Comparing three or more groups

Prof. Maria Tackett

STA 199

Click for PDF of slides

An old example...

ORIGINAL RESEARCH

Annals of Internal Medicine

Coffee Drinking and Mortality in 10 European Countries

A Multinational Cohort Study

Coffee	Died	Did not die
Non-drinker	1039	5438
Occasional drinker	4440	29712
Regular drinker	3601	24934

We have more than two samples! Non-coffee drinkers, occasional drinkers, and regular drinkers.

Is there an *association* between coffee drinking *status* and whether somebody died? Are the two independent?

STA 199

A new hypothesis test...

Coffee	Died	Did not die
Non-drinker	1039	5438
Occasional drinker	4440	29712
Regular drinker	3601	24934

- H_0: Coffee-drinking category and health outcome are independent; there is no association between the two variables
- H_a: Coffee-drinking category and health outcome are NOT independent; there is an association between the two variables

Review

Coffee	Died	Did not die
Non-drinker	1039	5438
Occasional drinker	4440	29712
Regular drinker	3601	24934

If H_0 were true, then we would expect:

- P(Non-Drinker) x P(Died) = P(Non-drinker AND Died)

Review

Coffee	Died	Did not die
Non-drinker	1039	5438
Occasional drinker	4440	29712
Regular drinker	3601	24934

If H_0 were true, then we would expect:

- P(Non-Drinker) x P(Died) = P(Non-drinker AND Died)
- P(Occasional Drinker) x P(Died) = P(Occasional drinker AND Died)

Review

Coffee	Died	Did not die
Non-drinker	1039	5438
Occasional drinker	4440	29712
Regular drinker	3601	24934

If H_0 were true, then we would expect:

- P(Non-Drinker) x P(Died) = P(Non-drinker AND Died)
- P(Occasional Drinker) x P(Died) = P(Occasional drinker AND Died)
- P(Regular Drinker) x P(Died) = P(Regular drinker AND Died)

STA 199

Review

Coffee	Died	Did not die
Non-drinker	1039	5438
Occasional drinker	4440	29712
Regular drinker	3601	24934

If H_0 were true, then we would expect:

- P(Non-Drinker) x P(Died) = P(Non-drinker AND Died)
- P(Occasional Drinker) x P(Died) = P(Occasional drinker AND Died)
- P(Regular Drinker) x P(Died) = P(Regular drinker AND Died)
- P(Non-Drinker) x P(Lived) = P(Non-drinker AND Lived)

STA 199

Review

Coffee	Died	Did not die
Non-drinker	1039	5438
Occasional drinker	4440	29712
Regular drinker	3601	24934

If H_0 were true, then we would expect:

- P(Non-Drinker) x P(Died) = P(Non-drinker AND Died)
- P(Occasional Drinker) x P(Died) = P(Occasional drinker AND Died)
- P(Regular Drinker) x P(Died) = P(Regular drinker AND Died)
- P(Non-Drinker) x P(Lived) = P(Non-drinker AND Lived)
- P(Occasional Drinker) x P(Lived) = P(Occasional drinker AND Lived)

Review

Coffee	Died	Did not die
Non-drinker	1039	5438
Occasional drinker	4440	29712
Regular drinker	3601	24934

If H_0 were true, then we would expect:

- P(Non-Drinker) x P(Died) = P(Non-drinker AND Died)
- P(Occasional Drinker) x P(Died) = P(Occasional drinker AND Died)
- P(Regular Drinker) x P(Died) = P(Regular drinker AND Died)
- P(Non-Drinker) x P(Lived) = P(Non-drinker AND Lived)
- P(Occasional Drinker) x P(Lived) = P(Occasional drinker AND Lived)
- P(Regular Drinker) x P(Lived) = P(Regular drinker AND Lived)

Observed vs. expected counts

Coffee	Died	Did not die
Non-drinker	1039	5438
Occasional drinker	4440	29712
Regular drinker	3601	24934

Let's investigate non-coffee drinking and dying:

- P(Non-Drinker) = $6477/69164 \approx 0.09365$
- P(Died) = $9080/69164 \approx 0.131$

If these were independent, we would *expect* P(Non-Drinker AND Died) to be $6477/69164 \times 9080/69164 \approx 0.012$. So, we expect approximately 850 study participants to be non-drinkers who died.

STA 199

Observed vs. expected counts

Coffee	Died	Did not die
Non-drinker	1039	5438
Occasional drinker	4440	29712
Regular drinker	3601	24934

The **observed** number is 1039, for a difference of 189 participants between the observed and expected counts.

