STA221

Neil Montgomery

Last edited: 2017-07-06 17:50

ever wonder why the sample variance is divided by $n - 1$?

Look at the formula for sample variance:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}$$

The numerator is a sum of n squares, but the denominator is $n - 1$. Why?

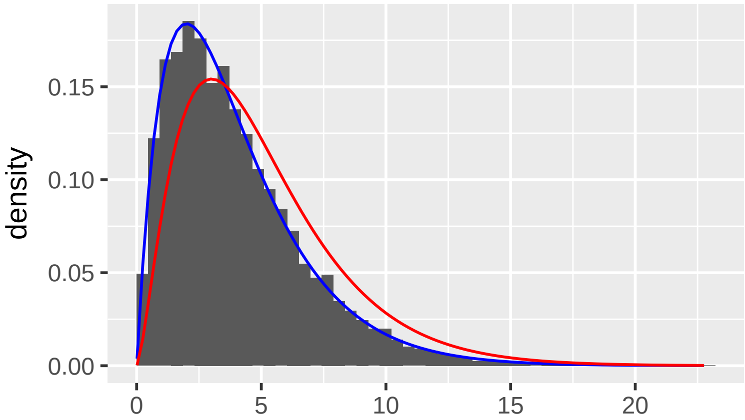
pictures of $\sum_{i=1}^{5}(x_i - \overline{x})^2$

I can simulate samples of size, say, 5 and compute that numerator, and make a histogram.

pictures of $\sum_{i=1}^{5}(x_i - \overline{x})^2$

I can simulate samples of size, say, 5 and compute that numerator, and make a histogram.

Here it is, with the χ^2_4 distribution in blue and the χ^2_5 in red:



a heuristic explanation

s^2 is calculated after fixing the value of $\overline{x}$

a heuristic explanation

s^2 is calculated after fixing the value of $\overline{x}$

So given $\overline{x}$ and *any* $n - 1$ of the n raw values, I can calculate that other raw value.

a heuristic explanation

s^2 is calculated after fixing the value of $\overline{x}$

So given $\overline{x}$ and *any* $n - 1$ of the n raw values, I can calculate that other raw value.

We say s^2 (given $\overline{x}$) only has $n - 1$ degrees of freedom.

is there evidence that something doesn't follow a given distribution?

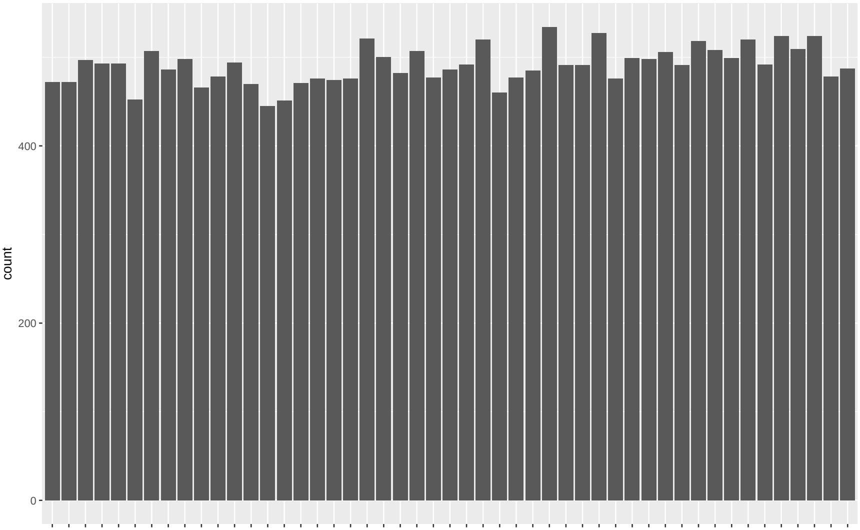
is a lottery “fair”

Lotto 6/49 is a Canadian lottery in which 49 identical balls are mixed together and 7 are selected, now twice per week. People can win money based on how many of the numbers they have out of the 6 on their ticket.

