MA22004

Seminar 9

Dr Eric Hall

3/12/2020

Announcements

Reminders

- It is week 9! You should have read the remainder of §7 Categorical data on **Perusall**.
- Feedback for Lab 5 and Class Test is forthcoming.

Upcoming

- Lab 6 due **Thursday 3 Dec** at **17:00** (late accepted until Sunday without penalty).
- Worksheet #9 Categorical data will be posted to Blackboard soon.
- Read §8 Quality control on **Perusall** before next seminar.

Testing for goodness of fit using χ^2

- Evaluate distribution of one categorical variable that has more than two levels
- Compare distribution of the categorical variable to a hypothetical distribution

Example: Scottish prison population

Ethnicity is not a factor that predisposes one to commit a crime. Therefore, we would expect the prison population of Scotland to mirror the true population of Scotland.

From Scotland's Census 2011, the total population by ethnicity (simplified):

Ethnicity	% Population
White	96.02
Asian	2.66
Black	0.68
Other	0.64
Total	100.00

Example: Scottish prison population

We would like to compare the distribution of ethnicity from a sample of the prison population to the distribution of ethnicity in Scotland.

H_0 (*nothing* going on): The prison population is a simple random sample from the total population. The observed counts of prisoners by ethnicity **follow the same** ethnicity **distribution** as population.

H_a (something *is* going on): The prison population is not a simple random sample from the total population. The observed counts of prisoners by ethnicity **do not follow the same** ethnicity **distribution** as population.

How do we evaluate these hypothesis?

- Quantify how different the observed counts are from the expected counts.
- Large deviations from what would be expected based on sampling variation alone would be strong evidence against H_0.

Goodness of fit test

How well do the observed data fit expected distribution?

Suppose we had a random sample of 2000 prisoners.

Ethnicity	% Population	Expected
White	96.02	
Asian	2.66	
Black	0.68	
Other	0.64	
Total	100.00	

Conditions

Independence - sample observations must be independent

- Random sampling
- If sampling w/o replacement, $m < 10$ population
- Each factor only contributes to one scenario (cell) in table

Sample size

Each scenario (cell) must have at least 5 expected cases

Anatomy of a test statistic

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

1. Identify difference between a point estimate and expected value assuming null hypothesis is true.
2. Standardizes the difference using the standard error of the point estimate.

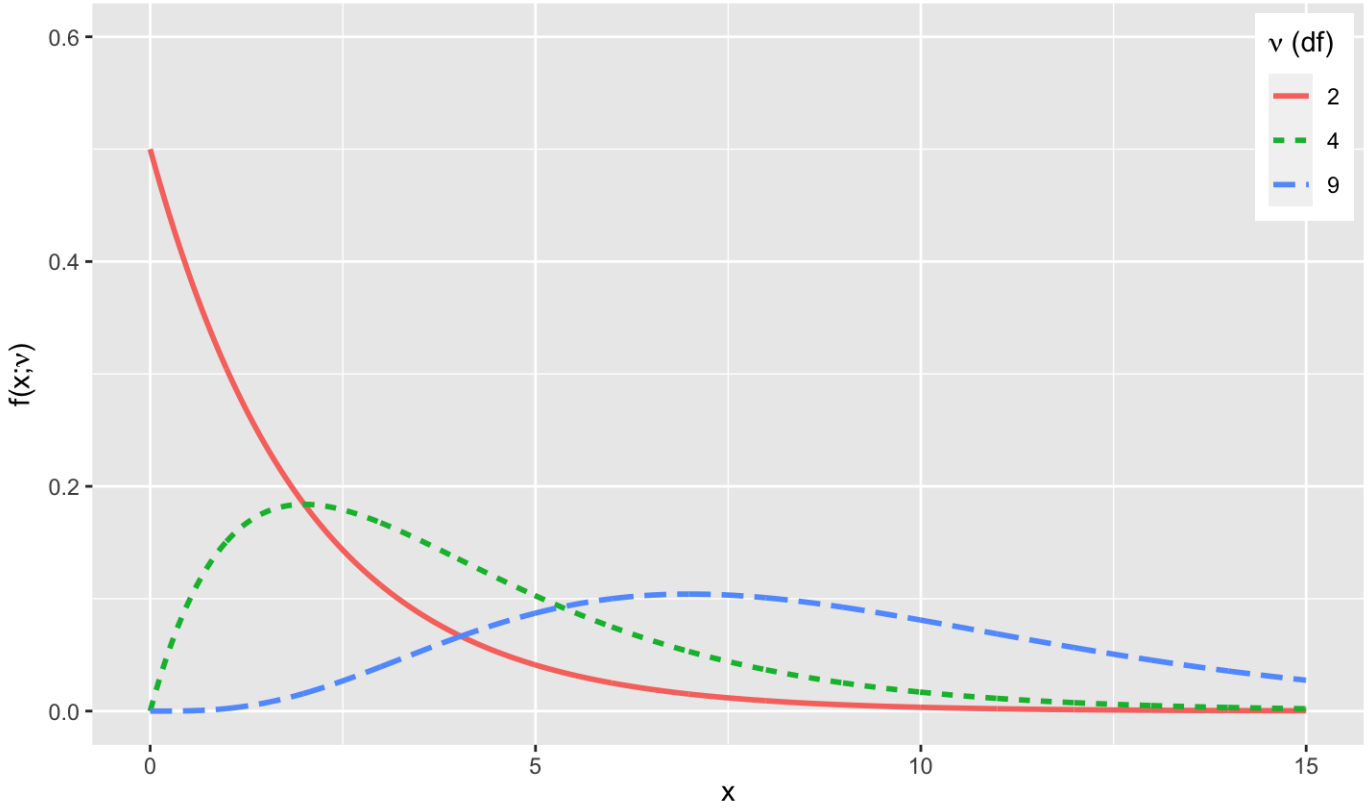
χ^2 statistic

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = V \qquad \sim \chi^2(k-1)$$

Why square?

1. Positive standardized difference
2. Unusual differences become even more unusual.

Recall the χ^2 distribution



Degrees of freedom (df) influences *shape*, *center* and *spread* of distribution.

Example: sample data

A random sample of 2000 inmates in year 2011-2012 yields:

Ethnicity	% Population	Expected	Observed
White	96.02	1920	1930
Asian	2.66	53	35
Black	0.68	14	27
Other	0.64	13	8
Total	100.00	2000	2000

Compare observed sample to hypothetical expected distribution to understand if the observed differences are due to chance variation alone.

Example: calculating test statistic

E_i	O_i
1920	1930
53	35
14	27
13	8

P-value

The P-value is always positive and a higher value of the test statistic means larger deviations from the null hypothesis.

The P-value is given by the **tail area above the calculated statistic**.

```
pchisq(20.1598, df = 4-1, lower.tail = FALSE)
```

```
[1] 0.0001572779
```

Since P-value of 0.00016 we would, e.g. at 0.05 level, **reject** the null hypothesis.

The data provide sufficient evidence that the prison population of Scotland is not representative of the general population. The ethnicity distribution of the prison population is different from the ethnicity distribution of the general population.

Summary

Today we discussed goodness of fit tests using the χ^2 distribution.