STA 199

Observed vs. expected counts

Coffee	Died	Did not die
Non-drinker	1039	5438
Occasional drinker	4440	29712
Regular drinker	3601	24934

The **observed** number is 1039, for a difference of 189 participants between the observed and expected counts.

Is this strong evidence against the claim of independence?

STA 199

Observed vs. expected counts

Well, that was just one cell! There are five more cells in which there may be differences between observed and expected counts.

Observed vs. expected counts

Well, that was just one cell! There are five more cells in which there may be differences between observed and expected counts.

How can we sum up these differences in a principled way, and use it to conduct statistical inference?

The chi-square test

The chi-squared test has a very nice motivation in terms of comparing observed vs. the expected counts that we would expect if H_0 were true.

If these total differences are **"large enough,"** then we **reject** the null hypothesis.

The chi-square test

The chi-squared test has a very nice motivation in terms of comparing observed vs. the expected counts that we would expect if H_0 were true.

If these total differences are **"large enough,"** then we **reject** the null hypothesis.

- To combine differences across table cells, we need to square them before adding them up (so that negative differences aren't canceled out by positive differences)

STA 199

The chi-square test

The chi-squared test has a very nice motivation in terms of comparing observed vs. the expected counts that we would expect if H_0 were true.

If these total differences are **"large enough,"** then we **reject** the null hypothesis.

- To combine differences across table cells, we need to square them before adding them up (so that negative differences aren't canceled out by positive differences)

- We will also scale these differences by the expected count (a difference of 189 participants isn't large when thinking about 100,000 total observations, but is huge when thinking about 300 total observations!)

STA 199

The chi-square test statistic

The chi-square χ^2 test statistic is

$$\sum_{i \in cells}^{r \times c} \frac{(O_i - E_i)^2}{E_i},$$

where $r \times c$ is the number of cells in the table (rows times columns), i indexes across all cells, O_i is the expected count in cell i, and E_i is the expected count in cell i.

STA 199

The chi-square test statistic

The chi-square χ^2 test statistic is

$$\sum_{i \in cells}^{r \times c} \frac{(O_i - E_i)^2}{E_i},$$

where $r \times c$ is the number of cells in the table (rows times columns), i indexes across all cells, O_i is the expected count in cell i, and E_i is the expected count in cell i.

This statistic is the total squared difference between the observed and expected cell counts, scaling by the expected cell count for each cell.

STA 199

The chi-square test statistic

The chi-square χ^2 test statistic is

$$\sum_{i \in cells}^{r \times c} \frac{(O_i - E_i)^2}{E_i},$$

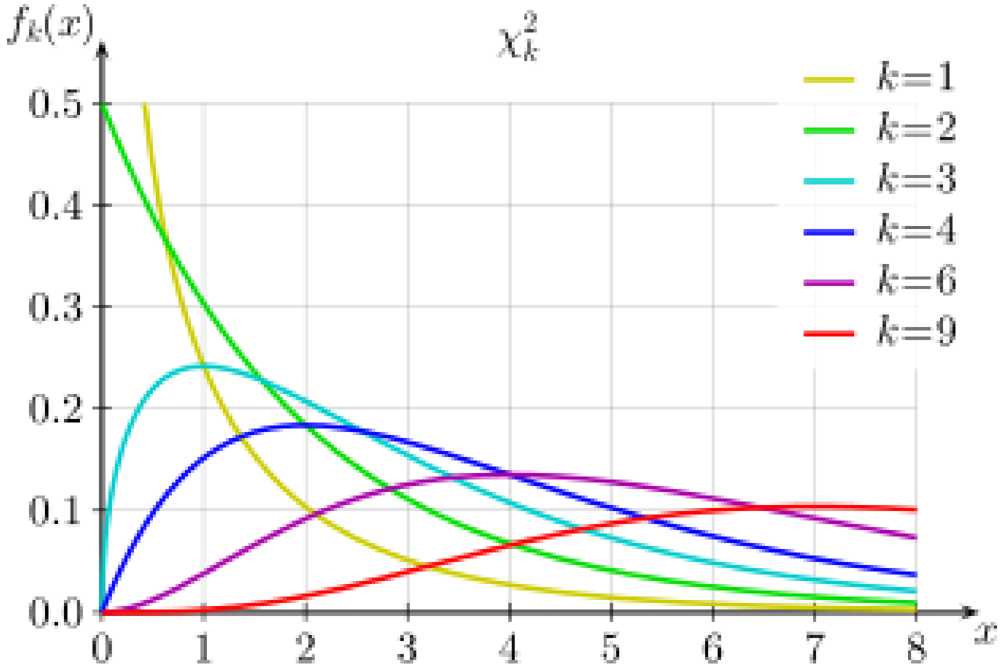
where $r \times c$ is the number of cells in the table (rows times columns), i indexes across all cells, O_i is the expected count in cell i, and E_i is the expected count in cell i.