I found a list of every number ever picked up to January, 2017, here:
`http://portalseven.com/lottery/canada_lotto_649.jsp`

```
## # A tibble: 3,437 × 8
##                 date  num1  num2  num3  num4  num5  num6 bonus
##                <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Sat, Jan 14, 2017     1     6    19    30    32    44    33
## 2 Wed, Jan 11, 2017    24    34    36    38    42    43    30
## 3  Sat, Jan 7, 2017     1    10    18    19    23    27    48
## 4  Wed, Jan 4, 2017     2    11    13    23    35    48    30
## 5 Sat, Dec 31, 2016     3     5    14    18    26    28    40
## # ... with 3,432 more rows
```

all 49 numbers should appear with roughly the same frequency



categorical data, cells, observed cell counts

The dataset (now) consists of one variable called `numbers`. This is a *categorical*, or *factor* variable with 49 possible *levels*. There are 24050 observations.

categorical data, cells, observed cell counts

The dataset (now) consists of one variable called `numbers`. This is a *categorical*, or *factor* variable with 49 possible *levels*. There are 24050 observations.

A categorical variable is summarized by producing a table of *observed cell counts* (notation: O_i). In this case:

```
## # A tibble: 49 × 2
##   numbers   O_i
##    <fctr> <int>
## 1       1   472
## 2       2   472
## 3       3   497
## 4       4   493
## 5       5   493
## # ... with 44 more rows
```

expected cell counts

If Lotto $6/49$ is actually fair, each number would appear with probability $1/49 = 0.0204$ each.

After 24050 numbers have been selected, we would expect to see:

$$24050 \cdot \frac{1}{49} = 490.82$$

of each number.

These are called *expected cell counts* — calculated under the assumption of fairness as defined in this example. (Notation: E_i)

measuring the deviation from the assumption of fairness

Each O_i is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

measuring the deviation from the assumption of fairness

Each O_i is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

How many? As long as $E_i \geq 10$, the approximation will be good.

measuring the deviation from the assumption of fairness

Each O_i is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

How many? As long as $E_i \geq 10$, the approximation will be good.

(Note: this is exactly like the $np \geq 10$ or $np \geq 5$ suggestion that is given for the accuracy of a normal approximation to a binomial.)

measuring the deviation from the assumption of fairness

Each O_i is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

How many? As long as $E_i \geq 10$, the approximation will be good.

(Note: this is exactly like the $np \geq 10$ or $np \geq 5$ suggestion that is given for the accuracy of a normal approximation to a binomial.)

The overall deviation is measured as:

$$\sum_{i=1}^{n} \left(\frac{O_i - E_i}{\sqrt{E_i}} \right)^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

measuring the deviation - compared to what?

The expected cell counts E_i are computed *for a given fixed sample size* N.

measuring the deviation - compared to what?

The expected cell counts E_i are computed *for a given fixed sample size* N.

So given n along with any of the $n-1$ expected cell counts, we could compute that other expected cell count.

measuring the deviation - compared to what?

The expected cell counts E_i are computed *for a given fixed sample size* N.

So given n along with any of the $n - 1$ expected cell counts, we could compute that other expected cell count.

(This might *seem* trivial in this example because all the expected cell counts are the same - but this is only because our hypothesis is that all the cell probabilities are the same.)

measuring the deviation - compared to what?

The expected cell counts E_i are computed *for a given fixed sample size* N.

So given n along with any of the $n-1$ expected cell counts, we could compute that other expected cell count.

(This might *seem* trivial in this example because all the expected cell counts are the same - but this is only because our hypothesis is that all the cell probabilities are the same.)

