{21}	p_{22}	p_{23}

The counts, given *fixed* marginal totals:

	1	2	3	Sum
1	n_{11}	n_{12}	n_{13}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	$n_{2.}$
Sum	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

the expected cell counts E_{ij}, under the null hypothesis

So we end up with:

$$E_{11} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{n_{\cdot\cdot}}$$

the expected cell counts E_{ij}, under the null hypothesis

So we end up with:

$$E_{11} = \frac{n_{1.} \cdot n_{.1}}{n_{..}}$$

$$E_{ij} = \frac{n_{i.} \cdot n_{.j}}{n_{..}}$$

The counts we actually observe are called O_{ij}. We evaluate the deviation from H_0 using the formula:

$$\chi^2_{obs} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

getting the p-value

No surprise that this sum of squares has a χ^2 distributions. But with how many degrees of freedom?

getting the p-value

No surprise that this sum of squares has a χ^2 distributions. But with how many degrees of freedom?

$(r-1)(c-1)$, where r and c are the numbers of rows and columns.

getting the p-value

No surprise that this sum of squares has a χ^2 distributions. But with how many degrees of freedom?

$(r-1)(c-1)$, where r and c are the numbers of rows and columns.

In the example, the observed and expected cell counts are:

	Low	Med	High	Sum
No	278	247	277	802
Yes	66	61	71	198
Sum	344	308	348	1000

	Low	Med	High	Sum
No	275.89	247.02	279.10	802.00
Yes	68.11	60.98	68.90	198.00
Sum	344.00	308.00	348.00	1000.00

the full analysis

```
##
##  Pearson's Chi-squared test
##
## data:  Leak and Pressure
## X-squared = 0.16116, df = 2, p-value = 0.9226
```

$$P(\chi^2_2 \geq 0.1611609) = 0.9225807$$

There is no evidence against the null hypothesis.

example (Q23.14 from book)

All non-editorial publications from the NEJM were classified according to `Publication Year` and whether or not it contained a statistical analysis.

	1978-79	1989	2004-05	Sum
No Stats	90	14	40	144
Stats	242	101	271	614
Sum	332	115	311	758

Question: "Has there been a change in the use of statistics?"

example (Q23.14 from book)

Expected cell counts:

```
##            Publication Year
## Statistics 1978-79    1989 2004-05 Sum
##   No Stats   63.07   21.85   59.08 144
##   Stats     268.93   93.15  251.92 614
##   Sum       332.00  115.00  311.00 758
```

example (Q23.14 from book)

Expected cell counts:

```
##            Publication Year
## Statistics 1978-79   1989 2004-05 Sum
##   No Stats   63.07  21.85   59.08 144
##   Stats     268.93  93.15  251.92 614
##   Sum       332.00 115.00  311.00 758
```

Results:

```
##
##  Pearson's Chi-squared test
##
## data:  doctor_know
## X-squared = 25.282, df = 2, p-value = 0.000003237
```

what's left to do?

- the test for independence,

what's left to do?

- the test for independence,
- consider requirements and assumptions in order for these procedures to work,

what's left to do?

- the test for independence,
- consider requirements and assumptions in order for these procedures to work,
- informal post-test assessments of the results