This statistic is the total squared difference between the observed and expected cell counts, scaling by the expected cell count for each cell.

Under H_0, the distribution of this sum is approximated by a χ^2 distribution with $(r-1) \times (c-1)$ degrees of freedom.

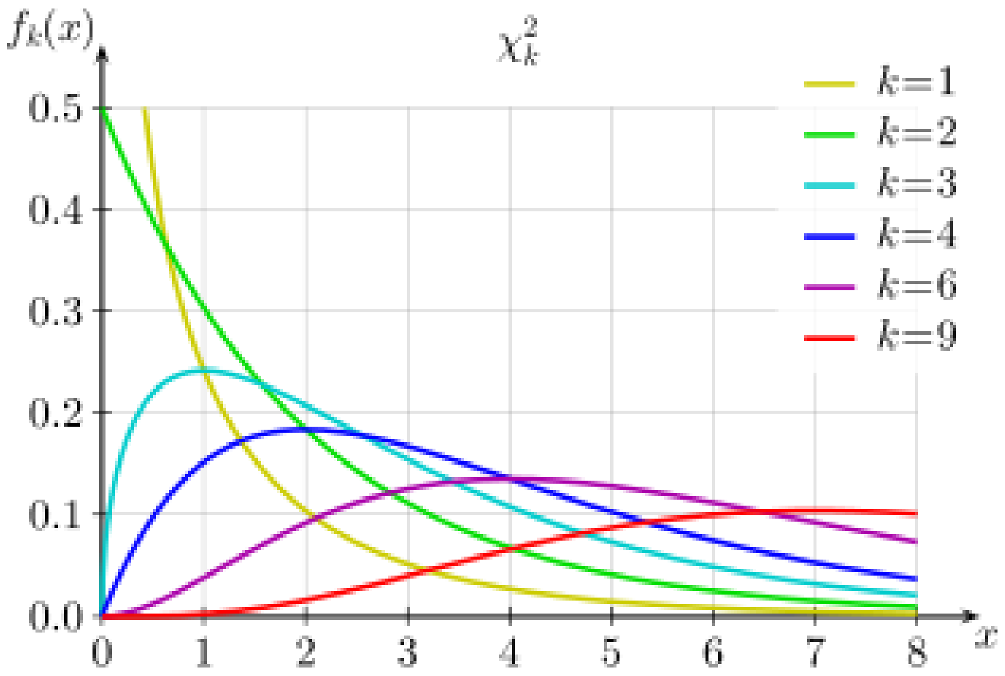
STA 199

Chi-squared distributions



STA 199

Chi-squared distributions



Remember, we only reject if the difference is "large enough." So, we only examine the right-tail. That is, the probability of seeing our χ^2 statistic **or larger** when calculating p-values.

STA 199

Implementation in R

Luckily, you don't have to calculate all the expected counts by hand, create the test statistic, and manually compare to a chi-square distribution.

```
coffee_data %>%
  slice(1:10)
```

```
## # A tibble: 10 x 2
##    coffee                     health_outcome
##    <chr>                      <chr>
##  1 Does not drink coffee Died
##  2 Does not drink coffee Died
##  3 Does not drink coffee Died
##  4 Does not drink coffee Died
##  5 Does not drink coffee Died
##  6 Does not drink coffee Died
##  7 Does not drink coffee Died
##  8 Does not drink coffee Died
```

STA 199

Chi-square test using infer

```
coffee_data %>%
  chisq_test(formula = health_outcome ~ coffee)
```

```
## # A tibble: 1 x 3
##   statistic chisq_df  p_value
##       <dbl>    <int>    <dbl>
## 1      55.2        2 1.05e-12
```

Chi-square test using infer

```
coffee_data %>%
  chisq_test(formula = health_outcome ~ coffee)
```

```
## # A tibble: 1 x 3
##   statistic chisq_df  p_value
##       <dbl>    <int>    <dbl>
## 1      55.2        2 1.05e-12
```

Formally assess the hypothesis that coffee drinking and health outcome are independent.

What might we conclude given these data?