We say

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

has $n-1$ degrees of freedom, and it follows (approximately) a χ^2_{n-1} distribution.

let's measure the deviation

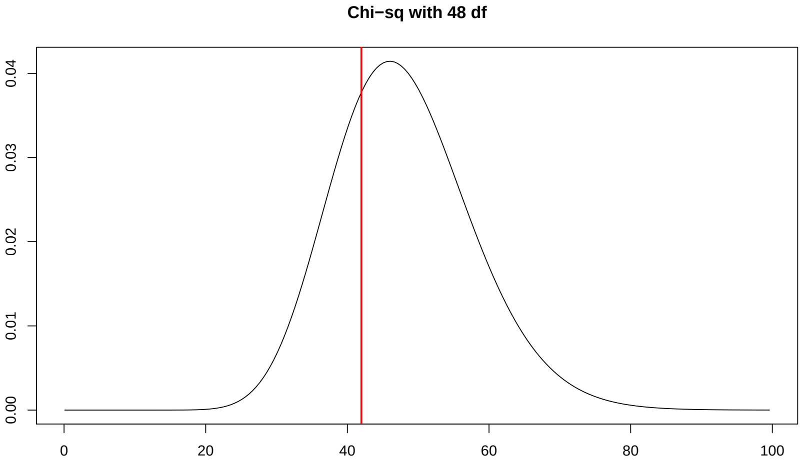
Here are the first few deviations (with $(O_i - E_i)^2/E_i$ called `D_i` for short):

```
## # A tibble: 49 × 4
##   numbers   O_i    E_i          D_i
##    <fctr> <int>  <dbl>        <dbl>
## 1       1   472 490.82 0.721634000
## 2       2   472 490.82 0.721634000
## 3       3   497 490.82 0.077813455
## 4       4   493 490.82 0.009682572
## 5       5   493 490.82 0.009682572
## # ... with 44 more rows
```

The sum of the `D_i` column is 41.99. Is this number surprising?

surprising, compared to what?

We know we should compare this number with the χ^2_{48} distribution. Here we can see we are not surprised. There is no evidence that Lotto 6/49 is unfair.



goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribution of interest.

goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribution of interest.

In the Lotto example, technically this statement is:

$$H_0 : p_1 = p_2 = \cdots = p_{49} = \frac{1}{49}$$

But usually we just make H_0 a simple written statement:

$$H_0 : \text{ the probabilities are all the same.}$$

goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribution of interest.

In the Lotto example, technically this statement is:

$$H_0 : p_1 = p_2 = \cdots = p_{49} = \frac{1}{49}$$

But usually we just make H_0 a simple written statement:

$$H_0 : \text{ the probabilities are all the same.}$$

The "distribution of interest" is technically the "discrete uniform distribution on the outcomes $\{1, 2, 3, \ldots, 49\}$"

goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribution of interest.

In the Lotto example, technically this statement is:

$$H_0 : p_1 = p_2 = \cdots = p_{49} = \frac{1}{49}$$

But usually we just make H_0 a simple written statement:

$$H_0 : \text{ the probabilities are all the same.}$$

The "distribution of interest" is technically the "discrete uniform distribution on the outcomes $\{1, 2, 3, \ldots, 49\}$"

The alternative hypothesis is the negation of the null. We don't normally bother to write it down.

goodness of fit as formal hypothesis test - II

Given a sample size N and the null hypothesis probabilities, compute the n expected cell counts. In this case:

$$E_i = Np_i = 24050 \cdot \frac{1}{49} = 490.82$$

goodness of fit as formal hypothesis test - II

Given a sample size N and the null hypothesis probabilities, compute the n expected cell counts. In this case:

$$E_i = Np_i = 24050 \cdot \frac{1}{49} = 490.82$$

Compute the *observed value of the test statistic*:

$$\chi^2_{\text{obs}} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 41.99$$

goodness of fit as formal hypothesis test - II

Given a sample size N and the null hypothesis probabilities, compute the n expected cell counts. In this case:

$$E_i = Np_i = 24050 \cdot \frac{1}{49} = 490.82$$

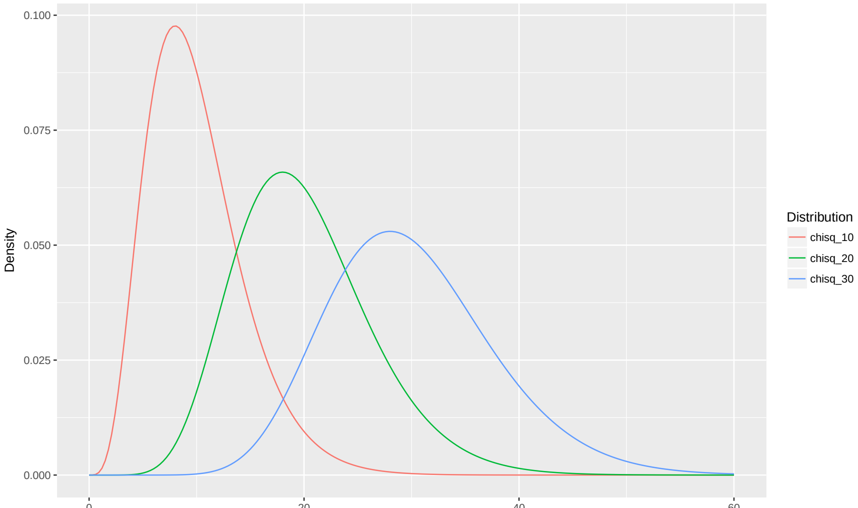
Compute the *observed value of the test statistic*:

$$\chi^2_{\text{obs}} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 41.99$$

Calculate the p-value based on χ^2_{obs} being approximately χ^2_{n-1}.

goodness-if-fit testing p-value

A p-value is the *probability of observing a more extreme value*, in the sense of being further from where the null hypothesis “lives”, which is where in this case?



goodness-of-fit testing p-value

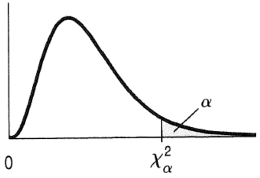
The p-value is $P(\chi^2_{48} \geq 41.99) = 0.7165747$

goodness-of-fit testing p-value

The p-value is $P(\chi^2_{48} \geq 41.99) = 0.7165747$

On tests you'll need to use a table. Here's a close-up of a table I found in a book:

Table X
Values of χ^2_α



Right-tail probability / df	0.10	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819

goodness-of-fit testing p-value (from table)

28	37.916	41.337	44.461	48.278	50.994
29	39.087	42.557	45.722	59.588	52.336
30	40.256	43.773	46.979	50.892	53.672
40	51.805	55.759	59.342	63.691	66.767
50	63.167	67.505	71.420	76.154	79.490
60	74.397	79.082	83.298	88.381	91.955
70	85.527	90.531	95.023	100.424	104.213

From a table the best you can do is to estimate the p-value.

All this together is called the "χ^2 goodness-of-fit test."

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”

The (only?) suitable numerical summary for two categorical/factor variables at a time is a so-called contingency table, or two-way table.

(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:

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

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 distribution. But with how many degrees of freedom?

getting the p-value

No surprise that this sum of squares has a χ^2 distribution. 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 distribution. 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 readings)

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 readings)

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 readings)

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 = 3.237e-06
```

χ^2 test for homogeneity

Do the rows (columns) have the same distributions? The mechanics were:

1. Because the row (column) probabilities are the same under H_0, we ended up with:

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

χ^2 test for homogeneity

Do the rows (columns) have the same distributions? The mechanics were:

1. Because the row (column) probabilities are the same under H_0, we ended up with:

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

2. Then we compared $\sum_{i,j} (E_{ij} - O_{ij})^2 / E_{ij}$ with a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

“homogeneity” versus “independence”"

We are going to learn a different test (of “independence”) that is mechanically identical. The difference is subtle, and is in how the data are collected.

“homogeneity” versus “independence”"

We are going to learn a different test (of “independence”) that is mechanically identical. The difference is subtle, and is in how the data are collected.

In a test of homogeneity, we are seeing if two or more sub-populations (the rows or columns) have the same distribution with respect to the levels of a factor variable.

“homogeneity” versus “independence””

We are going to learn a different test (of “independence”) that is mechanically identical. The difference is subtle, and is in how the data are collected.

In a test of homogeneity, we are seeing if two or more sub-populations (the rows or columns) have the same distribution with respect to the levels of a factor variable.

e.g. NEJM papers from subpopulations: 1978-79, 1989, and 2004-05. Factor variable was “did the paper have statistics in it, yes/no?”

“homogeneity” versus “independence”"

We are going to learn a different test (of “independence”) that is mechanically identical. The difference is subtle, and is in how the data are collected.

In a test of homogeneity, we are seeing if two or more sub-populations (the rows or columns) have the same distribution with respect to the levels of a factor variable.

e.g. NEJM papers from subpopulations: 1978-79, 1989, and 2004-05. Factor variable was “did the paper have statistics in it, yes/no?”

In a test of independence, we have one population and are seeing if categorization into levels of two factor variables is done independently.

“homogeneity” versus “independence”"

We are going to learn a different test (of “independence”) that is mechanically identical. The difference is subtle, and is in how the data are collected.

In a test of homogeneity, we are seeing if two or more sub-populations (the rows or columns) have the same distribution with respect to the levels of a factor variable.

e.g. NEJM papers from subpopulations: 1978-79, 1989, and 2004-05. Factor variable was “did the paper have statistics in it, yes/no?”

In a test of independence, we have one population and are seeing if categorization into levels of two factor variables is done independently.

Consider Q13 “Childbirth” from the readings. Researchers followed up on 1178 births, classifying them as “did mother have epidural” and “was child breastfeeding at six months.”

testing independence

What would it mean for epidural status and breastfeeding status to be independent?

testing independence

What would it mean for epidural status and breastfeeding status to be independent?

It means: if you know a woman had an epidural, it would not change the distribution of the two levels of breastfeeding.

testing independence

What would it mean for epidural status and breastfeeding status to be independent?

It means: if you know a woman had an epidural, it would not change the distribution of the two levels of breastfeeding.

Or vice-versa - independence is symmetric.

testing independence

What would it mean for epidural status and breastfeeding status to be independent?

It means: if you know a woman had an epidural, it would not change the distribution of the two levels of breastfeeding.

Or vice-versa - independence is symmetric.

The null hypothesis is best stated in words that describe the two variables, e.g. "H_0: epidural status is independent of breastfeeding status."

testing independence

What would it mean for epidural status and breastfeeding status to be independent?

It means: if you know a woman had an epidural, it would not change the distribution of the two levels of breastfeeding.

Or vice-versa - independence is symmetric.

The null hypothesis is best stated in words that describe the two variables, e.g. "H_0: epidural status is independent of breastfeeding status."

The test is done once again *with row and column totals taken as constant*.

independence—the details

The joint distribution along with the marginals (in a 2 by 3 example):

	1	2	3	Row Marginal
1	p_{11}	p_{12}	p_{13}	$p_{1\cdot}$
2	p_{21}	p_{22}	p_{23}	$p_{2\cdot}$
Column Marginal	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot 3}$	

independence—the details

The joint distribution along with the marginals (in a 2 by 3 example):

	1	2	3	Row Marginal
1	p_{11}	p_{12}	p_{13}	$p_{1\cdot}$
2	p_{21}	p_{22}	p_{23}	$p_{2\cdot}$
Column Marginal	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot 3}$	

Independence just means:

$$P(\text{Row Level 1 } \textit{and} \text{ Column Level 1}) = P(\text{Row Level 1})P(\text{Column Level 1})$$

and so on for all row and column levels.

independence—the details

The joint distribution along with the marginals (in a 2 by 3 example):

	1	2	3	Row Marginal
1	p_{11}	p_{12}	p_{13}	$p_{1\cdot}$
2	p_{21}	p_{22}	p_{23}	$p_{2\cdot}$
Column Marginal	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot 3}$	

Independence just means:

$$P(\text{Row Level 1 } \textit{and} \text{ Column Level 1}) = P(\text{Row Level 1})P(\text{Column Level 1})$$

and so on for all row and column levels.

In short:

$$p_{ij} = p_{i\cdot}p_{\cdot j}$$

independence—the details with fixed row/column totals

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

independence—the details with fixed row/column totals

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

Translating the last probability statement into a cell count version gives:

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

independence—the details with fixed row/column totals

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

Translating the last probability statement into a cell count version gives:

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

And from here we get the (same!) expected cell count as before:

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

Those are the cell counts one would get under perfect independence.

the χ^2 test for independence

Compare:

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

with a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

the χ^2 test for independence

Compare:

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

with a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

Childbirth example in full:

the χ^2 test for independence

Compare:

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

with a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

Childbirth example in full:

H_0 : epidural status is independent of breastfeeding status.

H_a : (usually omitted) ... is not independent ...

childbirth example

The observed data:

```
##           Breastfeeding
## Epidural    No  Yes  Sum
##       No   284  498  782
##       Yes  190  206  396
##       Sum  474  704 1178
```

childbirth example

The observed data:

```
##           Breastfeeding
## Epidural   No  Yes  Sum
##      No   284  498  782
##      Yes  190  206  396
##      Sum  474  704 1178
```

The expected cell counts:

```
##           Breastfeeding
## Epidural    No   Yes
##      No  314.7 467.3
##      Yes 159.3 236.7
```

childbirth example

The results:

```
##
##  Pearson's Chi-squared test
##
## data:  childbirth
## X-squared = 14.869, df = 1, p-value = 0.0001152
```

Conclusion: there is evidence that epidural status is not independent of breastfeeding status.

example "Twins" (Part VI Review Q15)

A JAMA paper studied the relationship between the quality of prenatal care and the birth circumstances.

example "Twins" (Part VI Review Q15)

A JAMA paper studied the relationship between the quality of prenatal care and the birth circumstances.

H_0: is quality of prenatal care independent of the birth circumstances?

example "Twins" (Part VI Review Q15)

A JAMA paper studied the relationship between the quality of prenatal care and the birth circumstances.

H_0: is quality of prenatal care independent of the birth circumstances?

Observed:

```
##               Twin Births
## Level of Care Preterm complex Preterm simple Term / postterm Sum
##    Intensive               18             15              28  61
##    Adequate                46             43              65 154
##    Inadequate              12             13              38  63
##    Sum                     76             71             131 278
```

"Twins"

Expected:

```
##                Twin Births
## Level of Care Preterm complex Preterm simple Term / postterm Sum
##    Intensive             16.7           15.6            28.7  61
##    Adequate              42.1           39.3            72.6 154
##    Inadequate            17.2           16.1            29.7  63
##    Sum                   76.0           71.0           131.0 278
```

“Twins”

Expected:

```
##               Twin Births
## Level of Care Preterm complex Preterm simple Term / postterm Sum
##    Intensive             16.7           15.6            28.7  61
##    Adequate              42.1           39.3            72.6 154
##    Inadequate            17.2           16.1            29.7  63
##    Sum                   76.0           71.0           131.0 278
```

Results:

```
##
##  Pearson's Chi-squared test
##
## data:  twins
## X-squared = 6.1437, df = 4, p-value = 0.1887
```

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.
- **and when expected cell counts are all over 5.**

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.
- **and when expected cell counts are all over 5.**
  - Other suggestions are common—all are fine.

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.
- **and when expected cell counts are all over 5.**
  - Other suggestions are common—all are fine.
  - Similar in spirit to the $np \geq 5$ suggestion in the "normal approximation to the binomial"

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.
- **and when expected cell counts are all over 5.**
  - Other suggestions are common—all are fine.
  - Similar in spirit to the $np \geq 5$ suggestion in the "normal approximation to the binomial"

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.
- **and when expected cell counts are all over 5.**
  - Other suggestions are common—all are fine.
  - Similar in spirit to the $np \geq 5$ suggestion in the "normal approximation to the binomial"

The second condition is the main thing to check.

χ^2 tests - other matters to ponder (or not)

The readings mention a few other things to “check” when it comes to goodness-of-fit tests. Here is some commentary:

- “counted data condition,” which even the readings admit is silly.

χ^2 tests - other matters to ponder (or not)

The readings mention a few other things to "check" when it comes to goodness-of-fit tests. Here is some commentary:

- "counted data condition," which even the readings admit is silly.
  - the summary tables with counts are not datasets - they are summaries.

χ^2 tests - other matters to ponder (or not)

The readings mention a few other things to "check" when it comes to goodness-of-fit tests. Here is some commentary:

- "counted data condition," which even the readings admit is silly.
  - the summary tables with counts are not datasets - they are summaries.
  - apparently some people try to apply the χ^2 procedures to other kinds of summary tables, which is why the readings emphasize this point as a warning.

χ^2 tests - other matters to ponder (or not)

The readings mention a few other things to "check" when it comes to goodness-of-fit tests. Here is some commentary:

- "counted data condition," which even the readings admit is silly.
  - the summary tables with counts are not datasets - they are summaries.
  - apparently some people try to apply the χ^2 procedures to other kinds of summary tables, which is why the readings emphasize this point as a warning.
- "randomization condition," which has more to do with the possibility of *inferring something about a larger population, or not* than anything to do with χ^2 tests per se.

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:
  - there was actually a very mild lack of independence among the rows in the dataset, but it wouldn't have effected the results.

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:
  - there was actually a very mild lack of independence among the rows in the dataset, but it wouldn't have effected the results.
  - **the expected cell counts were all much larger than 5.**

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:
  - there was actually a very mild lack of independence among the rows in the dataset, but it wouldn't have effected the results.
  - **the expected cell counts were all much larger than 5.**
  - I did indeed analyse counts.

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:
  - there was actually a very mild lack of independence among the rows in the dataset, but it wouldn't have effected the results.
  - **the expected cell counts were all much larger than 5.**
  - I did indeed analyse counts.
  - The analysis was not on any sample at all - I used *all* numbers ever drawn!

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:
  - there was actually a very mild lack of independence among the rows in the dataset, but it wouldn't have effected the results.
  - **the expected cell counts were all much larger than 5.**
  - I did indeed analyse counts.
  - The analysis was not on any sample at all - I used *all* numbers ever drawn!
- All other examples satisfied the $E_{ij} \geq 5$ condition, which is the main thing that should always be verified and commented on.

post-hoc investigations of χ^2 tests using residuals

The χ^2 tests are based on the following standardized deviation of *observed* from *expected*:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

(similar with *ij* subscripts.)

post-hoc investigations of χ^2 tests using residuals

The χ^2 tests are based on the following standardized deviation of *observed* from *expected*:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

(similar with *ij* subscripts.)

These can be called *standardized residuals*, where the $O_i - E_i$ are just the *residuals.*

post-hoc investigations of χ^2 tests using residuals

The χ^2 tests are based on the following standardized deviation of *observed* from *expected*:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

(similar with *ij* subscripts.)

These can be called *standardized residuals*, where the $O_i - E_i$ are just the *residuals.*

These are approximately $N(0, 1)$, so one can glance at the cell-by-cell residuals to get information about which cells had the largest deviation from expected.

standardized residuals example - I (pipeline)

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

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

##       Pressure
## Leak     High    Low    Med
##   No   -0.125  0.127 -0.001
##   Yes   0.253 -0.256  0.002
```

standardized residuals example - II (births)

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  childbirth
## X-squared = 14.388, df = 1, p-value = 0.0001487
```

```
##         Breastfeeding
## Epidural     No    Yes
##      No  -1.728  1.418
##      Yes  2.429 -1.993
```